

This E-Book and More

From

<http://ali-almukhtar.blogspot.com>

Subhas Chandra Mukhopadhyay
Yueh-Min Huang
Editors

Sensors

Advancements in Modeling,
Design Issues, Fabrication
and Practical Applications



Lecture Notes Electrical Engineering

Volume 21

S.C. Mukhopadhyay · R.Y.M. Huang (Eds.)

Sensors

Advancements in Modeling, Design Issues,
Fabrication and Practical Applications

Subhas Chandra Mukhopadhyay
School of Engineering and
Advanced Technology (SEAT)
Massey University (Turitea Campus)
Palmerston North
New Zealand
S.C.Mukhopadhyay@massey.ac.nz

Ray Y.M. Huang
Department of Engineering Science
National Cheng-Kung University
Tainan
Taiwan
huang@mail.ncku.edu.tw

ISBN: 978-3-540-69030-6

e-ISBN: 978-3-540-69033-7

Library of Congress Control Number: 2008929912

© 2008 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Contents

Part I Electromagnetic Sensors

Modern CMOS Hall Sensors with Integrated Magnetic Concentrators . . . 3
Christian Schott and Samuel Huber

Commercial Magnetic Sensors (Hall and Anisotropic Magnetoresistors) . . 23
Michael J. Haji-Sheikh

Improving the Accuracy of Magnetic Sensors 45
Pavel Ripka

Modelling Electromagnetic Field Sensors 61
Ian Woodhead

Dielectric Characterization of Biological Tissues: Constraints Related to Ex Vivo Measurements 75
Mustapha Nadi

Estimation of Property of Sheep Skin to Modify the Tanning Process Using Interdigital Sensors 91
V. Kasturi and S. C. Mukhopadhyay

Part II Fiber Optic/Optical Fibre Sensors

Fiber Bragg Gratings Evanescent Wave Sensors: A View Back and Recent Advancements 113
Andrea Cusano, Antonello Cutolo and Michele Giordano

Optical Fibre Humidity Sensors Using Nano-films 153
Jesus M. Corres, Ignacio R. Matias and Francisco J. Arregui

Overview of the OPTO-EMI-SENSE Project: Optical Fibre Sensor Network for Automotive Emission Monitoring	179
E. Lewis, G. Dooly, E. Hawe, C. Fitzpatrick, P. Chambers, J. Clifford, W.Z. Zhao, T. Sun, K.T.V. Grattan, J. Lucas, M. Degner, H. Ewald, S. Lochmann, G. Bramann, F. Gili, and E. Merlone-Borla	
Part III Wireless Sensors	
Wireless Sensor Networks and Applications	199
Yueh-Min Huang, Meng-Yen Hsieh and Frode Eika Sandnes	
Wireless Sensor Network Transport Layer: State of the Art	221
Md. Abdur Rahman, Abdulmotaleb El Saddik and Wail Gueaieb	
Part IV Sensors for Tracking and Navigation	
Real Time Tracking and Monitoring of Human Behavior in an Indoor Environment	249
Maki K. Habib	
Dynamic VRML-Based Navigable 3D Map for Indoor Location-Aware Systems	269
Wan-Young Chung and Chi-Shian Yang	
Part V Ultrasonic Sensor	
Ultrasonic Sensing: Fundamentals and its Applications to Nondestructive Evaluation	287
Ikuo Ihara	
Part VI Image Sensor	
Multimodal Image Sensor Fusion Using Independent Component Analysis	309
Nedeljko Cvejic, Nishan C. Canagarajah and David R. Bull	
Part VII Vision Sensing	
Fast Image Capture and Vision Processing For Robotic Applications	329
Gourab Sen Gupta and Donald Bailey	
Part VIII Sensors Based on Human Parameter	
Affection Based Multi-robot Team Work	355
Sajal Chandra Banik, Keigo Watanabe, Maki K. Habib and Kiyotaka Izumi	

Part IX Displacement Sensor

Displacement Sensor Using Magnetostrictive Wire and Decrease of its Hysteresis Error 379
Hiroyuki Wakiwaka

Part X THZ Sensor

Submillimeter-Wave Coherent and Incoherent Sensors for Space Applications 387
Goutam Chattopadhyay

Index 415

Guest Editorial

This special issue titled “Sensors: Advancements in Modeling, Design Issues, Fabrication and Practical Applications” in the book series of “Lecture Notes in Electrical Engineering” contains the invited chapters from experts working on different sensors related research in different parts of the world. A total of 19 chapters are presented in this book.

The first group of six chapters are on electromagnetic sensors. In the first chapter, Christian Schott and Samuel Huber have described modern CMOS Hall sensors with integrated magnetic concentrators which have dramatically enhanced magnetic field measurement performance. Michael Haji-Sheikh has described commercial magnetic sensors both Hall and anisotropic magneto-resistors in the second chapter. Pavel Ripka in the third chapter has described techniques for improvement of the accuracy of magnetic sensors. Many of these ideas can be used for other types of sensors and measurement systems in general. Sensors may use the attenuation, velocity and electrical dispersion of electromagnetic waves as a direct or intermediate step in the transduction process. Ian Woodhead in his chapter has discussed the modelling of the sensor response which provides the means to predict its spatial and temporal sensitivity and accuracy. In his chapter, Mustapha Nadi has briefly reviewed and discussed metrological aspects relating to electrical characterization of biological tissues. Experimental results obtained on different kinds of biological tissues (blood and bone) are presented and discussed as examples according to the influencing constraints specific to their physiological nature. V. Kasturi and S. C. Mukhopadhyay have reported a sensing system to improve the tanning process of sheep skin to produce better quality leather. The dielectric properties of the sheep skin are measured using a sensor system based on an interdigital sensor. Once the skin is converted into leather the process cannot be reversed. Over-treatment of the sheep skin can damage the quality of leather or inadequate tanning may not yield the desired level of quality, so it is important to know the appropriate tanning process required for the skin.

The second group contains three papers on fibre optic/optical fibre sensors. Andrea Cusano and his group has overviewed the most relevant milestones of the

technological evolution of Fiber Bragg Gratings Evanescent Wave Sensors in thirty years from the discovery of Kenneth Hill in 1978. They have also reviewed the advancement in the area of FBGs evanescent wave sensors as valuable technological platforms for chemical and biological applications. The emphasis have been placed on principles of operation, technological developments and overall performances discussing perspectives and challenges lying ahead. In the next chapter Jesus M. Corres and his colleagues have attempted to approach the fibre optic humidity sensing technology to scientists unfamiliar with the field. They have presented a general review of this type of sensors with emphasis in the techniques based on nanostructured coatings. These devices have been classified according to the sensing mechanism and taking also into account the different methods of fabrication and the sensing materials they are based on. Elfed Lewis and his group have reported the development an optical fibre based system capable of monitoring the presence of exhaust gas emissions and measuring their temperature on line in the exhaust system of a modern vehicle. There exists at present no commercial sensor, which is capable of providing online measurements of these exhaust gases as required by European legislation. The design of this sensor using low cost and compact optical components, which make it suitable for operation on board a vehicle, has been discussed.

The next two chapters are in the category of wireless sensors. Ray Huang and his group have reviewed some of the fundamental mechanisms of wireless sensor networks including their architecture, topology, data integration, routing techniques, and applications. Sensor network applications include both military and civilian monitoring in both rural and urban environments. Wireless sensor networks hold great potential for improving control, conservation, convenience, efficiency, reliability flexibility, and safety in network environments. In the next chapter Md Abdur Rahman and his colleagues have described the essence of a generic transport layer of a MultihopWireless Sensor Network (WSN). The transport layer of the Internet handles the congestion generated due to the network traffic and the end-to-end reliability of individual packets. Similar to the Internet, many WSN applications require a congestion control mechanism to regulate the amount of traffic injected within the WSN to avoid packet loss and to guarantee end-to-end reliable packet/event delivery. WSN researchers thus argue the presence of a transport layer for WSN similar to the Internet.

The following two chapters are on sensors for tracking and navigation. In his chapter Maki Habib has reported the development of a real time 3D sensor system and a new concept based on space decomposition by encoding its operational space using limited number of laser spots. The sensor system uses the richness and the strength of the vision while reducing the data-load and computational cost. The chapter presents the development and implementation of an intelligent 3D Fiber Grating (FG) based vision-system that can monitor and track human being status in real time for monitoring purposes to support wide range of applications. The 3D visual sensor is able to measure three-dimensional information with respect to human, objects and surrounding environment. The sensor system consists of a

CCD camera, a laser spot array generator (constitutes: laser diode and driver, lens, fiber gratings and holder), and a processing unit with alarm facilities and interfacing capabilities to a higher-level controller and decision-making along with a user-friendly interface. The system works by projecting a two-dimensional matrix of laser spots generated through two perpendicularly overlaid layers of FGs. In the next chapter W.Y. Chung and his group have developed a three Dimensional Navigation Viewer (3DNV), a convergence of location-aware application and three-dimensional (3D) graphics technology for a 3D visualization of location-aware information. The system allows visualization of situational information in a complete, 3D model of indoor environments equipped with instantly updated route results, synchronized with physical world. The approach is validated via indoor context-aware technologies, Cricket and Received Signal Strength Indication (RSSI). The overall results provide a valuable insight into the novel integration approach between 3D graphics standard, Virtual Reality Modeling language (VRML) and indoor location-aware systems.

In his chapter Ikuo Ihara has discussed the fundamentals of ultrasonic sensing techniques that can be used in the various fields of engineering and science. He has also included some advanced techniques used for non-destructive evaluations. At first, basic characteristics of ultrasonic waves propagating in media are described briefly. Secondly, basic concepts for measuring ultrasonic waves are described with introductory subjects of ultrasonic transducers that generate and receive ultrasonic waves. Finally, specialized results demonstrating the capabilities of using a buffer rod sensor for ultrasonic monitoring at high temperatures are presented.

In their chapter Nedeljko Cvejic and his group have presented a novel multi-modal image fusion algorithm using the Independent Component Analysis (ICA). Region-based fusion of ICA coefficients is implemented, in which the mean absolute value of ICA coefficients is used as an activity indicator for the given region. The ICA coefficients from given regions are consequently weighted using the Piella fusion metric in order to maximise the quality of the fused image. The proposed method exhibits significantly higher performance than the basic ICA algorithm and improvement over the other state-of-the-art algorithms.

In the next chapter G. Sen Gupta and his colleague have described a technique to significantly increase the speed of image processing for robot identification in a global-vision based system, targeted at real-time applications. Of major significance are the proposed discrete and small look-up tables for Y, U and V color thresholds. A new YUV color space has been proposed which significantly improves the speed of color classification. The look-up tables can be easily updated in real-time and are thus suitable for adaptive thresholding. The experimental results confirm that the proposed algorithm greatly improves the performance of the image processing system. The results are compared with other commonly used methods such as a composite look-up table which is indexed using RGB pixel values.

In their chapter Sajal Chandra Banik and his group have proposed approaches to multi-robot task allocation and cooperation in a chronological way such that they can be studied and compared for future development with affection based augmentation. In respect of some drawbacks (like high communication overhead, dead lock, etc.) with the existing approaches, they have presented the affection based task allocation and cooperation that has been used for a very few cases. They have also presented the complexity of the affective method and give some hints to compensate the complexity problems.

H. Wakiwaka has described a displacement sensor using magnetostrictive wire. It is a sensor which estimates the displacement from propagation time of an elastic wave that is caused and detected by using the magnetostrictive effect and the inverse-magnetostrictive effect. This sensor can be used for measurement up to 60 meters in simple structure, therefore it is appropriate for industry applications. Various methods for reducing the hysteresis error has been proposed.

In the last chapter G. Chattopadhyay has provided an overview of the state-of-the-art of submillimeter-wave sensors for a variety of space-borne applications and their performance and capabilities. Most of the radiation in the Universe is emitted at wavelengths longer than 10 m (30 THz), and this peaks at about 100 m (3 THz), excluding the contributions from the cosmic microwave background (CMB). Radiation in these wavelengths highlights warm phenomena, processes of change such as star formation, formation of planetary systems, and galaxy evolution; atmospheric constituents and dynamics of the planets and comets and tracers for global monitoring and the ultimate health of the earth. Sensors at far-infrared and submillimeter wavelengths provide unprecedented sensitivity for astrophysical, planetary, earth observing, and ground-based imaging instruments. Very often, for a spaced based platforms where the instruments are not limited by atmospheric losses and absorption, the overall instrument sensitivity is dictated by the sensitivity of the sensors themselves.

We do hope that the readers will find this issue interesting and useful in their research as well as in practical engineering work in the area of modern sensors and sensing technology. We are very happy to be able to offer the readers such a diverse special issue, both in terms of its topical coverage and geographic representation.

Finally, we would like to whole-heartedly thank all the authors for their contribution to this issue.

Subhas Chandra Mukhopadhyay, Guest Editor
School of Engineering and Advanced Technology (SEAT),
Massey University (Turitea Campus)
Palmerston North, New Zealand
S.C.Mukhopadhyay@massey.ac.nz

Ray Y. M. Huang, Guest Editor
Department of Engineering Science
National Cheng-Kung University
Tainan, Taiwan
huang@mail.ncku.edu.tw



Dr Subhas Chandra Mukhopadhyay graduated from the Department of Electrical Engineering, Jadavpur University, Calcutta, India in 1987 with a Gold medal and received the Master of Electrical Engineering degree from Indian Institute of Science, Bangalore, India in 1989. He obtained the PhD (Eng.) degree from Jadavpur University, India in 1994 and Doctor of Engineering degree from Kanazawa University, Japan in 2000.

During 1989–90 he worked almost 2 years in the research and development department of Crompton Greaves Ltd., India. In 1990 he joined as a Lecturer in the Electrical Engineering department, Jadavpur University, India and was promoted to Senior Lecturer of the same department in 1995.

Obtaining Monbusho fellowship he went to Japan in 1995. He worked with Kanazawa University, Japan as researcher and Assistant professor till September 2000.

In September 2000 he joined as Senior Lecturer in the Institute of Information Sciences and Technology, Massey University, New Zealand where he is working currently as an Associate professor. His fields of interest include Sensors and Sensing Technology, Electromagnetics, control, electrical machines and numerical field calculation etc.

He has authored 185 papers in different international journals and conferences, co-authored a book and written a book chapter and edited eight conference proceedings. He has also edited two special issues of international journals (IEEE Sensors Journal and IJISTA) as guest editor and a book with Springer-Verlag.

He is a Fellow of IET (UK), a senior member of IEEE (USA), an associate editor of IEEE Sensors journal and IEEE Transactions on Instrumentation and Measurements. He is in the editorial board of e-Journal on Non-Destructive Testing, Sensors and Transducers, Transactions on Systems, Signals and Devices (TSSD), Journal on the Patents on Electrical Engineering, Journal of Sensors. He is in the

technical programme committee of IEEE Sensors conference, IEEE IMTC conference and IEEE DELTA conference. He was the Technical Programme Chair of ICARA 2004 and ICARA 2006. He was the General chair of ICST 2005, ICST 2007. He is organizing the IEEE Sensors conference 2008 at Lecce, Italy as General Co-chair and IEEE Sensors conference 2009 at Christchurch, New Zealand as General Chair.



Yueh-Min Huang is a Distinguished Professor and Chairman of the Department of Engineering Science, National Cheng-Kung University, Taiwan, R.O.C. His research interests include Multimedia Communications, Wireless Networks, Embedded Systems, and Artificial Intelligence. He received his MS and PhD degrees in Electrical Engineering from the University of Arizona in 1988 and 1991 respectively. He has co-authored 2 books and has published about 160 refereed professional research papers. He has completed 10 PhD and over 80 MSES thesis students. Dr. Huang has received many research awards, such as the Best Paper Award of 2007 IEA/AIE Conference, Best Paper Award of the Computer Society of the Republic of China in 2003, the Awards of Acer Long-Term Prize in 1996, 1998, and 1999, Excellent Research Awards of National Microcomputer and Communication Contests in 2006. He also received many funded research grants from National Science Council, Ministry of Education, Industrial Technology of Research Institute, and Institute of Information Industry. Dr Huang has been invited to give talks or served frequently in the program committee at national and international conferences. Dr Huang is in the editorial board of the Journal of Wireless Communications and Mobile Computing, the Journal of Internet Technology, International Journal of Internet Protocol Technology, International Journal of Ad Hoc and Ubiquitous Computing, Journal of Security and Communication Networks and serves as an associate editor for Journal of Computer Systems, Networks, and Communications as well as International Journal of

Communication Systems. He was the Technical Programme Chair of Symposium on Digital Life Technologies (SDLT2007). He was the General chair of VIP2007. He is organizing the SDLT2008, PCM2008 and ICST2008. Huang is a member of the IEEE as well as IEEE communication, computer, and computational intelligence societies.

List of Contributors

Francisco J. Arregui

Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra. Pamplona, Spain

Donald Bailey

School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand, e-mail: d.g.bailey@massey.ac.nz

Sajal Chandra Banik

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan, e-mail: baniksajal@yahoo.com

G. Bramann

Department of Electrical Engineering & Computer Science, Hochschule Wismar, Germany

David R. Bull

Centre for Communications Research, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

Nishan C. Canagarajah

Centre for Communications Research, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

P. Chambers

Department of Electronic & Computer Engineering, University of Limerick, Ireland

Goutam Chattopadhyay

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, US, e-mail: goutam@jpl.nasa.gov

Wan-Young Chung

Department of Computer & Information Engineering, Dongseo University, Busan 617-716, Korea, e-mail: wychung@dongseo.ac.kr

J. Clifford

Department of Electronic & Computer Engineering, University of Limerick,
Ireland

Jesus M. Corres

Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de
Navarra, Pamplona, Spain, e-mail: jmcorres@unavarra.es

Andrea Cusano

Optoelectronic Division, Engineering Department, University of Sannio, C.so
Garibaldi 107 82100 Benevento, Italy, e-mail: a.cusano@unisannio.it

Antonello Cutolo

Optoelectronic Division, Engineering Department, University of Sannio, C.so
Garibaldi 107 82100 Benevento, Italy

Nedeljko Cvejic

Centre for Communications Research, University of Bristol, Merchant Venturers
Building, Woodland Road, Bristol BS8 1UB, UK, e-mail: nc332@cam.ac.uk

M. Degner

Department of Electrical Engineering & Information Technology, University of
Rostock, Germany

G. Dooly

Department of Electronic & Computer Engineering, University of Limerick,
Ireland

Abdulmotaleb El Saddik

Multimedia Communications Research Laboratory (MCRLab), University of
Ottawa, Canada, e-mail: abed@mcrlab.uottawa.ca

H. Ewald

Department of Electrical Engineering & Information Technology, University of
Rostock, Germany

C. Fitzpatrick

Department of Electronic & Computer Engineering, University of Limerick,
Ireland

F. Gili

Centro Ricerche Fiat, Strada Torino 50, 10043 Orbassano (TO), Italy

Michele Giordano

Institute for Composite and Biomedical Materials, National Research Council
(IMCB-CNR), P.le Enrico Fermi 1, 80055 Portici, Italy

K.T.V. Grattan

School of Engineering & Mathematical Sciences, City University, London EC1
0HB, UK

Wail Gueaieb

Machine Intelligence, Robotics, and Mechatronics (MIRAM) Laboratory,
University of Ottawa, Canada, e-mail: wgueaieb@site.uottawa.ca

Gourab Sen Gupta

School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand, e-mail: g.sengupta@massey.ac.nz

Maki K. Habib

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan. Currently with the American University in Cairo, Egypt, e-mail: maki@ieee.org

Michael J. Haji-Sheikh

Associate Professor of Electrical Engineering, College of Engineering and Engineering Technology, Northern Illinois University, De Kalb Illinois 60115 e-mail: mhsheikh@ceet.niu.edu

E. Hawe

Department of Electronic & Computer Engineering, University of Limerick, Ireland

Meng-Yen Hsieh

Department of Computer Science and Information Engineering, Providence University, Taiwan, ROC, e-mail: mengyen0501@gmail.com

Yueh-Min Huang

Department of Engineering Science, National Cheng-Kung University, Taiwan, ROC, e-mail: huang@mail.nck.edu.tw

Samuel Huber

Melexis Technologies SA, Switzerland

Ikuo Ihara

Department of Mechanical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan, e-mail: ihara@mech-nagaokaut.ac.jp

Kiyotaka Izumi

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan, e-mail: izumi@me.saga-u.ac.jp

V. Kasturi

School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand

E. Lewis

Department of Electronic & Computer Engineering, University of Limerick, Ireland

S. Lochmann

Department of Electrical Engineering & Computer Science, Hochschule Wismar, Germany

J. Lucas

Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 3GJ, UK

Ignacio R. Matias

Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra. Pamplona, Spain

E. Merlone-Borla

Centro Ricerche Fiat, Strada Torino 50, 10043 Orbassano (TO), Italy

S. C. Mukhopadhyay

School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand

Mustapha Nadi

Nancy University; L.I.E.N., BP 239 Faculty of Sciences and Techniques, 54506 Vandoeuvre les Nancy, e-mail: mustapha.nadi@lien.uhp-nancy.fr

Md. Abdur Rahman

Multimedia Communications Research Laboratory (MCRLab), University of Ottawa, Canada, e-mail: rahman@mcrmlab.uottawa.ca

Pavel Ripka

Czech Technical University, Faculty of Electrical Engineering, Department of Measurement, Prague, Czech Republic, e-mail: ripka@fel.cvut.cz

Frode Eika Sandnes

Faculty of Engineering, Oslo University College, Norway, e-mail: frodes@hio.no

Christian Schott

Melexis Technologies SA, Switzerland, e-mail: csc@melexis.com

T. Sun

School of Engineering & Mathematical Sciences, City University, London EC1 0HB, UK

Hiroyuki Wakiwaka

Shinshu University, 4-17-1, Wakasato, Nagano, 380-8553, Japan, e-mail: wakiwak@shinshu-u.ac.jp

Keigo Watanabe

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan, e-mail: watanabe@me.saga-u.ac.jp

Ian Woodhead

Lincoln Ventures Ltd, Lincoln University, Canterbury, New Zealand, e-mail: woodhead@lvl.co.nz

Chi-Shian Yang

Department of Ubiquitous IT, Graduate School of Design & IT, Dongseo University, Busan 617-716, Korea

W.Z. Zhao

School of Engineering & Mathematical Sciences, City University, London EC1 0HB, UK

Part I
Electromagnetic Sensors

Modern CMOS Hall Sensors with Integrated Magnetic Concentrators

Christian Schott and Samuel Huber

Abstract Combining modern CMOS Hall sensors with integrated magnetic concentrators dramatically enhances magnetic field measurement performance. A first key feature is that one, two, or all three magnetic field components can be measured in a small spot. On-chip digital signal processing allows for the evaluation of the flux density vector direction instead of merely the flux density strength. By this principle very robust and thermally insensitive contact-less angular and linear position sensors can be made. The second key feature is the passive magnetic flux amplification of the integrated concentrator by up to one order of magnitude. This feature is particularly interesting for the measurement of low flux density as for example around a current carrying conductor. The combination of low-field amplification and multi-axis capability allows even to address applications like the electronic compass, which have up to now been considered as far too demanding for Hall sensors.

Keywords Hall sensor · contactless position measurement · angle sensor · current sensor · electronic compass

1 Hall Effect and Hall Element

A Hall element is a type of galvano-magnetic sensor which transduces a magnetic flux density into an electrical signal. It is typically a plate-shape conductor with four equally spaced electrical contacts on its outer edge. When the device is exposed to a flux density B and a current is sent through the plate for example from contact 1 to 3, then a voltage appears between the other two contacts 2 and 4 (Fig. 1, *left*). The sign of this voltage depends on the orientation of B and it is proportional to the product of current I and magnetic flux density B . For constant current, the output voltage is a direct measure of the magnetic flux density.

$$V_{\text{Out}} = S \cdot I \cdot B \quad \text{where } S \text{ is the current related sensitivity}$$

Christian Schott
Melexis Technologies SA, Switzerland, e-mail: csc@melexis.com

Samuel Huber
Melexis Technologies SA, Switzerland

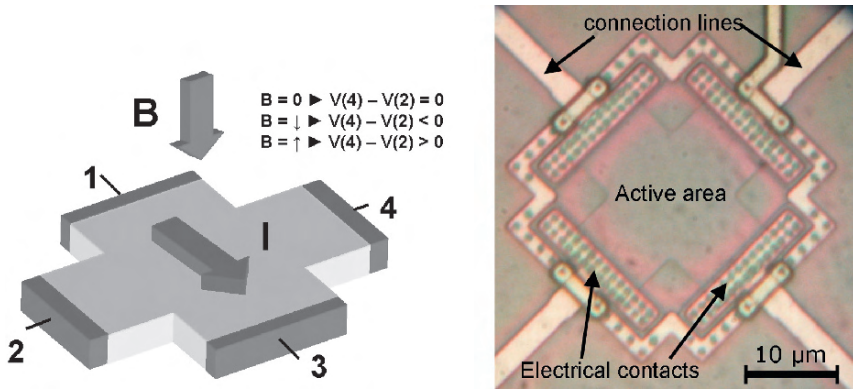


Fig. 1 A Hall sensor outputs a voltage which is proportional to the product of the current running through it and the flux density

The dimensions of a Hall plate in CMOS technology (Fig. 1, *right*) are typically a few 10 to 100 μm featuring a sensitivity between 70 and 300 V/AT . For a biasing current of 1 mA and an applied flux density of 10 mT the sensor yields a signal of a few mV. To amplify this low voltage and to provide for efficient compensation of residual offset and temperature drift, advanced electronics are implemented today onto the same silicon circuit. The final product is a very sophisticated and often application specific Hall sensor.

2 Contactless Magnetic Position Measurement

For contactless position measurement, typically a small magnet is either shifted or rotated in the vicinity of a magnetic field sensor as shown in Fig. 2. Conventional planar integrated Hall elements can only measure the magnetic flux density perpendicular to the chip surface with high precision.

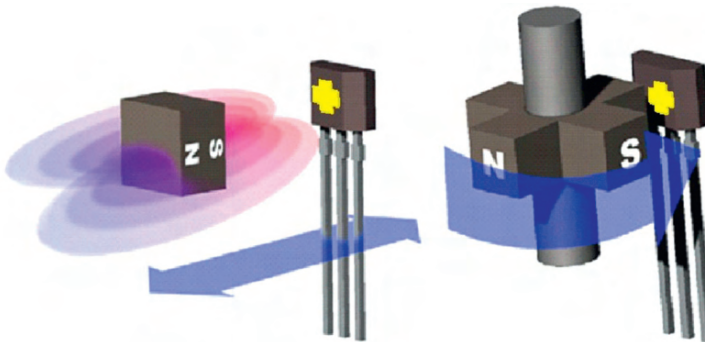


Fig. 2 Single-axis linear position sensor (*left*) and rotation sensor (*right*)

The perpendicular flux density seen by the sensor changes with the motion, and the sensor output signal can be used as position signal. Inherent to such single-axis systems is the problem, that the position information varies over time (t) due to ageing effects and it varies with temperature (T) due to a drift of magnetization of the magnet and due to a drift of sensitivity of the sensor.

$$V_H = S(T, t) \cdot I \cdot B(T, t)$$

with the sensitivity and magnetic flux density given by $S(T, t) = S_0 \cdot f_S(T, t)$ and $B(T, t) = B_0 \cdot f_B(T, t)$.

Therefore a conventional position measurement system must be temperature compensated with sensor and magnet and it must be frequently recalibrated to give a precise signal.

In the following section the principle of multi-axis sensing is explained. Then we illustrate the technology before we describe the magnetic benefits and drawbacks. Finally we point out the technology's potential by giving some application examples for position sensors, electrical current sensors and an electronic compass.

3 Multi-Axis Sensing

The drawbacks of sensor and magnet drift in single-axis position measurement can be virtually eliminated by using the field direction instead of the field amplitude as a measure for the position. It improves the robustness of the measurement on the magnet side as well as on the sensor side. The direction can be described by using the ratio of two non-collinear vector components. If now the drift of those two components with temperature and ageing is very similar, the ratio between them remains a constant value.

Fortunately, the direction distribution in space of the flux density emanating from a permanent magnet is very stable with temperature and ageing, although its strength changes. For the magnet we can then describe the field direction information as the ratio

$$\frac{B_1(T, t)}{B_2(T, t)} = \frac{B_{01} \cdot f_B(T, t)}{B_{02} \cdot f_B(T, t)} = \text{const}(T, t)$$

where B_1 and B_2 can be any two non-collinear vector components of the flux density, for example B_X and B_Y .

Let us now assume that we can manufacture a sensor with several axes of sensitivity which exhibit exactly the same drift over temperature and time. Then we may proceed in exactly the same way and describe the sensitivity direction information as:

$$\frac{S_1(T, t)}{S_2(T, t)} = \frac{S_{01} \cdot f_S(T, t)}{S_{02} \cdot f_S(T, t)} = \text{const}(T, t)$$

Here also, the sensitivity functions S_1 and S_2 describe the sensor's sensitivity along any two non-collinear directions, for example S_X and S_Y .

We can see from those two equations, that for magnetic position measurement through field direction we only need to make sure that the ratio of two functions is constant, not the functions themselves.

Exploiting these relationships in an angular position sensor with two orthogonal measurement axes X and Y and a permanent magnet giving an angle dependent magnetic signal we can now write for the ratio of the two signals

$$\frac{V_Y(T,t)}{V_X(T,t)} = \frac{B_0 \cdot \sin(\varphi) \cdot f_B(T,t) \cdot S_Y \cdot f_S(T,t)}{B_0 \cdot \cos(\varphi) \cdot f_B(T,t) \cdot S_X \cdot f_S(T,t)} = \frac{S_Y \cdot \sin(\varphi)}{S_X \cdot \cos(\varphi)}$$

Hence, by evaluation of the arctan function, an angular position information is generated. This value is not anymore dependent on temperature and ageing.

In the following we present a technology for implementing multi-axis measurement capability in CMOS integrated Hall sensors and how it is used for very accurate and stable angular and linear position measurement.

4 Integrated Magnetic Concentrator

To implement multi-axis capability into a CMOS sensor, the integrated magnetic concentrator (IMC) technology has been developed. It is based on the effect of local field deflection in the vicinity of a ferromagnetic body. The idea of combining such an IMC with silicon Hall devices is over 10 years old [1]. It was first used for precise current sensing and then for angular position sensing [2].

How this field deflection works is shown in the cross-section view of Fig. 3 (*left*) which depicts a CMOS Hall sensor with a ferromagnetic concentrator attached to its surface. A magnetic field which is parallel to the surface of the silicon chip is drawn into the IMC, which results in a local deflection of the field close to the IMC extremities. This is exactly the feature we need to sense the magnetic flux with conventional Hall elements. Since the flux lines go up on one side and down on the other side, we merely have to subtract the output voltages of those two Hall elements to obtain a voltage proportional to the external horizontal flux density.

For a flux density perpendicular to the sensor surface, the Hall chip works just as if no concentrator was present. This is indicated by the small deflection of the flux lines in Fig. 3 (*right*). The IMC is virtually ‘magnetically transparent’, because it is very thin in field direction.

When we use a magnetic concentrator in the shape of a circular disk and place four Hall elements at each 90 degrees angle under its edge, we are now able to measure two orthogonal in-plane magnetic field components at the same time. And moreover by designing the electronics in a way that the outputs of the different Hall elements can be added and subtracted, a three-axis Hall sensor can be implemented (Fig. 4) [3].

A different, but magnetically equivalent structure consists of a Hall sensor with two IMC pieces attached to its surface [4]. Both pieces are separated by an air gap (Fig. 5). Here again, the difference of the Hall elements’ output gives the horizontal flux component and the sum gives the vertical component. An advantage of this structure

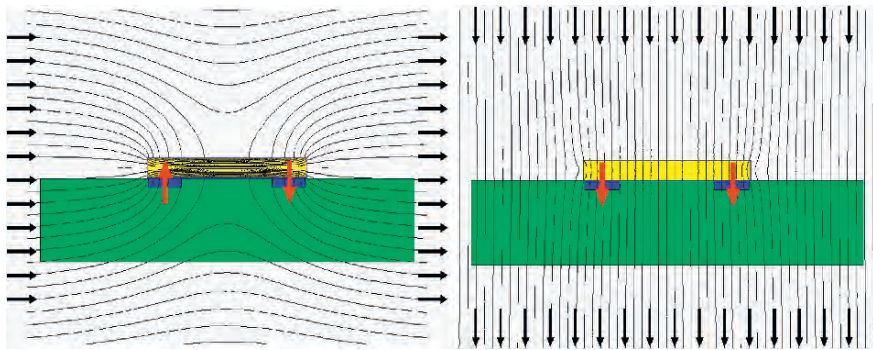


Fig. 3 Cross section simulation of a Hall sensor with two Hall plates and integrated magnetic concentrator for lateral field (*left*). Simulation with flux lines perpendicular to the surface (*right*)

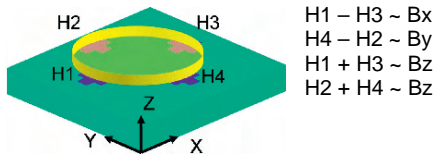


Fig. 4 By combining the four Hall elements differently, the three magnetic field components are measured

is that the IMC can be very big and all Hall elements can still be closely grouped in the sensor center.

We conclude in this section that IMC technology leads to very robust and stable position sensors turning single axis Hall sensors into multi-axis position sensors (Fig. 6). How the IMC is attached and structured onto the surface of the silicon die is discussed in the next section.

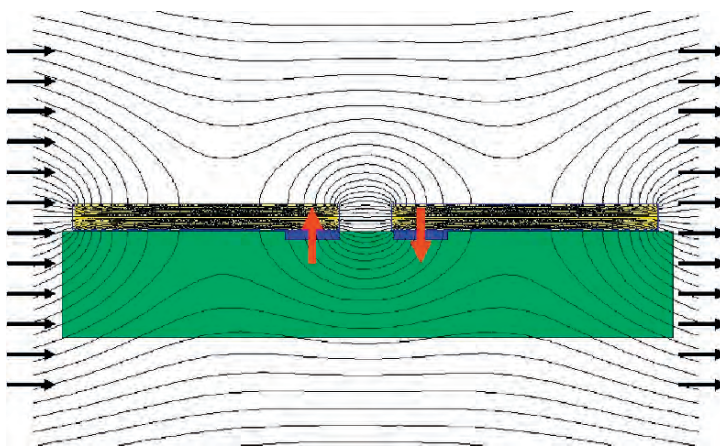


Fig. 5 Simulation of a Hall sensor with two IMC's with a gap in between

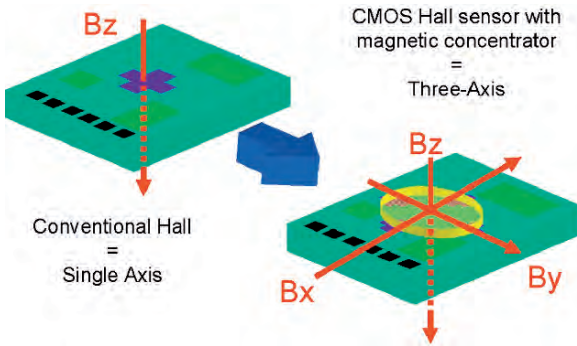


Fig. 6 The integrated magnetic concentrator brings three-axis measurement capability compared to a single axis conventional Hall sensor

5 IMC Process

The manufacturing process is a low cost batch process, so that thousands of sensors are realized in a precise, reliable and low-cost way. First the ferromagnetic layer is bonded to the wafer and then it is structured by photolithographic etching, leaving only the desired structures on the surface. Today many millions of such sensors find their way every year into automotive and industrial applications. With the introduction of IMC technology to modern deep submicron processes, such sensors are now equipped with powerful digital signal processing and interface circuitry making them very versatile for many applications.

The IMC process is a separate photolithographic batch process step between the CMOS microelectronic process and the device packaging (Fig. 7). An entire 8-inch CMOS wafer with several thousand sensors is processed at a time.

It consists of the following steps (Fig. 8):

1. The passivated CMOS wafer with open bond pads comes from the fab
2. a glue layer of a few microns thickness is dispersed onto the wafer
3. the 20 μm thick metal layer is bonded to the wafer and the glue is cured

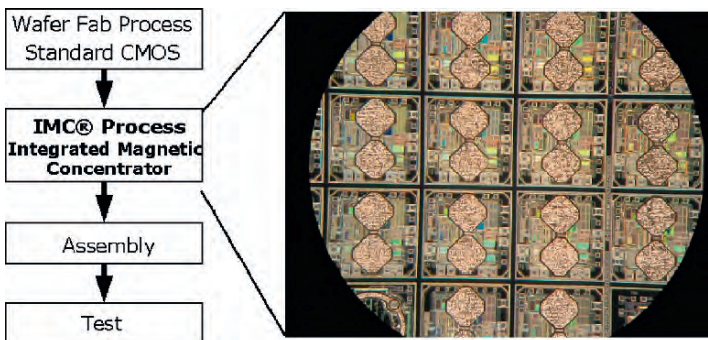


Fig. 7 The IMC process is an independent wafer-level postprocess

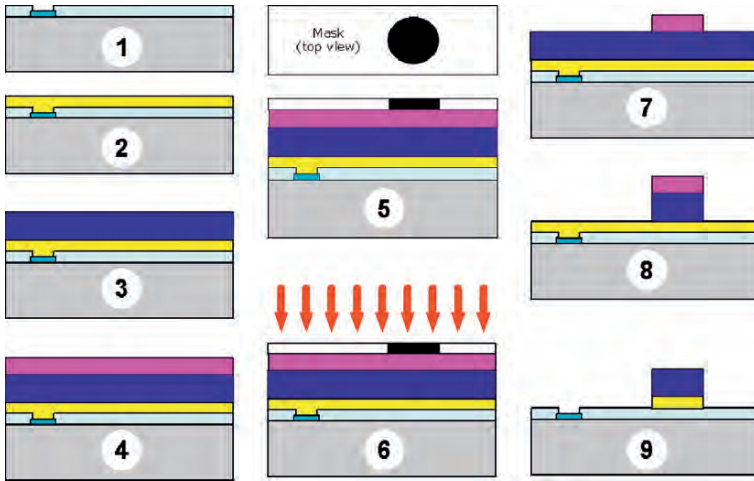


Fig. 8 Process steps of the IMC photolithographic postprocess

4. the photoresist is applied on top of the metal layer
5. the illumination mask is aligned above the wafer
6. the photoresist is illuminated by UV light
7. the photoresist is developed and stays where the metal shall remain
8. the unprotected parts of the metal layer is etched away
9. the remaining photoresist is stripped and the wafer is cleaned.

6 Magnetic and Geometric Properties of the IMC Layer

In the prior chapters we have described the working principle and the technology. In this section we now want to discuss the material and geometric properties of the integrated magnetic layer.

Ideally, an IMC in any application shall comply with the following three requirements:

1. High magnetic gain to increase the Hall output signal.
2. High saturation level to provide for a large linear operating range.
3. Low Hysteresis to reduce errors from the ‘magnetic history’.

The first parameter is almost exclusively related to IMC geometry and the last one to IMC material. The second parameter is related to both, material and geometry.

6.1 IMC Material Properties

The Hysteresis curve (Fig. 9) relates the flux density B to the magnetic Field H for ferromagnetic materials.

It is described by the following equation:

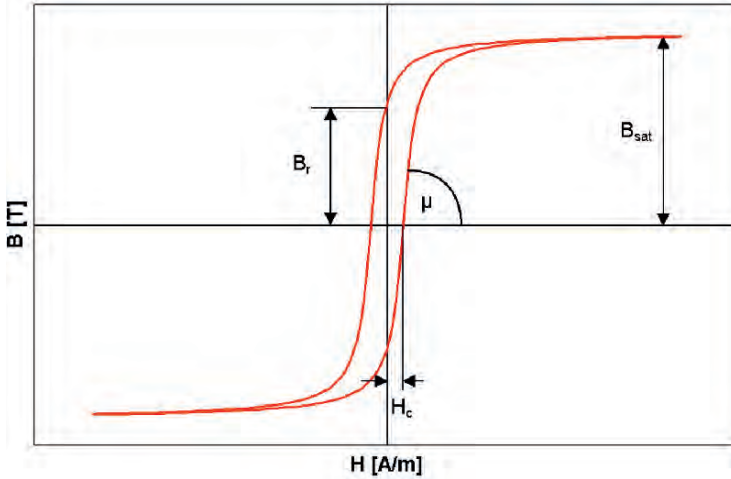


Fig. 9 Hysteresis curve of a ferromagnetic material

$$B = \mu_0 \cdot \mu_r \cdot H + B_r$$

B is the flux density which is actually measured by the Hall element and H is the magnetic field generated by the permanent magnet or electrical current. B_r is the remnant magnetization of the material, the value of which depends on the magnetic history. As we can recognize from the hysteresis curve, the relationship is not ideally linear and when using such material we are confronted to four phenomena:

- Flux density amplification: μ_r is typically much larger than 1.
- Linearity of the output signal: μ_r is typically not constant but changes with the magnetic field H .
- Saturation: For strong magnetic field, B approaches a constant level, the so-called saturation magnetization. μ_r approaches the value of 1.
- Hysteresis: Without external field, the material remains magnetized with the remnant flux density B_r and only becomes demagnetized if a coercive field $-H_c$ is applied.

A good IMC material therefore must have a small remnant magnetic flux density and a high saturation level. Those conditions are ideally met in amorphous iron nickel or iron cobalt alloys. Typical values are given in the following table.

Relative permeability μ_r	1500 ... 100000
Saturation flux density B_{sat}	0.5 .. 0.8 T
Remnant flux density B_r	10 μ T
Material thickness	20 μ m

The material thickness of approx. 20 μ m is given by the production procedure of amorphous metals and fits well for integration on CMOS circuits. Those values are very good, but for some applications not yet good enough as we will see later.

6.2 IMC Geometric Properties

Due to the high permeability of the IMC material, the flux density in the vicinity of the IMC edge can be significantly increased compared to the applied external field. In such a way the IMC works as passive magnetic amplifier of up to a factor of about 10 which directly impacts the signal-to-noise and signal-to-offset ratios. The achievable magnetic gain however not only depends on the length of the IMC structure, but also on its shape. As shown in Fig. 10, the gain of an ideally ‘short’ disk structure increase with about the square root of the disk diameter, where as the gain of a long and narrow rod is proportional to its length. Intermediate geometries as well as combinations of several of the above structures feature a characteristic in between those two.

For angular sensors a circular disk shape IMC is most appropriate so that properties are equal for any arbitrary in-plane angle of the magnetic flux density. A disk of 200 μm diameter with a magnetic gain of about 1.5 is a good match.

Current sensors have a single sensitivity direction perpendicular to the current flow. Here two IMC pieces are combined with a small air gap and a total length of 1.5 to 2 mm, which brings a magnetic gain of about 8.

The electronic compass contains a series of ring shapes which lead at a size of about 2 mm to a gain of 6.5.

Application	Magnetic gain	Nominal flux density	Remnant flux error ¹
Angular sensing	1.5	20 mT...80 mT	0.05%...0.01%
Current sensing	8	1 mT...20 mT	1%...0.05%
eCompass	6.5	10 μT...1 mT ²	1%...0.05% ³

¹ A remnant flux density of 10 μT is assumed.

² A high working field range is required due to parasitic magnetic fields.

³ With the use of an integrated demagnetization circuit.

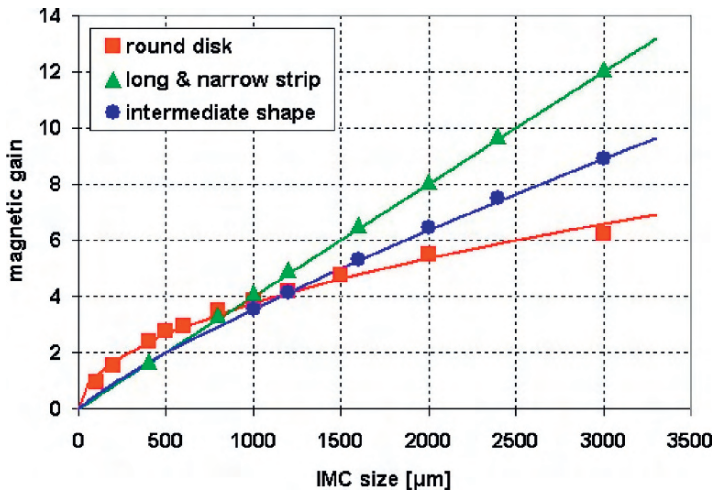


Fig. 10 Experimental gain values show that the magnetic gain increases with the size of the IMC in field direction

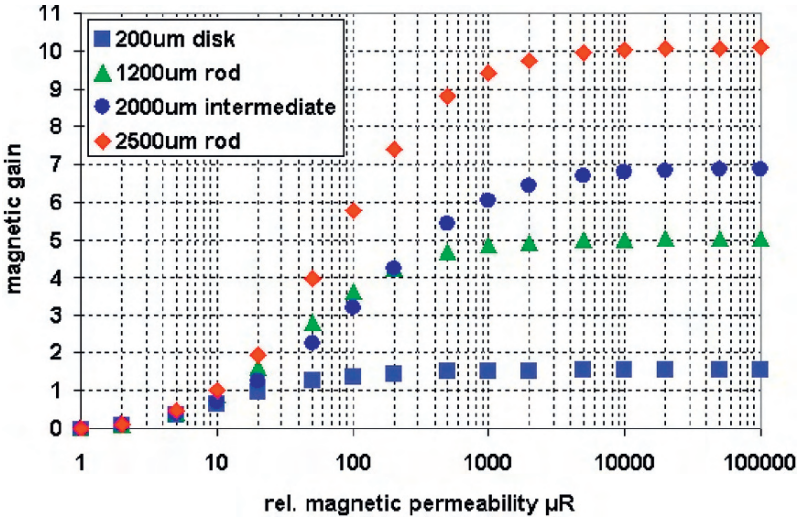


Fig. 11 Magnetic gain of different IMC structures depends on the relative permeability of the IMC material

Naturally, a high magnetic gain requires a high magnetic permeability of the IMC material (steep slope in the hysteresis curve). But due to the open magnetic circuit, for a relative permeability of approximately 5000 the gain saturation is reached for all investigated geometries. Thus the magnetic gain does not increase anymore for $\mu_R > 5000$ (Fig. 11).

The interested reader may find some more background information on IMC design in [5].

7 Saturation

For a material with high magnetic permeability the flux lines enter the IMC perpendicularly. If the magnetic field is increased the IMC starts to saturate and the permeability decreases. In consequence the flux density is reduced and the flux lines do not enter the IMC orthogonally anymore. Both effects reduce magnetic gain (Fig. 12).

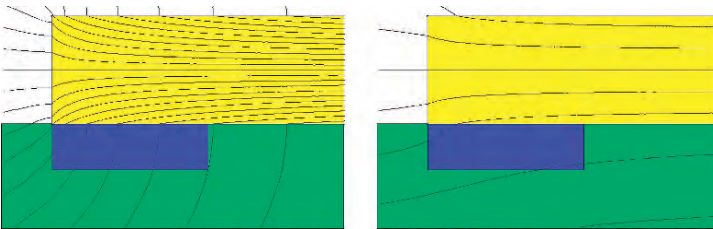


Fig. 12 Unsaturated IMC (left) and saturated IMC (right)

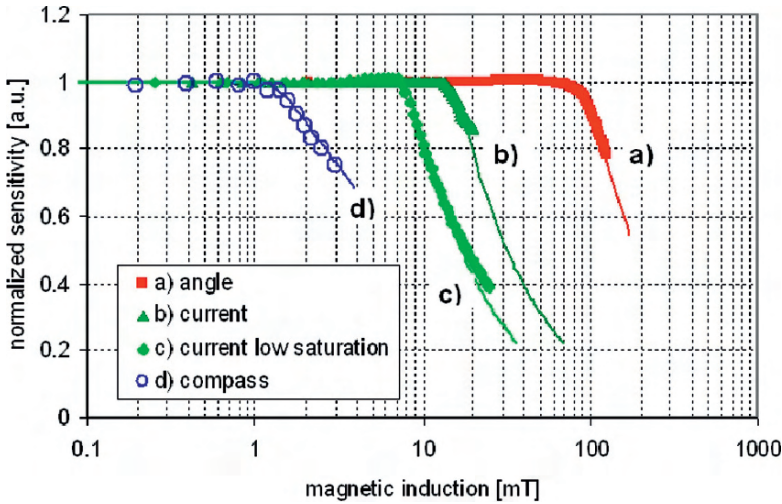


Fig. 13 Saturation characteristic of the shapes used in various applications. Besides the size and geometry, the material plays here also a role (see current sensor)

From an application point of view the linearity of the fullscale output signal is critical (see Fig. 13) and thus the IMC must not reach magnetic saturation for the specified full-scale range of magnetic flux density. Or in other words the magnetic gain must remain constant for the full range of the magnetic excitation field.

A more detailed study of saturation non-linear effects can be found in [6]

8 Hysteresis

Exposing the IMC sensor to a strong magnetic field by a permanent magnet leads to a material related remnant magnetization of the IMC after the magnet is removed (see Fig. 15 and Fig. 9). The remnant magnetization appears as magnetic signal, thus it is sensed by the Hall elements and therefore appears as an arbitrary offset voltage in the output signal.

Especially for electronic compass applications, where the full scale field is very small (about 20 μT), such magnetic offset of 10 μT is unacceptable, since it results in angle error of several 10° . To reduce remnant magnetization, a ‘degaussing’ procedure is integrated in the ASIC (Fig. 14).

The IMC is exposed to a strong magnetic field (a) which leads to a remnant magnetization and an arbitrary offset in the output signal (b). A current flowing through the center of the ring shape IMC will magnetize the IMC in a circular manner (c) and leave a circular magnetization that is not sensed by the Hall elements.

For various strengths of field exposure a remnant magnetization (perming) of up to 2.5 μT is left in the IMC. Through ‘degaussing’ the perming can be reduced to 0.1 μT (Fig. 15).

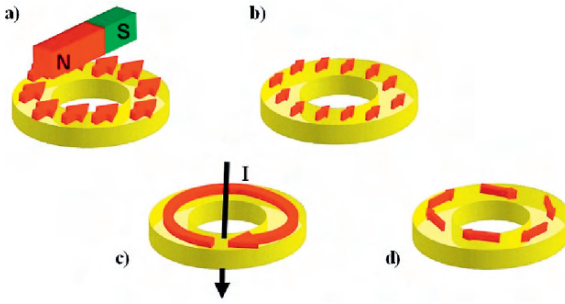


Fig. 14 Degaussing procedure of the eCompass sensor to remove a remnant magnetization of the IMC

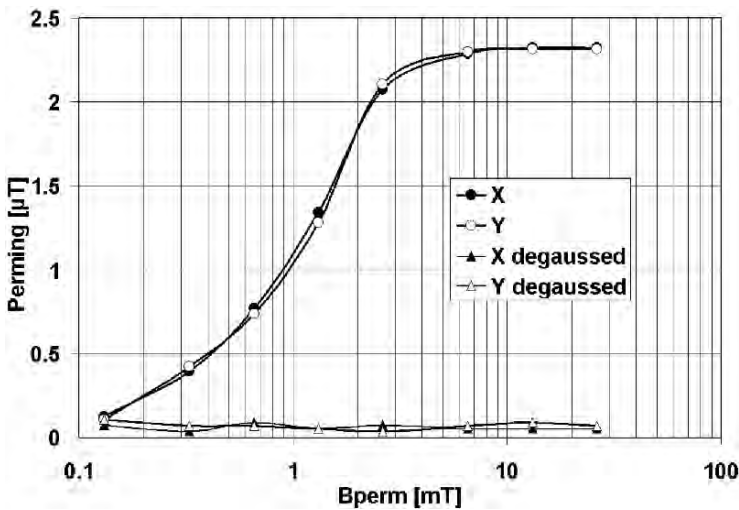


Fig. 15 Perming of the eCompass IMC for various strength of field exposure; with and without use of the degaussing circuit

9 Sensor Architecture

The sensor architecture comprises the Hall elements and also all necessary digital and analog electronic circuit parts which are needed for generating selected useful information from the applied magnetic field. The architecture diagram for a 3-axis IMC Hall sensor is given in Fig. 16.

The output voltage of a Hall device is typically in the order of microvolts to millivolts and must be amplified before being transmitted to the outside world. Typical amplification of a factor of 1000 is performed by a chain of amplifiers which are directly co-integrated on the same silicon where the Hall elements are. A special technique

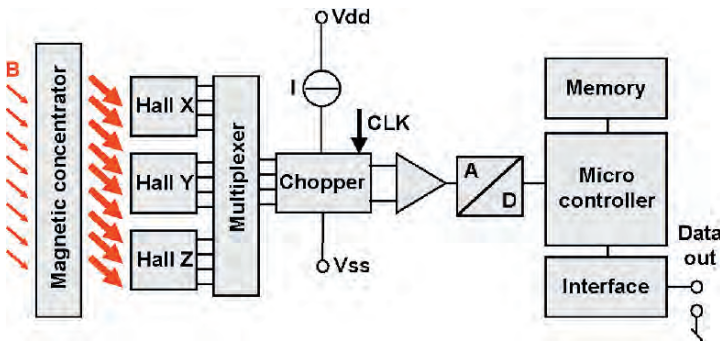


Fig. 16 Full architecture diagram of an IMC 3-axis sensor with microcontroller

called spinning current operation [7] allows at the same time to reduce inherent sensor offset considerably.

All circuitry for magnetic field sensing, amplification, analog and digital signal treatment and data storage are then designed onto a single piece of silicon (Fig. 17). The whole wafer with many thousands of those chips is then IMC post processed as described in Fig. 7 and finally the wafer is sawn and the sensors are packaged.

Some typical applications as we meet them in many industrial, automotive and consumer applications are illustrated in the following section. The attentive reader may recognize how the described technology of integrated magnetic field concentrator brings a strong enhancement of the sensor performance. At first we will show three position sensor applications where mainly the multi-axis sensing feature of the IMC technology is used. Then we show how the high magnetic gain allows for the implementation of a high-quality current sensor. And finally we demonstrate the advantage of combining both key features within an electronic compass.

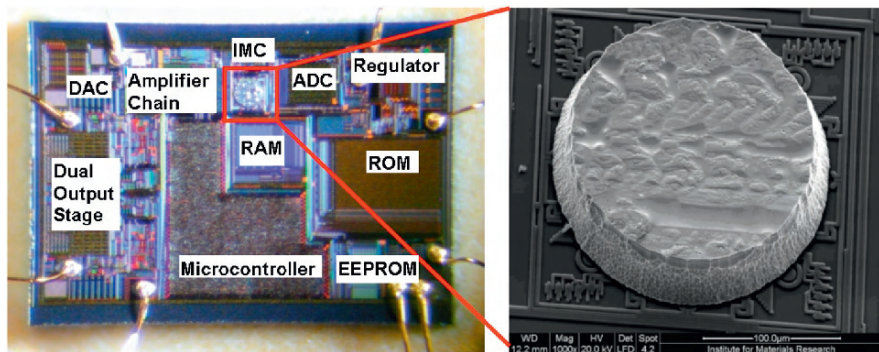


Fig. 17 Die photograph of the 3-axis sensor of $1.8 \times 2.8 \text{ mm}^2$ dimension (left) and electron-microscope picture of the IMC of $200 \mu\text{m}$ diameter (right)

10 Angular Position Sensor

A very straight forward application of a two-axis magnetic sensor is the angular position measurement (Fig. 18). In this case a small magnet is mounted on a rotating axis above the sensor. The magnetization of the magnet is perpendicular to the axis and parallel with the sensor plane. With a rotation of the axis the flux density vector rotates in the sensor plane and the output signal of the two sensor axes X and Y yield a sine and cosine signal with the rotation.

The analog sine and cosine signals are first converted into the digital domain. Then the inverse tangent function of their ratio is computed by a digital algorithm (Fig. 19).

This output represents the position angle of the magnet on the rotating axis with a typical accuracy of better than 0.3° over a wide temperature range. By more advanced programming the sensor can calculate functions from this angle, so that application specific angle signals can be generated.

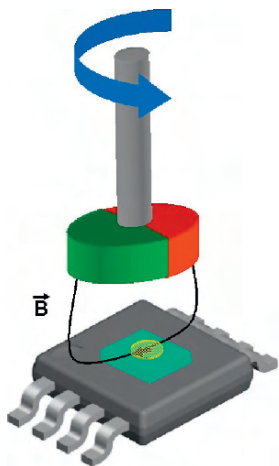


Fig. 18 Magnetic Contactless Angle Sensor

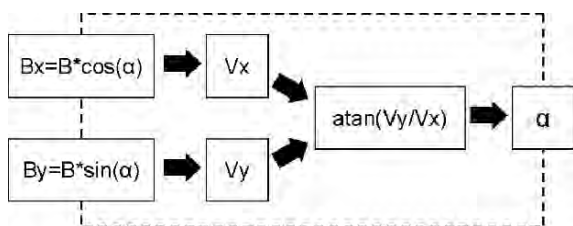


Fig. 19 Computation of the angular position from the two magnetic flux components B_x and B_y

11 Linear Position Sensor

For linear position measurement, the two-axis Hall sensor is combined with a magnet which is now magnetized orthogonally to the sensor plane Fig. 20. The magnet is now shifted parallel to one axis at a certain distance above and sideways of the sensor. This again leads to a sine a co-sine reading of the direction of the field angle. The ratio of both values is directly proportional to the magnet position.

The computation of the position output is very simple for this application since it contains merely a division between the two Hall voltages along the two measurement axes (Fig. 21). The practical working range is typically about two to three times the diameter of the used magnet. The accuracy is about 0.5% of the working range and the system is very robust versus temperature changes.

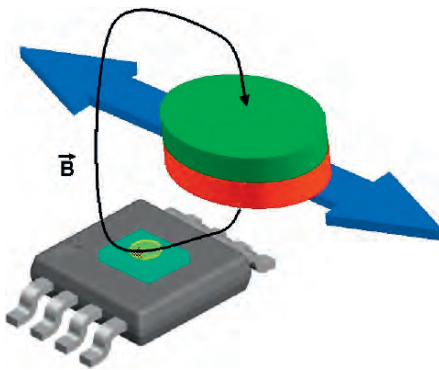


Fig. 20 Magnetic Contactless Linear Position Sensor

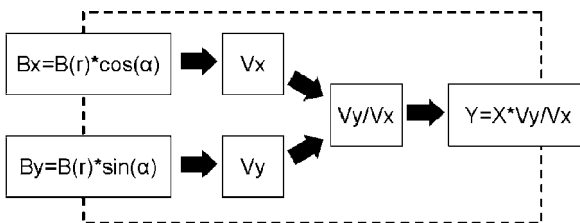


Fig. 21 Computation of the linear position from the two magnetic flux density components B_x and B_y

12 Contactless Joystick Sensor

By using all three magnetic field sensitive axes, a contactless magnetic joystick can be implemented (Fig. 22). The two tilt angles in the XZ and YZ planes can be computed very accurately by using all three magnetic field components as shown in Fig. 23. Here again, the position information is derived from the ratio of magnetic field components, so that temperature and ageing drift of magnet and sensor are cancelled out.

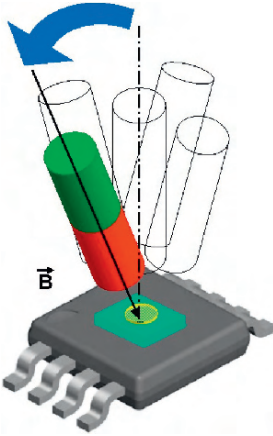


Fig. 22 Contactless magnetic joystick

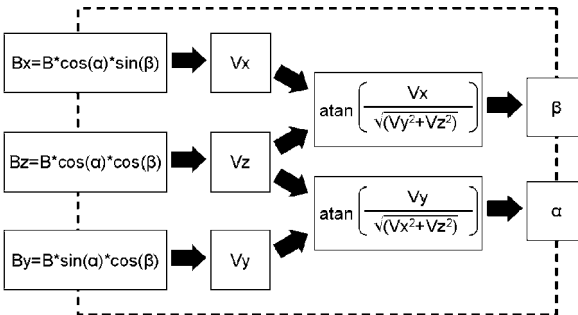


Fig. 23 Computation of the two tilt angles α and β from the three magnetic flux density components B_x , B_y and B_z

13 Electrical Current Sensor

The electrical current flowing through a wire, busbar or PCB track leads to a magnetic field around the wire which can be sensed by a single axis IMC current sensor (Fig. 24,

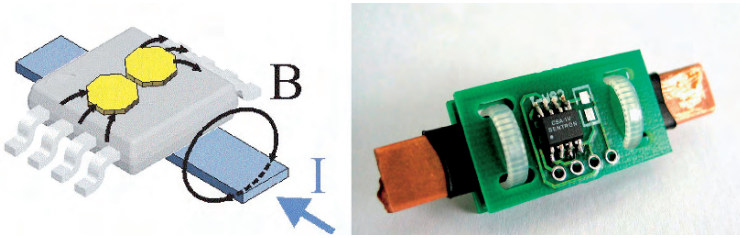


Fig. 24 single axis IMC current sensor working principle (*left*); IMC current sensor on a copper busbar (*right*)

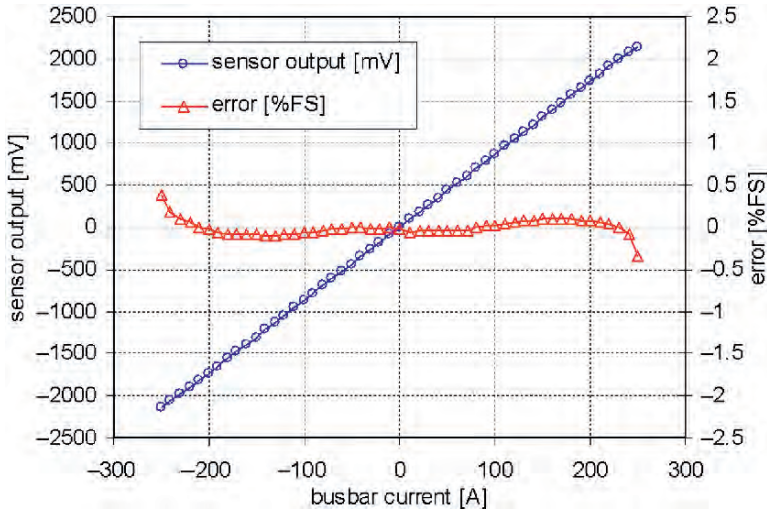


Fig. 25 Measurement of an IMC current sensor on busbar for a current of ± 200 A. The linearity error is smaller than 0.5%

left). This application takes primarily advantage from the high magnetic gain of IMC technology.

In order to be able to measure currents up to several hundred Amps the IMC current sensor is attached to a cable or mounted on a copper busbar (Fig. 24, *right*). In this particular case an IMC current sensor is designed for up to ± 200 A of busbar current leading to an output signal of ± 1.75 V and an error of less than 0.2% FS (Fig. 25).

14 Electronic Compass

The most challenging application for IMC technology is the electronic compass which takes maximum advantage from multi-axis capability AND magnetic gain [8, 9]. The earth's flux density is about 100 times smaller than in a typical current sensor application and 1000 times smaller than in a typical position sensor application. Besides a

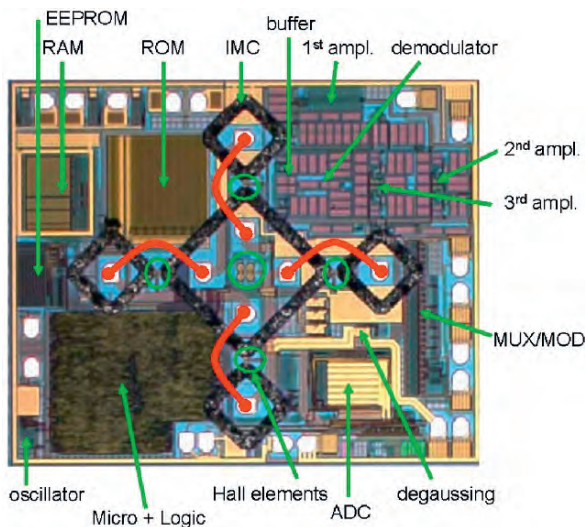


Fig. 26 Die photograph of the electronic compass with its various electronic blocks. The degaussing loop bonding wires are connecting the bonding pads in the center of the external IMC octagon structures with those of the central octagon

high magnetic gain we therefore also need a high electronic gain of over 1000. Additionally at least two orthogonal components of the magnetic flux must be measured to determine the direction. Figure 26 shows a photograph of the silicon die with about 7 mm^2 surface area. The circuit contains the analog and digital electronics for amplification, biasing, filtering and ADC conversion as well as a 16-Bit microcontroller with memory on the same chip.

It features an angular accuracy of $\pm 2^\circ$ and is packaged after manufacturing in a very small plastic package of $5 \text{ mm} \times 5 \text{ mm} \times 1 \text{ mm}$ dimension.

References

1. H. Blanchard, L. Chiesi, R. Racz, R.S. Popovic, "Cylindrical Hall Device", Proc. of International Electron Devices Meeting, IEDM 1996, 8–11 Dec. 1996, pp. 541–544.
2. R.S. Popovic, C. Schott, P.M. Drljaca, R. Racz, "A new CMOS Hall angular position sensor", *Technisches Messen*, tm Vol. 6, June 2001, pp. 286–291.
3. C. Schott, R. Racz, S. Huber, "CMOS three axis Hall sensor and joystick application", Proceedings of IEEE Sensors, Vienna, Austria, 24–27 Oct. 2004. Vol. 2, pp. 977–980.
4. H. Blanchard, F. De Montmollin, J. Hubin and R. S. Popovic, "Highly sensitive Hall sensor in CMOS technology", *Sensors and Actuators A: Physical*, Vol. 82, Issues 1–3, 15 May 2000, pp. 144–148.
5. P. Drljaca, F. Vincent, P.-A. Besse and R. S. Popovic "Design of planar magnetic concentrators for high sensitivity Hall Devices", *Sensors and Actuators A: Physical*, Vol. 97–98, 1 April 2002, pp. 10–14.
6. P. Drljaca, M. Demierre, C. Schott, R.S. Popovic, "Nonlinear Effects in magnetic angular position sensor with integrated Flux Concentrator", Proc. of 23rd International Conference on Microelectronics (MIEL 2002), Vol. 1, Nis, Yugoslavia, 12–15 May, 2002.

7. P.J.A. Munter, "A low-offset spinning-current Hall plate", *Sensors and Actuators A* 21–23 (1990) 743–746.
8. R. Racz, C. Schott, S. Huber, "Electronic Compass Sensor", *Proc. of IEEE Sensors*, 24–27 Oct. 2004, Vol. 3, pp. 1446–1449.
9. C. Schott, R. Racz, A. Manco, N. Simonne, "CMOS Single-Chip Electronic Compass With Microcontroller", *IEEE Journal of Solid-State Circuits*, Vol. 42, Issue 12, Dec. 2007, pp. 2923–2933.

Commercial Magnetic Sensors (Hall and Anisotropic Magnetoresistors)

Michael J. Haji-Sheikh

1 Introduction

Magnetic sensors account for a significant portion of the sensing market. Manufacturers such as Honeywell, Phillips, Optek, Cherry, and Infineon primarily make commercial and automotive sensors while Fujitsu, IBM, Maxtor and Seagate control the information sector additionally Asahi Chemicals has a significant position in fan speed sensing. There are two types of magnetic sensors. The first type of magnetic sensor commonly used is the Hall-effect sensor. The Hall-effect sensor is a device that depends on the mobility of carriers in a semiconductor material such as silicon. The second type of magnetic sensor is the magnetoresistor. Magnetoresistors come in different types. These are ordinary magnetoresistors (MR), anisotropic magnetoresistors (AMR), giant magnetoresistors (GMR) and colossal magnetoresistors (CMR). Of these, the CMR has yet to move out of the research phase. These commercial devices, manufactured by the above companies, have diverse applications such as proximity sensors, gear-tooth sensors, and read head sensors. An example of how important the magnetic sensor is, a search of the United States Patent Data Base shows over three thousand patents using hall-elements. Additionally there are over four hundred patents using magnetoresistors. There is an art and a science to building commercial sensors. Often it takes a diverse group of scientists and engineers to characterize and model these sensors prior to committing a design to production. This is due to the secondary nature of the sensing mechanisms which are commonly used. Most sensing mechanisms only show up as small perturbations in a larger property such as resistivity or permittivity. Often these perturbations are only a few percent of a typical full scale output for a given sensor. The difficulty in using these properties as sensing mechanisms is that they often interact with their surroundings in ways which obscure the measurements of interest. What is meant with this statement is that the structures often use to mount the sensors will have just as much influence as the external signal as the output of the sensor.

Michael J. Haji-Sheikh

Associate Professor of Electrical Engineering, College of Engineering and Engineering Technology, Northern Illinois University, De Kalb Illinois 60115. e-mail: mhsheikh@ceet.niu.edu

Nanotechnology related advances will effect how we view and develop a new generation of sensors. The interaction of magnetic materials and the patterning of these materials will eventually lead to devices that we have not yet conceived. Most magnetic thin film devices are nanoscale in the thickness but newer research areas include nanoscale patterning and nanoscale self-assembly.

2 Hall Sensor Design

The design of Hall-effect sensors has been well detailed in Popovic's book [1] so we will only lightly treat the physics and concentrate on the procedures required to build a production sensor. The design of Hall-effect sensors is an exercise in geometry and device physics. The typical Hall-effect sensor, or Hall cell, is built into a lightly doped n-type epitaxial layer due to the high electron mobility. The classical mechanics of the Hall-effect depends on the Lorentz force. The Lorentz force states that there is a force placed by a magnetic field onto a moving charged particle. The Lorentz equation is

$$\vec{F}_m = -q(\vec{v} \times \vec{B}), \quad (1)$$

where F_m is the Lorentz force, q is the charge on an electron, v is the electron velocity, and B is the external magnetic field. A simple schematic of a Hall cell is shown in Fig. 1. The magnetic field is normal to the top surface. The electric field set up by this external magnetic field. There is a counter-balancing electric field which is

$$\vec{E}_h = -\vec{v} \times \vec{B} \quad (2)$$

set up when the magnetic field is applied.

The hall voltage set up at the side taps is

$$V_h = \int_m^n \vec{E}_H \cdot d\vec{z} \quad (3)$$

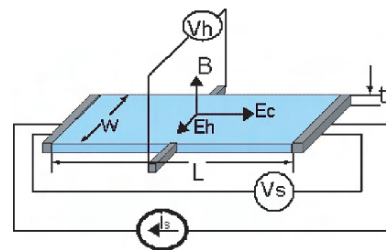


Fig. 1 Schematic of a rectangular Hall-effect element

and as defined in Popovic [1]

$$V_{hp} = \mu_p E_x B_y w \tag{4}$$

and

$$V_{hn} = \mu_n E_x B_y w \tag{5}$$

for the p -material and the n -material respectively. Additionally, the Hall electric field is

$$\vec{E}_H = -R_H(\vec{J} \times \vec{B}) \tag{6}$$

where the Hall Coefficient R_H for p -material and n -material is

$$R_{Hp} = 1/qp \quad R_{Hn} = 1/qn \tag{7}$$

where p is the number of p -carriers per cubic centimeter in the semiconductor and n is the number of n -carriers per cubic centimeter in the semiconductor. The hall voltage can be rewritten

$$V_{hp} = \frac{R_H}{t} I B_{\perp} \tag{8}$$

where I is the source current, t is the Hall cell thickness and B is the perpendicular magnetic field. For finite contact Hall cells

$$V_{hp} = G \frac{R_H}{t} I B_{\perp} \tag{9}$$

where G is a geometrical correction factor. The most common design for the for commercial Hall-effect sensor is euphemistically called the ‘band-aid’ contact Hall. Fig. 2 is a schematic of a ‘band-aid’ device. The name ‘band-aid’ comes from the resemblance of the contact to the medical gauze.

To effectively manufacture and calibrate a Hall-effect sensor, it requires an integrated manufacturing concept. These sensors required to have a high degree of manufacturing accuracy and a high degree measurement accuracy. This is accuracy

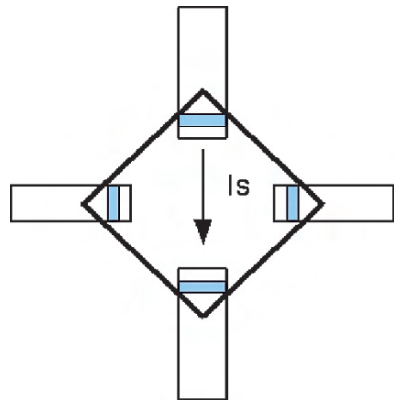


Fig. 2 The ‘band-aid’ hall cell. It comprises of a n -epitaxial area with a high concentration n -contact area and metal interconnect. This cell is demonstrated without a field plate

is important, since many of the applications are critical for system functionality. An important issue for the Hall-effect sensor is the effect of piezoresistance on the offsets. There are many methodologies for the balancing out the stress effects on the Hall-effect sensor. These stress balancing methods must be applied in the concept stage. These range from using (100) silicon to using dual and quad Hall cells. An important resource for research of the Hall-element is the United States Patent Office database. This resource is free and is not only able to be searched by the www.uspto.gov search engine but also is open to searches by popular commercial search engines.

One of the more recent innovations in Hall devices is the use of field plates to adjust offsets. Cohen [2] extends the patent by Plagens [3], which proposes to use metal or polygates placed in critical positions on the field oxide. These gates modulate the resistivity at the surface of the Hall-element to throw in a counter-offset.

Alexander's [4] patent uses a temperature compensated power supply to hold the offset constant which is a more standard method in sensing. Steiner et al. [5] looks at a more novel method of Hall-element design. The first 3 patents [2, 3] deal with traditional rectangular hall elements whereas Steiner deals with a circular element and spins the current in a circle to subtract the offset.

The traditional way to correct offsets is the addition of Hall-elements with currents running in multiple paths. The dual Hall cell has current running at ninety degrees to each other to neutralize the stress offsets due to piezoresistance effects. This though, increases the current required by a factor of two. The quad Hall shown in Fig. 3 not only balances out the stress induced offsets but additionally removes alignment induced offsets. The unfortunate side effect is that the quad hall cell has four times the current of the single hall cell.

The location of the Hall cell in the chip i.e. whether the sensor is on the edge of the chip versus the center can also effect the signal to mechanical noise ratio. With the introduction of digital technologies, the use of CMOS switches allows for the electrical rotation of the supply and sense leads to minimize the stress effects. The switches used for these type of sensors have to have a significantly large enough area so as to minimize the on-resistance. The Steiner patent [5] is representative of the offset adjustment methodology. Another invention that has affected how Hall-effect sensors are designed is the chopped Hall sensor which was invented by Bilotti [7] as shown in Fig. 5. Unlike the Steiner patent, the current is not rotated to all four possible positions but to two positions, ninety degrees apart.

To minimize packaging stresses, not only should the layout of the chip be considered but the layout of the package and chip both should be considered for a complete design pictures. The total stresses of the mounting on the die attach flag plus the over molding will effect the offsets of the Hall-effect sensor. As stated earlier, the offsets generated by the packaging stresses cannot be separated from the signal of interest. There are subtle effects of die coatings which can also effect stress induced offsets due to over molding. The proposed method of designing a Hall-effect

sensor, or any sensor, is to combine modeling with experimental verification. It is important to test and calibrate any model using some form of test structures. This calibration will pay off in the long run since it allows the designer to perform software designs of experiment (or DOE). These software experiments allow for a rapid minimization of the design options and design cycles. Due to the fact that Hall-effect devices are generally in *n*-type silicon (due to the significantly higher mobility), the piezoresistance coefficients of *n*-type silicon need to be obtained. The definition of piezoresistance is the change in the resistance of a material with an induced strain. The work of Matsuda et al. [8] details the piezoresistance mechanisms and coefficients for *n*-type silicon. The tensor relationship using both first and second order effects are

$$\frac{\partial \rho}{\rho_o} = \sum_j \pi_{ij} T_j + \sum_{j,k} \pi_{ijk} T_j T_k \tag{10}$$

where *T* is the stress $\partial\rho/\rho_o$ is the normalized resistance and π_{ij} is the first order piezoresistance coefficient and π_{ijk} is the second order coefficient. This equation in combination with a finite element analysis program such as ANSYS will allow for the calculation of voltage offsets generated by packaging stresses. Additionally a detailed graphical analysis was done by Kanda [9] in 1982. Once the best possible candidates for a particular application are chosen, then a test structure layout is made using the modeled design. The tests structures are then manufactured in the particular technology i.e. CMOS or Bipolar. After the silicon is finished, the measurement and evaluation starts.

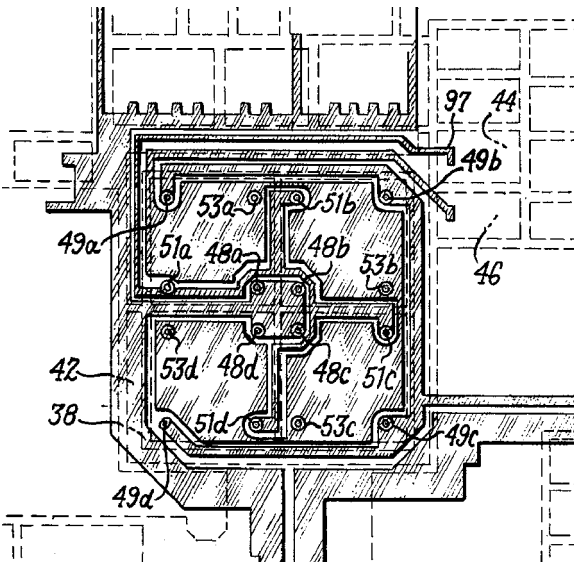


Fig. 3 Quad Hall cell with the supply voltage in the center of the cell. This patented method of quad-hall cell was invented by Higgs and Humenick [6]

3 AMR (Anisotropic Magneto-Resistive) Sensors

There are many different magnetic field sensing applications. Unfortunately, Hall sensors cannot cover all these applications. Magnetoresistive sensors for these applications offer higher sensitivities and superior performance. These applications range from engine position sensing to hard drive read heads. Newer AMR read heads are often combined with GMR sensors to help increase the aerial density. Additional applications range from proximity sensing to wheel speed detection.

There are many inventions based on magnetoresistivity. One of the first sensors in the U.S. Patent files is Nepela and Potter [10] and Lee's [11]. Potter [12] is also the author of one of the first comprehensive papers on the magnetoresistive effect in the general literature. These devices used the properties inherent in magnetoresistance to read recorded data from a magnetic storage media. This concept moved from the analog recording industry to data storage over the following years. A picture of the Lee's [11] invention is shown in Fig. 4. This sensor requires a field of ninety degrees to the direction of the serpentine section in the center.

The history of the magnetic recording using AMR sensors can be shown in the history of applicable inventions [13, 14, 15, 16, 17, 18]. Recent work in the recording and read head area has focused more on GMR than AMR. Industrial and automotive applications for AMR sensors range from proximity sensors to rotational speed sensors. Initially, AMR sensors were used as sensitive magnetometers as shown by Paul et al. [19], but now have been shown to be much more useful and flexible. One application that has recently come into its own is the magnetic rotary encoder.

One of the first recorded invention of this type is the rotary encoder invented by Ito et al. [20] and shown in shown in Fig. 5. This concept is expanded on by Haji-Sheikh et al. [21] to determine both speed and direction and is shown in Fig 6. Previous approaches for AMR models approach magnetoresistance in a relatively piecewise manner [22]. The normal procedure for design of magnetoresistors outlined in Tummanski [23] have not changed significantly since the mid-eighties and can be difficult to use. The design equations require fixing the field angle to resistor direction which means that the user must have multiple design equations. There are no above saturation models for sensor design, until recently, that accurately model a sensor element. The method presented here is an approach which

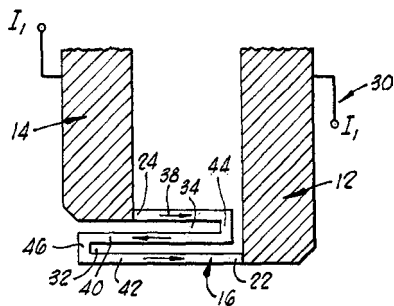


Fig. 4 Early invention showing the primary sensing element in F. Lee's [11] patent

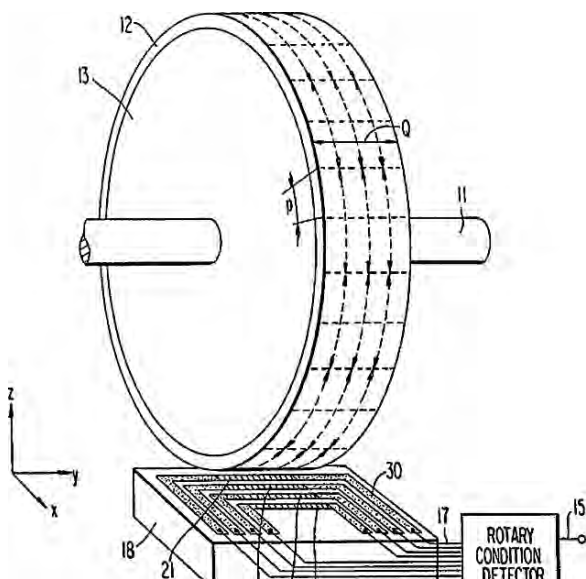


Fig. 5 Magnetic rotary encoder schematic from Ito et al. [20]

has been successfully used for applications ranging from a high current sensor to wheel-speed sensors which use saturation mode magnetoresistors.

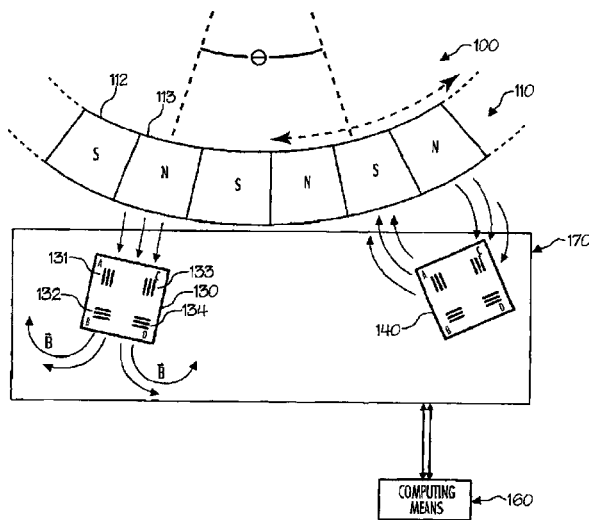


Fig. 6 Magnetic encoder using two sensors to be able to determine speed and direction of a ring magnet from Haji-Sheikh et al. [21]

4 AMR Model

AMR sensors can be used in two basic modes i.e. above magnetic saturation and below magnetic saturation. The curve shown in Fig. 7 shows the response of an anisotropic magnetic resistor with a transverse magnetic field. This figure breaks up the model into the above saturation value and the below saturation value. To understand the total behavior of the magnetoresistor, it is important to understand the behavior of the magnetoresistor in saturation. Since the above saturation resistivity is purely angle dependent, this model should be done first and then used to model sub-saturation behavior. To generate model information, test samples of Kelvin connected magnetoresistors need to be made at specified widths and thicknesses so as to sample the possible design space. Figure 8 is a schematic of a Kelvin connected resistor with the current in the outer connections and the voltage measure in the inside connections. To characterize the AMR film, it is important to understand some of the material properties of the film itself. The most common method of production of both anisotropic magnetoresistive films is by the use of plasma deposited materials. The films used to make the AMR sensors are oriented generally on the (111) plane and can be modeled as a single domain film.

To evaluate the crystallite orientation, samples were taken to Argonne National Laboratories Advanced Photon Source. Figures 9 and 10 are the results of zone plate measurements. The results show that the initial permalloy deposition aligns with the TaN (111) then the NiFe (111) increases in intensity with thickness. This increasing intensity demonstrates that the permalloy films are strongly (111) oriented. This matches data generated by Yeh [24] where he demonstrates the strength of the film orientation depending on the seeding layer.

The films deposited by Yeh [24] on SiO_2 were weakly oriented whereas the films deposited on Si_3N_4 and TaN were strongly oriented. The exact mechanisms of magnetoresistance is well beyond the scope of this analysis though an interesting paper

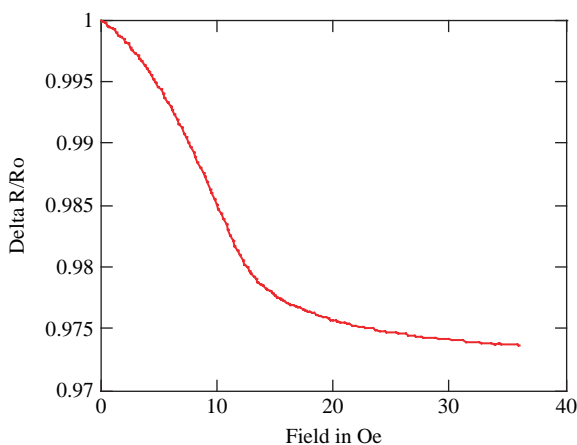


Fig. 7 The transverse magnetoresistance curve for a 37.5 nm thick magnetoresistor with a 35 μm wide resistor

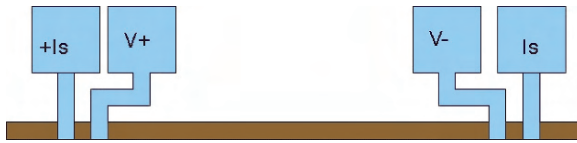


Fig. 8 Kelvin connected magnetoresistor

by Berger [25] was performed to determine the mechanisms of magnetoresistance. Berger’s experiment analyzed the saturation value of magnetoresistance as it is related to crystal orientation. Berger’s experiment took three single crystal nickel resistors (100,110,111) and oriented them so that the resistor was ninety degrees to the magnetic field direction. Notably the (111) oriented single crystal nickel sample showed no magnetoresistance variation as the sample was rotated so that the maximum change stayed constant. Berger assumed that his measurement were not accurate enough to sense the resistance variation in the six fold symmetry plane. This could explain some of the single domain behavior of the permalloy film.

Historically, all analysis of magnetoresistance has started with the Voigt-Thompson equation [25, 26]

$$\frac{\Delta R}{R_o} = \frac{\Delta R_{max}}{R_o} \cos^2 \theta. \tag{11}$$

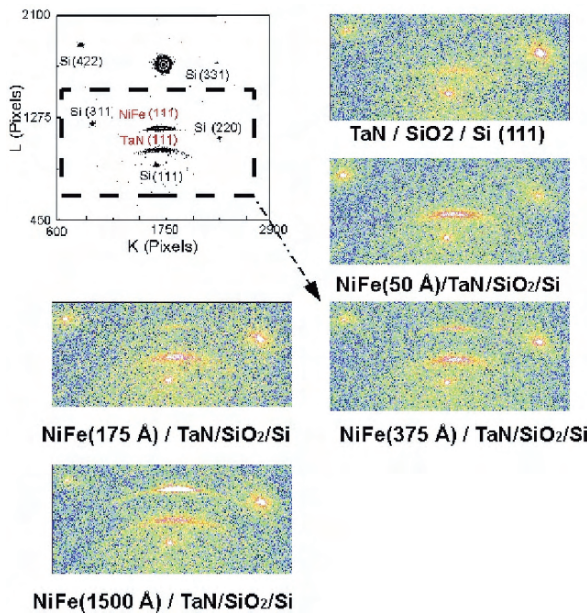


Fig. 9 Synchrotron x-ray reciprocal mapping of permalloy thin films using an image plate. Courtesy of Y. Yoo taken at Argonne National Labs

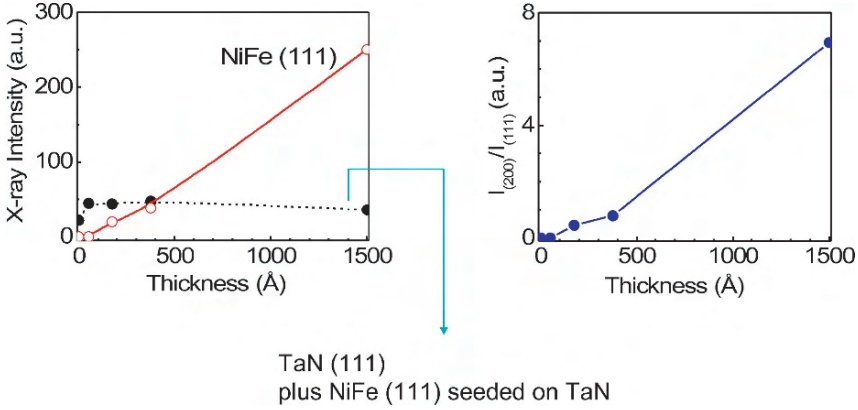


Fig. 10 Plot of x-ray intensity versus film thickness for a TaN/NiFe film. Courtesy of Y. Yoo taken at Argonne National Labs

Unfortunately, this relationship does not match the behavior of strongly oriented magnetoresistive films. To determine the saturated resistor behavior it is necessary to apply the basic tensor relationships as outlined in Nye [27]. This model can then be used to model the effect of a transverse applied field on the AMR of individual sensor elements. The results of solving the minimum energy equation (2) results in (12),

$$\Delta\theta = \frac{MH}{2K_u} \cos\theta, \quad (12)$$

where M is the magnetization, H is the external field, and K_u is the anisotropy constant. This change in angle can be used to calculate the change in resistance for a given applied field. To solve for the magnetoresistance of a thin film, it is necessary to set the proper initial conditions. Initial conditions for the AMR effect often assume that the resistance is completely anisotropic and that there are no isotropic scattering centers, This truly cannot be the case so a modified version of these initial conditions are as follows,

$$\rho_{||}' = \rho_{||} + \rho_o \quad (13)$$

$$\rho_{\perp}' = \rho_{\perp} + \rho_o \quad (14)$$

$$\Delta\rho' = \frac{\rho_{||}' - \rho_{\perp}'}{2}, \quad (15a)$$

$$\rho' = \frac{\rho_{||}' + \rho_{\perp}'}{2}, \quad (15b)$$

where the resistivities are related to the magnetization and ρ_o is the isotropic resistivity. Isotropic resistivity has many contributors such as grain boundaries and other conduction electrons. From equations (13)–(15b) the following tensor relationship can be defined,

$$P'_{\text{total}} = P'_o + P'_m, \quad (16)$$

where P'_m is the magnetic portion of the resistivity, P_0 is the isotropic portion and the total is now

$$P'_{\text{total}} = \begin{bmatrix} \rho_0 & 0 \\ 0 & \rho_0 \end{bmatrix} + \begin{bmatrix} \rho' + \Delta\rho' \cos(2\theta) & \Delta\rho' \sin(2\theta) \\ \Delta\rho' \sin(2\theta) & \rho' - \Delta\rho' \cos(2\theta) \end{bmatrix} \quad (17)$$

By solving the following relationship,

$$\vec{E} = \rho \vec{J} \quad (18)$$

where E is the electric field and J is the current density. The modified AMR relationship can be shown to be similar to the Mohr's circle as described in Nye [27] and is as shown in (19),

$$\rho_{\text{eff}} = \rho_0 + \rho' \left[\left(1 + \frac{\Delta\rho'}{\rho'} \cos 2\theta \right)^2 + \left(\frac{\Delta\rho'}{\rho'} \sin 2\theta \right)^2 \right]^{\frac{1}{2}} \quad (19)$$

This derivation of this equation is detailed in Haji-Sheikh et al. [28]. Equation (19) cannot be calculated directly but the data can be arrived at by an equivalent voltage form. The equation shown in (20) is the measurable form for which the AMR data can be fit,

$$V_{\text{total}} = I_s R_o \left(A + B \left((1 + C \cos 2\theta)^2 + (C \sin 2\theta)^2 \right)^{\frac{1}{2}} \right). \quad (20)$$

To develop an accurate AMR relationship, it is important to make detailed measurements of magnetoresistance versus magnetization angle. This measurement is best done using purely electrical methods since mechanical methods can have significant issues with lash. Screw lash is a mechanical hysteresis which is difficult to overcome and creates an inaccuracy in the rotation angle measurement in the mechanically positioned measurement systems. Alternately, inaccuracies using electrical methods can be as small as a few hundredths of a percent. The development of a measurement system which can apply magnetic field angles carefully and accurately is a critical step. This starts with using a wafer prober which is made from non-magnetic materials then designing an x - y Helmholtz coil which can apply at least 30 Oe in each direction. Unfortunately, to make such a system some mechanical jigg is necessary to align precise right angles.

$$\vec{B} = \vec{a}_z \frac{\mu_o I b^2}{2(z^2 + b^2)^{\frac{3}{2}}} \quad (21)$$

and

$$\vec{B} = \vec{a}_z \frac{N \mu_o I b^2}{2(z^2 + b^2)^{\frac{3}{2}}} \quad (22)$$

for a Helmholtz coil with multiple wraps. The magnetic field is calculated from equation (22) where N is the number of turns per coil, I is the current, z is the halfway point between the coils, b is the radius of the coils and μ_o is the permeability

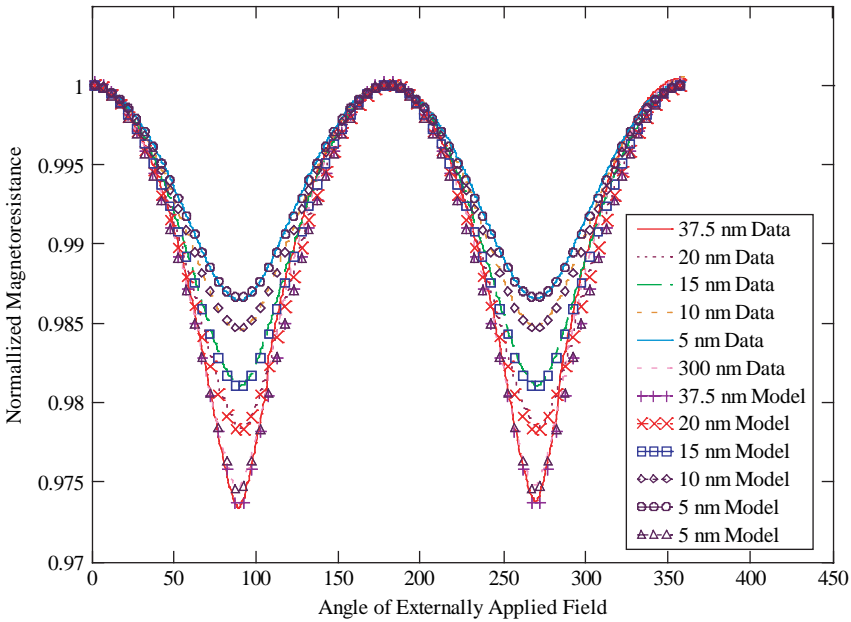


Fig. 11 Magnetoresistance versus angle for individual AMR resistors. This demonstrates the rhombohedral tensor model against the actual data. The model shown in this plot is necessary to extract the angle of magnetization rotation for a given hard axis field [28]

of free space (a permeable pole piece is undesirable due to the remnant field effect). The farther away the coil from the test structure and the larger the coil the higher the current and or the more coils necessary. The more wraps the more wire the higher the resistance in the coil itself which in turn increases the voltage and the temperature. The test temperature needs to be closely monitored due to the high temperature coefficient of resistivity of the permalloy material. Test results, in saturation, of a TaN/NiFe film used in commercial sensing is shown in Fig. 11. The data was taken using an automated test system which was programmed to rotate the field in the horizontal plane. The tensor model shown here can accurately track the resistance as the magnetization angle is rotated. Table 1 shows the coefficients used to fit the resistances plotted in Figure 11. It is apparent that the magnitude of the resistance change varies with the thickness of the film. Not only does the magnitude change but the shape of the curve changes. The C coefficient tracks the shape of the resistance

Table 1 Table of coefficients of fit calculated for the graph in Fig. 11

	5.0 nm	10.0 nm	15.0 nm	20.0 nm	25.0 nm	30.0 nm	37.5 nm
A	.97923	.97580	.97340	.97090	.97050	.96968	.97000
B	.01420	.01640	.01695	.01695	.01726	.01722	.01630
C	.47647	.47700	.57200	.57200	.71800	.76944	.84500

rather than the magnitude. The value of C increases with thickness which leads to one to make the next logical extension. If C is equal to one the equation reduces to,

$$\rho_m = \rho||' * |\cos \theta| \quad (23)$$

which is an interesting result. This result seems to verify the lack of magnetoresistance variation in the (111) which was noted by Berger [25]. The results of the preceding analysis have been applied to a Wheatstone bridge sensor (shown in Fig. 12) against a forty-eight pole-pair ring magnet. A forty-eight pole pair magnet is a composite magnet which has alternating magnetic poles. A ring magnet is the same 48 pole pairs in a ring drive off of a shaft (or spindle).

The bridge is two voltage dividers in parallel and can be solved by the following relationship

$$\Delta V = V_0 \frac{R_2}{R_2 + R_1} - V_0 \frac{R_4}{R_3 + R_4} \quad (24)$$

and

$$R_1 + R_2 = R_3 + R_4 \quad (25)$$

so

$$\Delta V = V_o \frac{R_2 - R_4}{R_2 + R_1}. \quad (26)$$

Each of the individual resistor elements has the same nominal resistance so that the resistors can be represented by

$$R_1 = R_o \left(A + B \left((1 + C \cos 2\theta)^2 + (C \sin 2\theta)^2 \right)^{\frac{1}{2}} \right) \quad (27)$$

$$R_1 = R_o \left(A + B \left((1 + C \cos 2(\theta + 90^\circ))^2 + (C \sin 2(\theta + 90^\circ))^2 \right)^{\frac{1}{2}} \right) \quad (28)$$

$$R_2 = R_3 \quad (29)$$

$$R_1 = R_4 \quad (30)$$

Replacing the individual resistors with the above equations creates a sensor bridge whose output is dependent on the angle of the external field. An important concept in magnetoresistance is the idea of hard-axis and easy-axis film behavior. Figure 12 shows a schematic of the ring magnet test setup. Results using the equations (24) through (30) are plotted in Fig. 13 along with a Voigt-Thompson model. The easy-axis is by description the natural zero energy orientation state of the magnetization. This orientation is governed by deposition conditions and by geometry. Figure 14 shows a easy-axis curve along with a hard-axis curve. To sort out the below saturation behavior it is necessary to solve (10) for the magnitude of the change in angle for a given resistance.

This rearranging results in (31),

$$|\cos \theta| = \left[\frac{1}{4C} \left[\left[\left[\frac{V_o}{I_s R_o} - A \right] \frac{1}{B} \right]^2 - C^2 - 1 + 2C \right] \right]^{\frac{1}{2}}. \quad (31)$$

Fig. 12 Schematic of the ring magnet test set up. The field at the sensor is above the saturation field

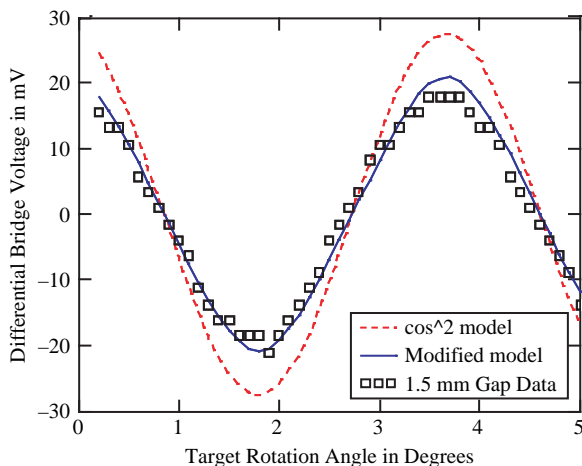
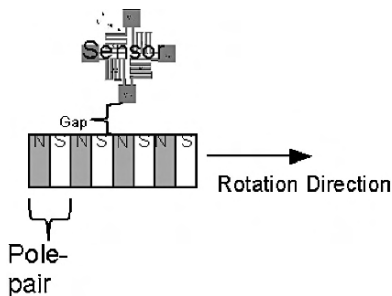


Fig. 13 Comparison of experimental magnetoresistor bridge and the bridge model versus the angle of rotation for a ring magnet on a spindle. The model assumes that the magnetic field saturates the magnetoresistor and the external field rotates 360° every 7.5° of spindle rotation. The $\cos^2\theta$ (Voigt-Kelvin) model over-predicts the sensitivity of the 200° sensor [28, 29]. The gap between the sensor and the ring magnet is 1.5 mm

Fig. 14 Hard and easy axis curves for a single domain magnetoresistive films

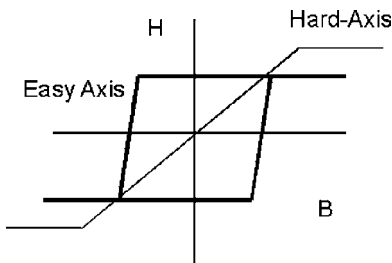
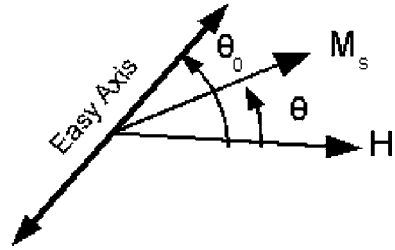


Fig. 15 Domain magnetization rotation off of the easy axis



It is important to relate the change in angle to a given external field. Figure 15 is a schematic of the field rotation vectors.

From Chikazumi and Charap [30] we get,

$$E = -K_u \cos^2(\theta - \theta_0) - M_s H \cos \theta, \tag{32}$$

where E is the energy of the system, K_u is constant of uniaxial anisotropy and M_s is the saturation magnetization. To minimize the energy, the derivative of the energy with the angle of rotation is taken. This derivative is

$$\frac{dE}{d\theta} = -K_u \sin 2(\theta - \theta_0) - M_s H \sin \theta. \tag{33}$$

For a weak magnetic field which is $H \ll K_u/M_s$ and θ_0 is nearly equal to θ then (33) is

$$2K_u \Delta\theta = M_s H \sin \theta_0 \tag{34}$$

then

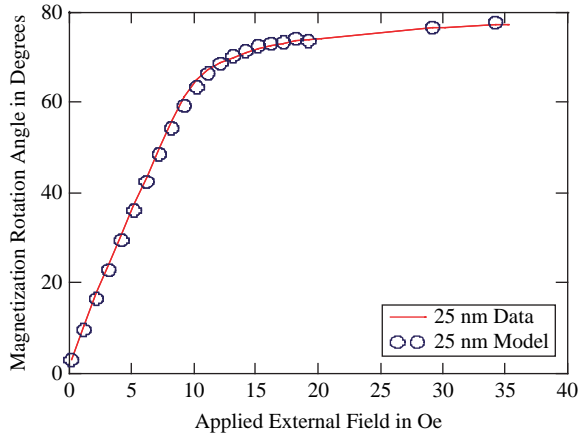
$$\Delta\theta = \frac{M_s H}{2K_u} \sin \theta_0. \tag{35}$$

The modification to match actual off-axis magnetization rotation requires that the geometry of magnetization vector (in the lattice) and the geometry of K_u be considered. If both are initially assumed to be orthorhombic in nature (i.e. rectangular in the 2-d plane) then

$$\Delta\theta = \frac{M}{2K_u} H = \frac{M_o \left[(1 + \alpha \cos 2\theta)^2 + (\alpha \sin 2\theta) \right]^{\frac{1}{2}}}{2K_u \left[(1 + \delta \cos 2\theta)^2 + (\delta \sin 2\theta) \right]^{\frac{1}{2}}} \tag{36}$$

Figures 16 and 17 show the application of (36) to the magnetoresistance equation (20). The results show that we have a reasonable amount of correlation. The sensitivity to the applied field angle for a single strip sensor is fairly high as shown in Fig. 18. Ten degrees of rotation will result in 20% decrease in the positive field direction but a slight increase in the negative field direction. If the resistor is rotated 45°, the sensitivity in the negative field direction is significantly higher than the positive field direction. This holds true until the magnetization reversal happens and then the behavior reverses direction. Figure 19 shows this behavior with sev-

Fig. 16 Hard axis magnetization rotation versus external applied field for an actual 25 nm resistor and a model of a 25 nm resistor. This model is tensor based



eral different resistor widths for a 37.5 nm thick resistor. This behavior is caused by magnetization reversal and can cause problems when using the Permalloy sensor to trigger at a particular field level. Another important effect to consider is the influence of proximity. This proximity effect is due to the below saturation sensitivity increase when two AMR resistors are placed in close proximity with each other. This spacing between the adjacent resistors is called the ‘gap’. Previously we have shown the effect of changing thickness on the sensitivity of the AMR element but there also is an effect of element width on the sensitivity.

This thickness to width ratio is one measure of the sensitivity while proximity is another measure. As shown in the preceding section the thickness effects both the initial slope and the maximum sensitivity at saturation. Additionally the width of

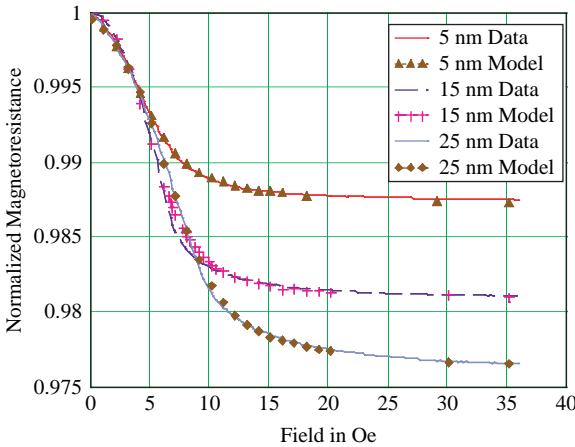


Fig. 17 HA magnetoresistance versus applied field for various thickness magnetoresistors with a constant 35µm width. The tensors used for this model assumes rhombohedral geometry

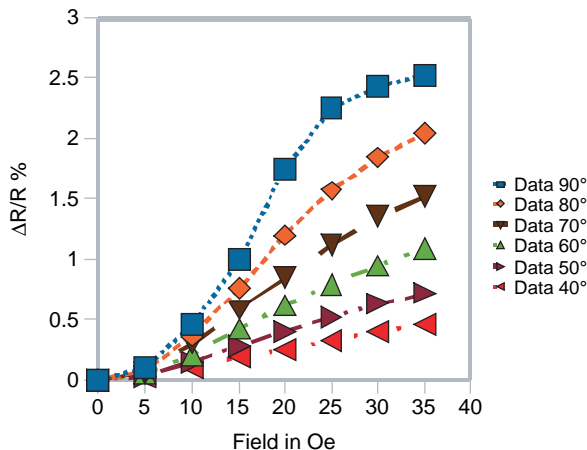


Fig. 18 Effect of field rotation on a single 20µm resistor. This reduction of sensitivity shows the importance of sensor alignment to the external applied field

the resistor for a given resistor thickness can effect the slope but not the maximum sensitivity at saturation. The demagnetizing field is described by Dibern [22] and also Pant [32] as $H_d \approx (t/w) M_s / 4\pi$ where t is the film thickness and w is the width of the resistor.

When more than one resistor is placed in close proximity the adjacent resistors tend to influence each other. The closer the spacing between elements the stronger the proximity effect on the elements. Figure 20 is a schematic representation of such a resistor array. Pant [32] defines a scale factor for the adjacency of the resistors with the relationship of spacing ‘ g ’ with respect to the resistor width, w . The equation is for the proximity effect, calculated from the electrostatic model, is

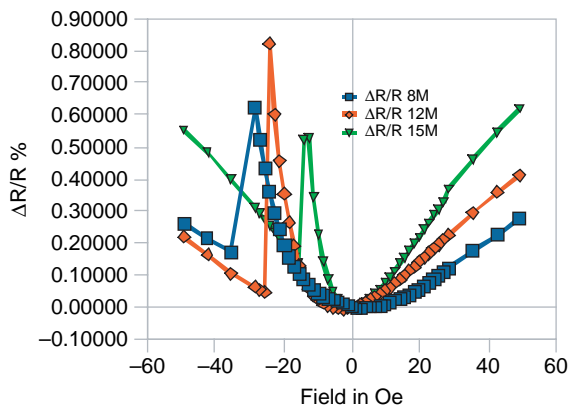


Fig. 19 Single resistor elements of different widths with a 45° applied field

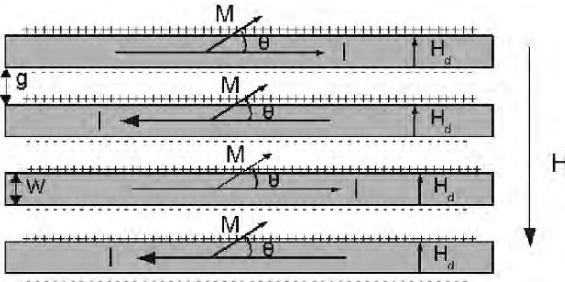


Fig. 20 Schematic of sensor elements using the proximity effect for a serpentine resistor array. The analysis of the effect uses the electrostatic model for magnetics and can be solved numerically [31]

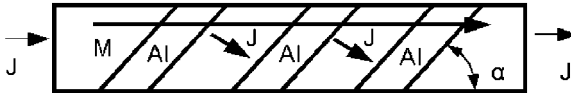


Fig. 21 This figure shows a typical barber-pole sensor element. The angle α and the width of the aluminum shorting straps are determined using finite element methods

$$\alpha(g/w) = \frac{2(g/w)}{1 + 2g/w} + \frac{g/w}{2(1 + g/w)^2} (\pi/2 - 4) \tag{37}$$

so that

$$H_s \approx H_k + t/w \frac{M_s}{4\pi} \alpha(g/w). \tag{38}$$

Each different resistor width behaves as if it was actually a wider resistor. The easy axis behavior is not affected by proximity so that the hysteresis remains the same as a single element of permalloy. A common method of AMR sensor design is the barber-pole sensor. This sensor manipulates the current orientation versus

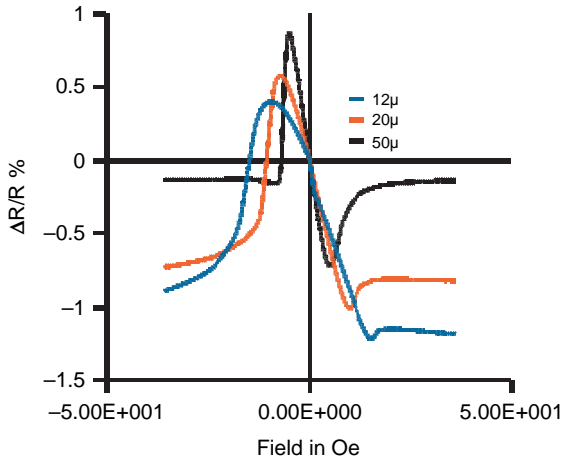


Fig. 22 Response of Barber pole sensor with a 90° applied field

magnetization orientation to create a asymmetrical i.e. odd-function sensor. This odd-function sensor response is completely dependent on the same issues as the single strip element and also is dependent on the orientation of the magnetization orientation at zero applied field. An additional issue with the barber-pole sensor is the element resistance. Since the aluminum shorting straps are significantly lower in resistance then the permalloy, the sensor elements designed using this method are generally much larger than the non barber-pole sensor. Figure 21 shows the important components of the barber-pole sensor. These type of sensors are produced commercially by Philips and Honeywell. The actual design of the shorting straps requires the use of a finite element program such as ANSYS. The models shown in Tummanski [23] for this type of sensor are focused on the linear region of the sensor. It is possible using equation (20), equations (35) and (36) to model the entire sensor behavior. Another complication of this type of sensor is the sense of direction for the magnetization. Both Honeywell and Phillips approach this problem from different directions. In Bharat B. Pants patent [32] the method proposed to keep the magnetization sense is the use of ‘high-current’ straps.

The data in Fig. 22 is generated for barber pole elements using three different resistor widths with a constant shorting strap design. The offset of the resistor data from the 50 μ m and the 20 μ m and 12 μ m is due to the crowding of the current lines at the edge.

5 Future Progress

The uses of magnetic sensors over the last four decades has continually evolved as manufacturing technology has evolved. The introduction of giant magnetoresistors made from magnetoresistive materials has further increased the applications and scope of the basic effect. The introduction of nano-technology will further increase the use of these materials. Nanostructures for data storage applications and

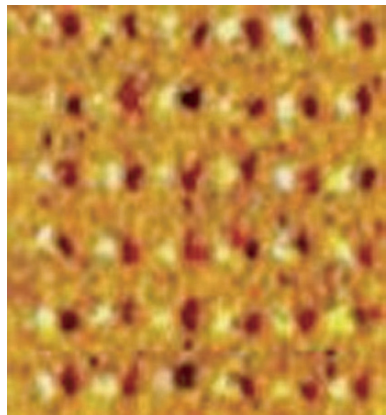


Fig. 23 A MFM image in which the dot magnetizations are organized. Image was produced in air, spacing 1 μ m center to center dot size 200 nm. Image obtained using a Quesant Q-Scope 350 MFM with a cobalt tip

possibly for computing applications may well use ferromagnetic materials such as permalloy. Additionally, nanomagneto-resistive structures have been proposed. Work by Kanparthy [34] in Fig. 23 shows the interaction of nanodots using a Magnetic Force Microscope (MFM). The magnetization patterns may be used to store information in hard drives though a more efficient method to write the data to the nanodots must be developed. This behavior at 200 nm is similar to images made by Zhu et al. [35]. Similar structures may be useful for memory or sensing applications. These structures have complex interactions in which the patterns have some meaning but at this point are still in the process of being understood. Zhu et al. [35] have demonstrated the ability to change the magnetic orientation with their MFM tip. Another interesting development in the area of magnetic nanostructures is the study of magnetic properties of permalloy nanowires [36, 37, 38]. Permalloy retains much of its behavior in the nano-scale making it a good candidate for nano-dimensional sensors.

Several different methods for the manufacturing of nanowires have been reported. These are traditional photo-patterning, e-beam lithography, and nano-templating by the use of anodic nanoporous aluminum oxide to form the nano-wires. All these methods have merit and will continue to have importance for the next decade. The future of magnetic sensing, in the nano-scale, may be a material that has been in use since the 1930s.

References

1. R. S. Popovic, *Hall Effect Devices*, CRC Press, 2004 2nd edition
2. Cohen, Isaac, 'Self aligned hall with field plate', US Patent 7,002,229, February 21, 2006
3. M. Plagens, M. Haji-Sheikh, and W. Matzen, 'Hall-effect element with integrated offset control and method for operating hall-effect element to reduce null offset', US Patent 6,492,697, December 10, 2002.
4. A. Alexander, P. Nickson, and D. Foley, 'Monolithic magnetic sensor having externally adjustable temperature compensation', US patent 6,154,027, November 28, 2000.
5. R. Steiner, A. Haeberli, F. Steiner, and Christoph Maier, 'Spinning current method of reducing the offset voltage of a hall device', US Patent 6,064,202, May 16, 2000.
6. J. Higgs and J. Humenick, 'Integrated circuit with stress isolated Hall element', US Patent 4,578,962, March 25, 1986.
7. A. Bilotti and Gerardo Monreal, 'Chopped hall sensor with synchronously chopped sample-and-hold circuit', US Patent 5,621,319, December 8, 1995.
8. K. Matsuda, Y. Kanda, and K. Suzuki, 'second-order piezoresistance coefficients of *n*-type silicon', *Jpn. J. Appl. Phys.*, 28, L1676–L1677 (1989).
9. Y. Kanda, 'A graphical representation of the piezoresistance coefficients in silicon', *IEEE Trans. on Electron Devices*, ED-29, 1, January (1982).
10. D. A. Nepela and R. I. Potter, 'Head assembly for recording and reading, employing inductive and magnetoresistive elements', US patent 3,887,945, June 3, 1975.
11. F. Lee, 'Supersensitive magnetoresistive sensor for high density magnetic read head', US Patent 4,047,236, Sept. 6, 1977.
12. T. McGuire and R. Potter, 'Anisotropic magnetoresistance in ferromagnetic 3d alloys', *IEEE Trans. Magn.*, 11, 4, 1018–1038, (July 1975).

13. K. Kanai, 'Magnetic head with thin sheet exhibiting magnetoresistive property', US Patent 4,051,542, Sept. 27, 1977.
14. K. Kanai, 'High sensitivity magnetic head using magneto-resistive effect element', US Patent 4,068,272, Jan. 10, 1978.
15. J.-P. Lazzari, 'Magnetic transduction device with magnetoresistances', US Patent 4,315,291, Feb. 9, 1982.
16. J.-P. Lazzari, 'Magnetic reading and writing head with magnetoresistant element', US Patent 5,168,408, Dec. 1, 1992.
17. T. A. Schwarz, P. G. Bischoff, C. M. Leung, J. C. Chen, and P. Thayamballi, 'Method of making a magnetoresistive head with integrated bias and magnetic shield layer', US Patent 5,312,644, May 17, 1994.
18. M. T. Krounbi, J. H.-T. Lee, 'Simplified method of making merged MR head', US Patent 5,779,923, July 14, 1998.
19. M. C. Paul, G. F. Sauter, and P. E. Oberg, 'Thin-Ferromagnetic-Film magnetoresistance manometer sensitive to easy axis field components and biased to be insensitive to hard axis components', US Patent 3,546,579, Dec. 8, 1970.
20. S. Ito, M. Nagao, K. Toki, and K. Morita, 'Magnetic rotary encoder for detection of incremental angular displacement', US Patent 4,319,188, Mar. 9, 1982.
21. M. J. Haji-Sheikh, M. Plagens, and R. Kryzanowski, 'Magnetoresistive speed and direction sensing method and apparatus', US Patent 6,784,659, August 31, 2004.
22. U. Dibbern, 'Magnetic field sensors using the magnetoresistive effect', *Sensors and Actuators*, 10, 127–140, (1986).
23. S. Tummanski, *Thin Film Magnetoresistive Sensors*, IOP 2001, pp. 19–30.
24. T. Yeh, M. Sivertsen, and C.-L. Lin, 'Preferred Crystal Orientation of NiFe Underlayers and its effect on Magnetostriction of Co/Cu/Co Thin Films', *IEEE Trans. on Magn.*, 34, (4) (July 1998).
25. L. Berger and S. A. Friedberg, 'Magnetoresistance of a permalloy single crystal and effect of 3d orbital degeneracies', *Phys. Rev.*, 165, (2) pp. 670–679, (1968).
26. Th. Rijkers and S. Lenczowski, 'In-plane and out of plane magnetoresistance i...', *Phys. Rev. B*, pp. 362–366, (1997).
27. J. F. Nye, *Physical Properties of Crystals*, Oxford Science Publications, first publication 1959.
28. M. J. Haji-Sheikh et. al., 'Anisotropic Magnetoresistive Model for ...', *IEEE Sens. J.*, pp. 1258–1263, Dec. 2005.
29. M. J. Haji-Sheikh and Y. Yoo, 'An accurate model of a highly ordered 81/19 Permalloy AMR Wheatstone bridge sensor against a 48 pole pair ring-magnet', *IJISTA*, 3, No (1/2), 95–105, (2007).
30. S. Chikazumi and S. Charap, *Physics of Magnetism*, Robert E. Krieger Publishing Company, pp. 260–263, (1978).
31. Michael Haji-Sheikh, 'TaN/NiFe/TaN anisotropic magnetic sensor element', US Patent 5,667,879, September 16, 1997.
32. B.B. Pant, *J. Appl. Phys.* 79, 6123 (1996).
33. B. B. Pant, D. R. Krahn, and R. B. Fryer, 'Magnetic field sensing device', US Patent 5,247,278, September 21, 1993.
34. S. Kanparthy, Thesis, Northern Illinois University, Fall (2007).
35. X. Zhu, P. Grütter, V. Metlushko, and B. Ilic, 'Magnetic force microscopy study of electron-beam-patterned soft permalloy particles: Technique and magnetization behavior', *Phys. Rev. B*, 66, 024423, (2002).
36. Y. Rheem, B.-Y. Yoo, B.K. Koo, W.P. Beyermann, and N. V. Myung, 'Synthesis and magneto-transport studies of single nickel-rich NiFe nanowire', *J. Phys. D: Appl. Phys.* 40, 7267–7272, (2007).
37. L. Piraux, K. Renard, R. Guillemet, S. Mtéfi-Tempfli, M. Mtéfi-Tempfli, V. A. Antohe, S. Fusil, K. Bouzehouane, and V. Crosb, 'Template-Grown NiFe/Cu/NiFe Nanowires for Spin Transfer Devices', *Nano Lett.*, 7(9), 2563–2567, (2007).
38. A. O. Adeyeye, R. L. White, 'Magnetoresistance behavior of single castellated Ni₈₀Fe₂₀ nanowires', *J. Appl. Phys.* 95, 2025 (2004).

Improving the Accuracy of Magnetic Sensors

Pavel Ripka

Abstract This chapter describes techniques for improvement of the accuracy of magnetic sensors. Many of these ideas can be used for other types of sensors and measurement systems in general.

Keywords Magnetic sensors · feedback compensation · temperature stability · noise measurement · noise reduction

1 Introduction

Besides the well-established magnetic sensor types (Hall for high fields, Fluxgate for lower fields, SQUID for small field gradients and proton for absolute scalar measurements), new types of magnetic sensors continuously appear – from AMR, GMR, CMR, SDT to GMI, HT SQUID and a variety of novel semiconductor sensors [1, 2].

Sensor noise is very popular parameter for the characterisation of performance of magnetic sensors. But the noise value is not always the parameter which really limits the sensor performance. Sometimes the temperature stability or linearity are of higher importance.

2 Temperature Stability

For many applications, temperature stability is the critical issue, not sensitivity, noise or linearity. As the temperature properties are difficult to measure and interpret, they are not often discussed in research papers. Two basic temperature parameters are usually given: temperature dependence of sensitivity (usually given in

Pavel Ripka
Czech Technical University, Faculty of Electrical Engineering, Department of Measurement,
Prague, Czech Republic, e-mail: ripka@fel.cvut.cz

ppm/K) and temperature drift of the offset (in nT/K). Using the two mentioned basic units is a good practice as it allows direct comparison. However some authors and manufacturers prefer mimics and they give these parameter indirectly (e.g. they express offset drift in millivolts of the output voltage or even in ppm of the full scale).

In wider temperature range these parameters are often not constant or even not monotonic. This is also the case when the dominant source of the temperature dependence is compensated. In such case the stability is expressed by parameter limits.

Temperature dependence of sensitivity usually has simple sources such as temperature changes in permeability, resistivity, coil and core dimensions. Sensitivity temperature coefficients ('tempcos') of the sensors from one batch are similar. Origins of the temperature dependence of the offset are more complicated and even the sensors from the same production batch have these parameters different (including different sign).

Besides other techniques, temperature dependence of the sensitivity can be effectively suppressed by the negative feedback. High temperature dependence of the sensitivity of the sensor is replaced by much lower temperature dependence of the feedback coil (typically 30 to 200 ppm/K). It is worth stressing that this technique cannot improve the offset stability.

2.1 Hall Sensors

The sensitivity of Hall sensors depends on a number of parameters as the density and mobility of charge carriers, which are strongly temperature dependent [3]. The temperature coefficient of the magnetic sensitivity of Si devices is about 0.1%/K. Thin-film InSb Hall sensors are much more sensitive, but their temperature coefficient of sensitivity is about $\pm 1\%$ /K.

Semiconductor materials of Hall sensors suffer from unwanted piezoelectric and piezoresistive effect. Mechanical stress is mainly caused by the plastic package, which has different temperature dilatation than the chip inside. This stress creates piezoelectric voltages and inhomogeneity of the chip resistivity, causing changes in the sensor offset. Offset temperature dependence of Hall sensors can be suppressed by avoiding the package stress, but this is not easy for low-cost volume production. Another technique is to use several sensors in one package, each being positioned at different part of the chip and having different current direction. And probably the most effective is the spinning-current technique: the sensor element is symmetrical and current terminals are periodically swapped with voltage contacts. Other measurements are made with reversed current direction. In the basic case of rectangular sensor all four readings are averaged. Some sensors use more contacts. Typical values of the sensor offset are $B_{\text{off}} \approx 10 \text{ mT}$, 1 mT, 0.1 mT for Si, InGaAs, and InSb Hall devices, respectively. Long-term fluctuations correspond to $B_{\text{off}} \approx 10 \mu\text{T}$ in high-quality silicon Hall devices. [4]

2.2 Magnetoresistors

The most stable magnetoresistors are AMRs (Anisotropic MagnetoResistors) [5]. Although the influence of basic tempo of AMR material resistance is suppressed by the bridge configuration, the temperature coefficient of sensor sensitivity is typically as high as 0.25%/K (Philips KMZ 51). By using of the feedback it can be reduced to the TC of the feedback coil, which changes from piece to piece and often can be below 100 ppm/K. [6].

Only AMR sensors with periodical flipping have low and stable offset. Flipped sensors have the temperature offset drift at the order of 10 nT/K. This can be further suppressed by additional balancing of the bridge.

GMR sensors (Giant MagetoResistors) may withstand higher temperatures than AMRs, but they have in general higher both offset and sensitivity temperature dependence and their characteristics are non-linear.

SDT (Spin-Dependent Tunneling) magnetoresistors are also called TMR (tunneling magnetoresistors). They have similar properties as GMRs, but can be made smaller with high resistance and thus low power consumption. Fig. 1 shows temperature dependence of sensitivity for prototype SDT sensor (NVE). The corresponding temperature coefficient is as high as $-670 \text{ ppm}/^\circ\text{C}$, but the linear dependence allows corrections.

Periodical de- or re-magnetization to increase the stability is not possible in case of GMR and SDT sensors. Some GMR sensors can be even magnetically destroyed by strong external magnetic field.

2.3 Fluxgate Sensors

Although the typical offset drift of fluxgate sensor ranges from 0.1 to 0.5 nT/K, in the best magnetometers the overall offset stability of 100 pT to 1 nT in the temperature

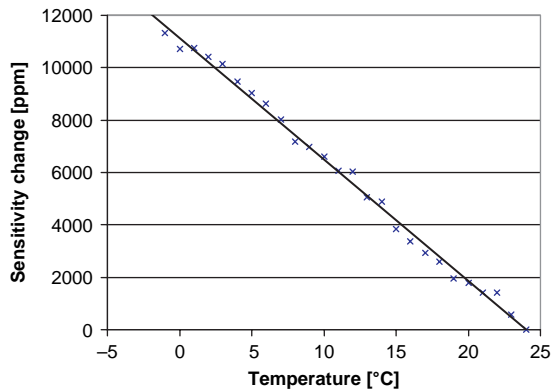
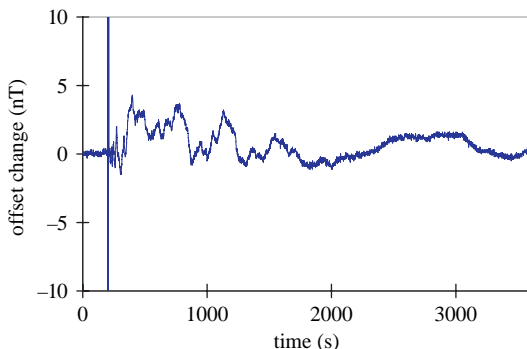


Fig. 1 Sensitivity change vs. temperature for SDT sensor

Fig. 2 Offset recovery after temperature shock



range from -40 to $+70^{\circ}\text{C}$ was achieved. Recovery time after temperature shocks is important parameter for field instruments. Temperature gradients cause mechanical stresses, which have long relaxation times. Fig. 2 shows the offset recovery of the sensor subjected to rapid cooling from $+22^{\circ}\text{C}$ to -25°C .

Low offset of fluxgate sensor is a consequence of the fact that positive and negative saturation fields of core magnetic material have exactly the same value. Non-zero offset and its changes are mainly caused by:

1. Magnetically hard regions in the sensor core associated with structural or surface defects
2. Thermally induced mechanical stresses coupled through non-zero magnetostriction
3. Changes in the magnetic properties of the core material and the parameters of the excitation field coupled through inhomogenities of the core and the winding.

The universal recipe to increase the temperature stability is to match the thermal expansion coefficients of all the sensor parts.

The sensitivity tempco of fluxgate sensors is usually around 30 ppm/K, but some fluxgate magnetometers are compensated up to 1 ppm/K [7]. In the Earth's field of 50 000 nT and after temperature change of 20K this corresponds to 1 nT error, which is sometimes comparable with field change measured.

2.4 GMI Sensors

Experimental data on temperature effects in GMI sensors are rarely published. Although the sensitivity can be quite stable (200 ppm/K without feedback was reported in [8]), uncompensated offset drift is large. Fig. 3 shows temperature offset drift of permalloy GMI sensor described in [9].

As the measured quantity in GMI sensors is usually modulus of impedance, the offset of the single-core sensor is directly influenced by temperature dependence of the material DC resistivity and also through inductance by temperature dependence

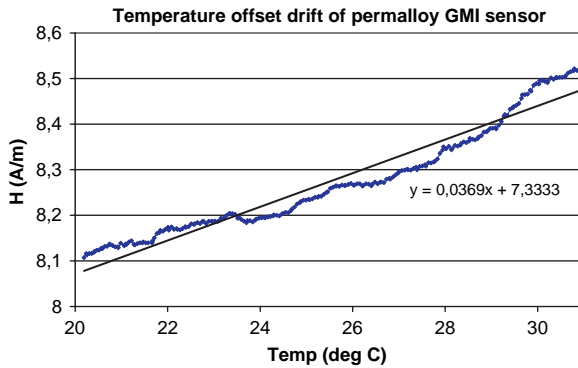


Fig. 3 Temperature offset drift of GMI sensor [9]

of permeability. Symmetrical characteristics require bias, which couples sensitivity coefficient to offset drift and which can make the characteristics very complicated.

We expect significant improvement by using compensated double-core sensor. However, offset stability of GMI sensors is also affected by perming effect (remance), which is not negligible even for very soft materials; thus for the precise applications the sensor should be periodically de- or re-magnetized.

2.5 Proton Magnetometer

Proton magnetometer is almost absolute instrument, based mainly on gyromagnetic ratio of proton, which is temperature independent. Thus the temperature dependence is given by the electronics (e.g. the stability of the reference frequency). Achievable absolute error is 0.1 nT (in 60 000 nT range) in a broad temperature range [10]. Overhauser and optically-pumped resonant magnetometers are faster and more resistant to field gradients, but they cannot achieve absolute accuracy (including temperature effects) better than 1 nT [11].

3 Linearity

Imagine that you should detect a small magnetic signature from a moving platform. The field to be detected may be as small as 0.1 nT in the presence of the Earth's field of 50 000 nT. You can use multi-sensor gradiometric array to suppress the background field, but due to the movements of the platform the whole array is exposed to large changing field components. The required linearity error is less than 0.1 out of $\pm 50\,000$, which is 1 ppm. The only vectorial sensor, which can achieve ppm linearity is compensated fluxgate.

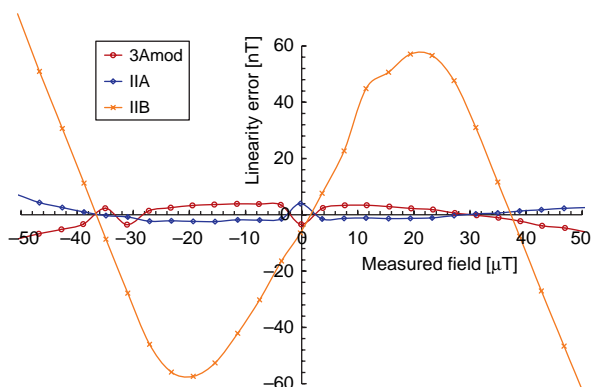


Fig. 4 Linearity error of feedback compensated PCB fluxgate. Sensor IIB has short feedback coil (covering only part of core) which creates non-homogenous compensation field. This results in degraded linearity [12]

Linearity of any magnetic sensors can be substantially improved by using a feedback compensation. For this two criteria should be met:

1. enough gain in the open loop
2. high linearity of the compensation coil

In highly linear systems care should be taken to select linear enough analog components. Push-pull power stage for boosting the feedback current is typical Achilles heel, when ppm linearity is in question.

In low-cost fluxgate sensors the pick-up coil simultaneously serves for the feedback compensation. In this case the coil design should be a compromise. Figure 4 shows three types of PCB fluxgate sensors: sensor IIB has too short compensation/pick-up coil, which results in degraded linearity.

3.1 How to Measure the Linearity

Linearity measurement for feedback compensated sensors is rather difficult. You need a computer-controlled system consisting of a precise current source (usually DC current calibrator), stable calibration coil and two precise dc voltmeters. First condition is that no ferromagnetic object should be close to the measurement setup, as the non-linearity of the magnetization characteristics of such object could influence the measured results. Further, the variations of the external field or the sensor noise should not affect the results. Another source of error is the noise and stability of the testing field. This is usually caused by noise and variations of the testing current, but also mechanical vibrations and temperature dilatation may cause changes of the constant of the testing coil or in the position between the coil and the sensor under test. Good technique how to suppress some of the mentioned influences is

- avoid vibrations
- make the measurement in thermostated room without an air flow
- avoid ferromagnetic objects in the vicinity of the testing setup (including parts of the building)
- use testing coil much larger than the sensor (for 25 mm long sensor 50 cm diameter Helmholtz coil is a minimum)
- check the coil heating during the test: temperature dilatation changes the coil constant
- measure the testing current simultaneously with the sensor output – synchronous trigger is a must
- use integrating voltmeter (6,5 digit resolution is a minimum). Use long measurement time, which is a multiple of power line cycle (50 PLC or more) to suppress AC fields

4 Crossfield Effect

Imagine a magnetic sensor which has sensitive direction in y -axis. Crossfield effect (or crossfield error) is an unwanted sensitivity to fields in x direction. It is not so that the sensor is not correctly aligned. This is a non-linear effect, e.g. in AMRs if $H_y = 0$, sensitivity to H_x is zero. Crossfield error in AMRs happens only in case that both H_x and H_y are non-zero. The situation in fluxgates is different.

4.1 AMR

Crossfield sensitivity is inherent to AMR sensor. It is caused by the anisotropic character of the sensor. One can reduce the effect by changing the sensor design, but this also reduces the sensitivity and finally increases the sensor noise. It is important to understand this effect in order to use AMR sensors correctly. Finally we will show that the effect can be (1) numerically corrected, (2) suppressed using flipping or (3) almost annihilated by feedback. This description is based on AMR crossfield model derived in [13].

Let us consider the ideal single-domain AMR strip with current flowing in x direction and only coherent rotation of magnetization. The strip has uniaxial anisotropy with easy direction x and effective field of H_0 . The physical meaning of H_0 is that perpendicular field of this strength will rotate the magnetization from easy axis by 90° (to the direction of the external field).

Let us first examine only a presence of the magnetic field to be measured, $H = H_y$, applied along the y -axis and in the plane of the strip so that it is perpendicular to easy direction x . This field causes the internal magnetisation of the strip to rotate by an angle φ , which can be calculated solving the minimum energy equation as:

$$\sin \varphi = \frac{H_y}{H_0} \quad \text{for } H_y < H_0 \quad (1)$$

and $\varphi = 90^\circ$ for $H_y \geq H_0$

As derived in [5], the resistance in x direction is then:

$$R(H_Y) = R_0 + \Delta R \left[1 - \left(\frac{H_Y}{H_0} \right)^2 \right] = R_0 + \Delta R \cos^2 \varphi \quad (2)$$

By means of Barber poles we deflect the current by 45° , so that the resistance equation becomes:

$$R = R_0 + \Delta R \cos^2 (\varphi + 45^\circ) \quad (3)$$

$$(4)$$

and because

$$\cos(45^\circ + \varphi) = \frac{\sqrt{2}}{2} (\cos \varphi - \sin \varphi) \quad (5)$$

$$\begin{aligned} \cos^2(45^\circ + \varphi) &= \frac{1}{2} (\cos \varphi - \sin \varphi)^2 \\ &= \frac{1}{2} (\cos^2 \varphi + \sin^2 \varphi - 2 \cos \varphi \sin \varphi) \\ &= \frac{1}{2} - \cos \varphi \sin \varphi = \frac{1}{2} - \sin \varphi \sqrt{1 - \sin^2 \varphi} \end{aligned}$$

we get

$$R = R'_0 \pm \Delta R \frac{H_Y}{H_0} \sqrt{1 - \left(\frac{H_Y}{H_0} \right)^2} \quad (6)$$

where $R'_0 = R_0 + \frac{\Delta R}{2}$ and '±' represents two different possible orientations of the barber poles (aluminium stripes on top of the Permalloy strip). For $H_y \ll H_0$, we get linear field dependence:

$$R = R'_0 + \Delta R \frac{H_Y}{H_0} \quad (7)$$

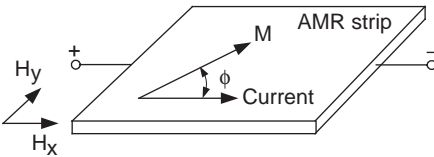


Fig. 5 Magnetization in AMR strip is deflected from x direction by measured field H_y . The cross-field H_x is also present [1]

Nonlinearity which appears for higher fields is effectively suppressed if four strips with proper directions of barber poles form a Wheatstone bridge. This configuration also suppresses the effect of temperature dependence of the resistivity. For linearity it is the best to supply the bridge from a current source; for the best temperature stability it is better to use voltage source.

Now let us consider a more complicated case: the total external field H is no longer in y direction. This means that there is a crossfield H_x present simultaneously to the measured field H_y . We suppose that the external field H is much smaller than the anisotropy field: $|H_x|, |H_y| \ll |H_0|$.

For the **total energy density** of a single domain of anisotropic material in the magnetic field H we may write

$$E = E_A + E_H = \frac{1}{2}\mu_0 M_S H_0 \sin^2 \varphi - \mu_0 M_S H \cos \alpha$$

where M_S is a saturation magnetization, α is an angle between the magnetization \mathbf{M}_s and external field $\mathbf{H} = (H_x, H_y)$ and φ is an angle between the magnetization \mathbf{M}_s and easy direction x

In order to find the energy minimum we solve the equation $dE/d\varphi$ and (still for $|H_x|, |H_y| \ll |H_0|$) we will find

$$\sin \varphi = \frac{H_y}{H_x + H_0} \quad (8)$$

Thus for the strip resistance we find [13]:

$$R = R'_0 + \Delta R \frac{H_y}{H_x + H_0} \sqrt{1 - \left(\frac{H_y}{H_x + H_0}\right)^2} \quad (9)$$

so that the sensor output is approximately proportional to

$$V_1 \approx \frac{H_y}{H_x + H_0} \quad (10)$$

and the sensitivity to the crossfield H_x is

$$\frac{\partial V_1}{\partial H_x} = \frac{-H_y}{(H_x + H_0)^2} \quad (11)$$

Also in this case the bridge improves the linearity at high fields. From eq. 9 it is clear that there is no response to crossfield H_x when the measured field H_y is zero. That is why the crossfield is erased by feedback compensation of the measured field. We also see that increasing H_0 improves the linearity in uncompensated mode, but decreases sensitivity. H_0 is an effective anisotropy field, which is a result of both shape anisotropy and induced anisotropy. Thus we can change H_0 by changing the strip thickness or by magnetic field annealing.

If we cannot use the feedback compensation, better method than increasing H_0 is to correct the crossfield error by calculation. For this we should also have another sensor, which measures the orthogonal component. If we know the value of H_0 , we can calculate H_y from the reading of these two sensors. The equation has no analytical solution, but it can easily be solved numerically [14]. The only problem is the knowledge of H_0 . It is not easy to measure it, but fortunately the sensitivity to this parameter is not critical [15].

But how we can correct the crossfield by flipping? If we flip the sensor (i.e. reverse the strip magnetization), we change the sign of H_0 in the formula and get the output voltage

$$V_2 \approx \frac{H_y}{H_x - H_0} \quad (12)$$

Standard processing is to subtract V_1 and V_2 during the analog signal processing:

$$V \approx \frac{H_y}{H_x + H_0} - \frac{H_y}{H_x - H_0} = \frac{-2H_y H_0}{H_x^2 - H_0^2} \quad (13)$$

The sensitivity of V to the crossfield H_x is reduced for typical situation when $H_x < 0.1 H_0$, and the remaining error is :

$$\frac{\partial V_{\text{out}}^{\text{flip}}}{\partial H_x} = \frac{4H_y H_x H_0}{(H_x^2 - H_0^2)^2} \quad (14)$$

4.2 Fluxgate

Crossfield error in fluxgate sensors was analyzed in [16]. This effect is very small in sensors having long core with low demagnetization with respect to the measured field and large demagnetization in perpendicular direction. When the core has ring shape, this is no longer true – however, proper design can keep this error in the range of 1 to 5 nT in the Earth's field (50 000 nT) [17].

A fundamental difference from crossfield error in AMR is that in fluxgate this error appears even for zero measured field. It is simply not possible to find 'insensitive direction' for fluxgate sensor. More precisely, insensitive direction depends on the amplitude of the perpendicular field. This is shown in Fig. 6 – the sensor output change with perpendicular field is highly non-linear.

A decrease in sensitivity with crossfield can be observed even for perfectly homogenous core. Possible reasons may be small gain in the feedback loop, non-homogeneous compensation field or sensor non-linearities. Figure 7 shows the sensitivity change for ring-core and race-track sensor as a function of the perpendicular field. In this case the high sensitivity change of the race-track sensor was caused by the lack of a homogenous compensation field. The race-track's pick-up coil which covered only 70% of the core length was also used for the feedback. This compromised the sensor performance.

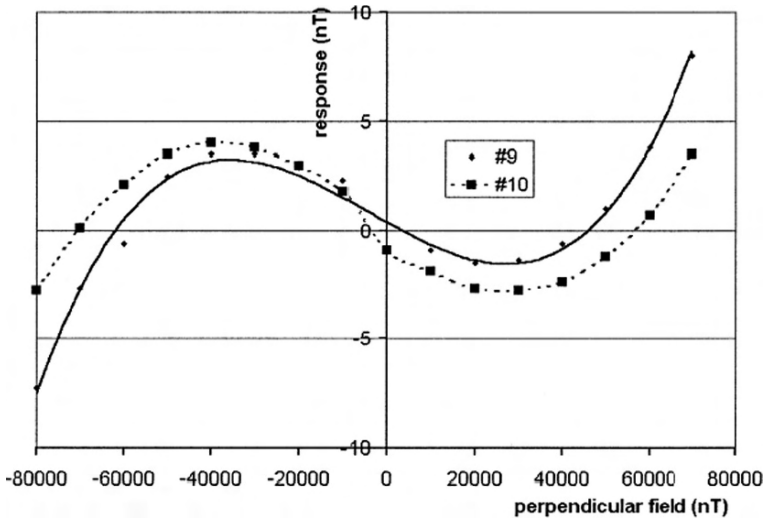


Fig. 6 Perpendicular field response of ring core sensor when the measured field was zero. Results for two sensors from the same batch are shown – from [16]

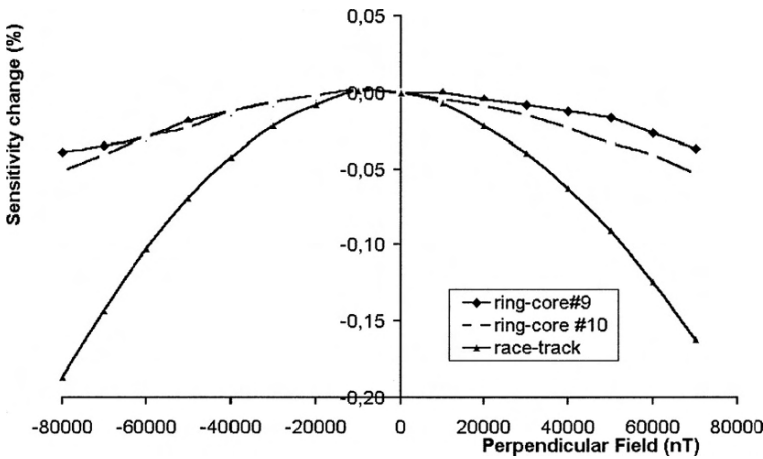


Fig. 7 Sensitivity change as a function of the crossfield for ring-core and race track sensors. The sensitivity was tested by a field step of $50\mu\text{T}$ – from [16]

5 Noise

Noise of magnetic sensors has usually $1/f$ character, which at some cut-off frequency (usually kHz) changes to white. Noise power spectrum density in field units at 1 Hz (in $\text{pT}/\sqrt{\text{Hz}}@1\text{ Hz}$) is a commonly accepted parameter, which allows comparing sensors. However some authors express the noise figures in units of output voltage

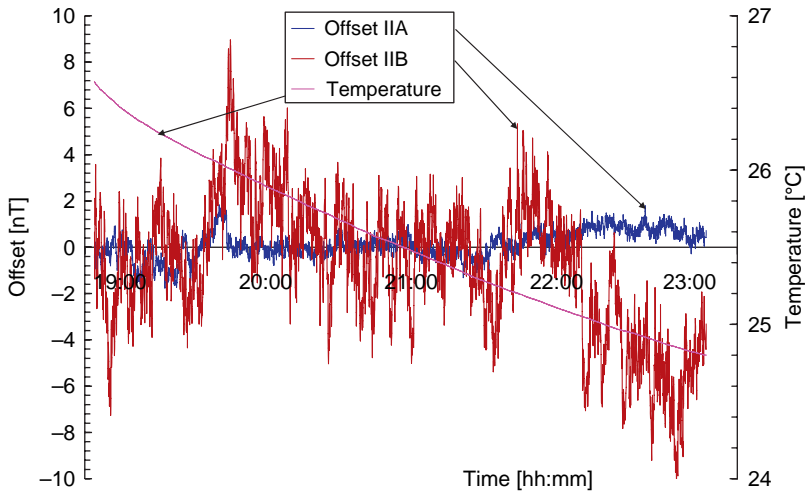


Fig. 8 Offset stability of PCB fluxgate. Sensor IIA with homogenous excitation is more stable [12]

or give the noise spectrum density at higher frequency. This is not fair, unless clearly specified.

In general the noise is magnetic or electrical. The origin of the magnetic noise is still a mystery, but we know well how to reduce it. In magnetoresistors it is important to achieve single-domain state or to reduce the closure domains to minimum. This is achieved by proper geometry (thin film), chemistry and field annealing [5]. In fluxgate sensors the single domain state cannot be achieved, although there were efforts in this direction. However, noise can be reduced to units of $\text{pT}/\sqrt{\text{Hz}}$ @ 1 Hz using the well-known recipes:

- low coercivity, high permeability master alloy for the core
- induce anisotropy in the hard direction to make the magnetization characteristics smooth and the magnetization process rotational
- homogenous excitation and compensation
- no mechanical stress in the sensor structure
- stable excitation with very high peak value

Long-term offset stability (or ultralow-frequency noise) is of equal importance as 1 Hz noise. One of the key requirements is a stability and homogeneity of the excitation field (Fig. 8).

5.1 How to Measure Noise Values

The noise measurement should be performed in a good magnetic shielding. The suggestions are following:

- demagnetize the shielding prior to the measurement, but wait one hour after each demagnetization for domains to relax
- keep the room temperature constant with no airflow (thermoelectric currents in the shielding wall can generate field which mimics the sensor noise)
- reduce mechanical vibrations (together with the shielding remanence they cause magnetic noise)
- reduce AC fields
- switch-off all redundant instruments
- wait for the night time – external interfering fields are lower
- measure power density spectrum using RMS average ($N = 50$ to 100) – this makes the spectrum smooth (avoid vector averaging that suppresses the very noise we want to measure).
- adjust reasonable overlap for averaging – this reduces the measurement time (suggested value is 95%)
- use window (e.g. Hanning, not ‘flat’)
- estimate the noise spectrum density at 1 Hz: do not take the value from one point (narrow cursor)
- monitor the time plot of the sensor output: use the same spectrum analyzer, not extra oscilloscope. Notice that some analyzers use signal stored before you start the measurement! Steps in the data caused by switching the instrument range or spikes of the external origin can spoil the noise measurements. These events are easy detectable as artifacts in the time plot.
- repeat the measurement with sensor excitation powered off to be sure that the noise of your instruments is well below what was measured as a sensor noise

Some magnetoresistors have very low noise, but simultaneously low sensitivity [18] so that what we measure is a noise of electronics, not a sensor noise itself. These sensors are usually configured as Wheatstone bridge. There are two techniques to make these measurements:

- (1) to supply the sensor bridge by AC voltage or current – this transforms the signal to higher frequency, where the amplifier noise is lower. After amplification the sensor signal should be demodulated back to its original base frequency band (or if possible the FFT spectrum analyzer should be set accordingly).
- (2) to use cross-correlation technique: instead of differential amplifier at the bridge output, two independent single-ended amplifiers are used and their outputs are processed by two-channel FFT spectrum analyzer. By using cross-spectrum the (uncorrelated) amplifier noise can be suppressed by 6 to 20 dB. This technique was used in the study, which compared AMR, GMR and SDT sensor noise [19].

6 Perming Effect

Remanence is often a weak point of magnetic sensors containing ferromagnetic component as a functional core or a field concentrator.

In general low-coercivity soft magnetic alloys should be used. But this is not enough. A magnetic field shock changes the sensor offset and the only possibility to erase it is to demagnetize the sensor – or magnetize it into the well defined saturated magnetic state.

In fluxgates this is achieved automatically by using the excitation field of very high intensity. In order to reduce the power consumption and also the heating, the excitation current is in the form of narrow pulses.

Flux concentrators may have demagnetization winding, but this is used only very rarely. Most of the sensors have no such winding. The amplification factor of the concentrators is usually moderate and its remanence is often ignored.

AMR sensors use flipping. Modern AMRs have integrated flat flipping coil. Although the efficiency of such coil is low, as it has poor coupling to the core, the power consumption can often be tolerated and integration lowers the price. Optimizing the flipping pulse shape and amplitude reduces the perming [20].

7 Remarks on Digitalization

A/D converter is often a weak point of the magnetometer. Requirements on the digitizer noise, linearity and stability are so demanding, that alternative solution was found in integration of the converter into the feedback loop [21]. However this was not successful and digitally assisted analog feedback loop with external ADC is still the best solution. The assisting circuits configure the parameters of the magnetometer in order to change the scale, sensitivity, time constant and dynamic range to fit best the current requirements.

8 Peculiarities of Current Sensors

Current sensors are often based on magnetic sensing. They were recently reviewed in [22].

A specific requirement for current sensors is geometrical selectivity: some currents should be measured while some others should be ignored. The requirement on current sensor is also to be insensitive to ambient magnetic field and its variations. This can be achieved by using a magnetic yoke. If this is not possible (e.g. the yoke would be too bulky to avoid saturation), magnetic gradient is measured using a couple of magnetic sensors. Sometimes more complicated circular sensor fields are used.

Current transformers with nanocrystalline cores can be made very small and precise [23]. A resistance to DC current component and external magnetic field component is often required (e.g. by the producers of power meters) – this can be achieved using low-permeability core; the accuracy can be maintained by

compensation of the amplitude and phase error, which can be made constant over wide range of the measured current [24].

Rogowski coils (also called dI/dt sensors) should be used with integrator – either classical analog or by single-chip digital. The device is also being used in power meters – the advantage is that contains no magnetic material so that it cannot be saturated [25].

The most popular are **Hall current sensors** – either open-loop or compensated [26]. The main common problem is a limited zero stability and a remanence of the yoke. The latter can be reduced by periodical demagnetization of the yoke. However, **fluxgate current sensors** are typically 10-times better in zero stability than Hall current sensors due to the fact that their core is periodically saturated by the excitation current. DC current transformers and comparators often work on fluxgate principle.

9 Peculiarities of Position Sensors

Magnetic field sensors are extensively used in industry for construction of position sensors: either two-state (proximity sensors, rotational sensors) or linear output sensors (which can measure linear or rotational position. These position sensors often use permanent magnet attached to the sensor or to the target. Leading market for sensors is automotive industry. It is well known that temperature stability is a critical issue and position sensors must be temperature characterized or compensated. Heremans has shown that at high fields semiconductor magnetoresistors are more temperature stable than Hall sensors [27]. However, temperature dependence of the sensor itself is only a part of the problem. Smart Hall sensors allow to calibrate and correct for temperature dependencies of not only the sensor itself, but the whole magnetic circuit which contains permanent magnet and soft parts of the magnetic circuit (mainly the target itself, which is often toothed wheel) [28]. AMR sensors are more sensitive and immune to vibrations than Hall sensors. Thus they allow using cheaper or smaller permanent magnets and/or the increased distance between the sensor and the target.

References

1. P. Ripka (ed): *Magnetic sensors and Magnetometers*, Artech 2001.
2. P. Ripka, K. Záveta: *Magnetic sensors in J. Buschow (ed.): Handbook of Magnetic Materials*, Elsevier 2008.
3. R.S. Popovic: *Hall effect devices*, 2nd edition, CRC Press 2004.
4. P. Ripka, A. Tipek (eds.): *Modern Sensors Handbook*, ISTE 2007.
5. S. Tumanski: *Thin Film Magnetoresistive Sensors*, IoP Publ, 2001
6. P. Ripka, M. Vopálenský, A. Platil, M. Döscher, K.-M. H. Lenssen, H. Hauser: AMR magnetometer. *J. Magn. Magn. Mater.*, 254–255 (2003), 639–641.

7. O.V. Nielsen et al.: Development, construction and analysis of the 'Orsted' fluxgate magnetometer. *Meas. Sci. Technol.*, 6 (1995), 1099–1115.
8. M. Malátek, A. Platil, P. Ripka: GMI Current Sensor, Proc. Eurosensors XVII., Portugal, 2003, 418–419.
9. M. Malátek, A. Platil, A. Cerman, P. Ripka: Permalloy GMI Sensor. *J. Electr. Eng.*, 10/s, 2002, 195–198.
10. F. Primdahl, J.M.G. Merayo, P. Brauer, I. Laursen, T. Risbo: Internal field of homogeneously magnetized toroid sensor for proton free precession magnetometer. *Meas. Sci. Technol.* 16(2): 590–593, February 2005.
11. Budker D., M. V. Romalis: Optical magnetometry. *Nat. Phys.*, 3, 227–234 (2007).
12. J. Kubik, M. Janosek, P. Ripka: Low-power fluxgate sensor signal processing using gated differential integrator. *Sens. Lett.*, 5(1): 149–152 Mar 2007.
13. P. Ripka, M. Butta: Origin of the crossfield effect in AMR sensors, Proc. EMSA conference, 2008, to appear in *Sensors Letters*.
14. B.B. Pant, M. Caruso: Magnetic Sensor Cross-Axis Effect, AN 205, Honeywell.
15. J. Kubik, J. Vcelak, P. Ripka: On cross-axis effect of the anisotropic magnetoresistive sensors. *Sens. Actuators A*, 129, 15–19, 2006.
16. P. Ripka, S.W. Billingsley: Crossfield effect at fluxgate. *Sens. Actuators A: Phys.* 81, 176–179, 2000.
17. P. Brauer, J.M.G. Merayo, O.V. Nielsen, F. Primdahl, J.R. Petersen: Transverse field effect in fluxgate sensors. *Sens. Actuators A*, 59, 70–74, 1997.
18. Veerdonk R. J. M van de et al.: 1/f noise in anisotropic and giant magnetoresistive elements, *J. Appl. Phys.* 82, 6152–6164, 1997.
19. N.A. Stutzke, S.E. Russek, D.P. Pappas, et al.: Low-frequency noise measurements on commercial magnetoresistive magnetic field sensors. *Journal of Applied Physics* 97 (10): Art. no. 10Q107 Part 3 May 15 2005
20. M. Vopalensky, P. Ripka, A. Platil: Precise magnetic sensors. *Sens. Actuators A: Phys.* 106 (1–3): 38–42, 2003.
21. A. Cerman, A. Kuna, P. Ripka, J.M.G. Merayo: Digitalization of highly precise fluxgate magnetometers. *Sens. Actuators A*, 121/2, 421–429, 2005.
22. Ripka P: Current sensors using magnetic materials. *J. Optoelectr. Adv. Mater.*, 6(2): 587–592, 2004.
23. K. Draxler, Styblikova, R. Use of nanocrystalline materials for current transformer construction, *J. Magn. Magn.*, 157/158, 447–448, 1996.
24. Mlejnek P., P. Kašpar, K. Draxler: Measurement of ratio error and phase displacement of DC tolerant current transformers. *Sens. Lett.* 5, 289–292 2007
25. Current sensing for energy metering, William Koon, Analog Devices, Inc. www.analog.com
26. L. Dalessandro, N. Karrer, J. W. Kolar: High-performance planar isolated current sensor for power electronics applications. *IEEE Trans. Power Electron.* 22, 1682–1692, 2007.
27. J. Heremans: Narrow – gap semiconductor magnetic field sensors and applications. *Semiconductor Sci. Technol.* 8, S424–S430, 1993.
28. www.micronas.com

Modelling Electromagnetic Field Sensors

Ian Woodhead

Abstract Sensors may use the attenuation, velocity and electrical dispersion of electromagnetic waves as a direct or intermediate step in the transduction process. Modelling the sensor response provides the means to predict its spatial and temporal sensitivity and accuracy. The modelling may use differential or integral equation techniques. Each method has situations where its advantages dominate. As an example, a typical integral equation solution is described, along with discussion of cell size and shape considerations and selection of an appropriate basis function.

Keywords Electromagnetic modelling · integral equation · microwave sensor

1 Introduction

This chapter is concerned with the use of electromagnetic wave (EM) propagation to supply information for a sensor. Commonly, the sensor will use a change in attenuation and/or propagation time or phase, as a surrogate for the required measure. For example a change in reflected frequency provides, via Doppler shift, a measure of velocity; a change in moisture content and other material properties will be reflected in the propagation velocity and attenuation of an EM wave passing through the material; and the propagation velocity of a surface wave on an air-substrate interface, may represent the permittivity of the substrate. Here we discuss modelling EM wave propagation, outline differential equation and integral equation methods, and then provide an example solution using integral equations.

Ian Woodhead

Lincoln Ventures Ltd, Lincoln University, Canterbury, New Zealand, e-mail: woodhead@lvl.co.nz

2 Modelling Plane Wave Propagation

The propagation of EM waves in a material is described by Maxwell's equations. Assuming that the electrical properties of the material do not change during propagation, the two Maxwell equations that describe EM propagation may be written as

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad (1)$$

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \quad (2)$$

The right hand side of Eq. 2 comprises the conduction current due to the conductivity σ of the propagating medium, and the displacement current due to the energy storage or real component ε , of the permittivity. Solving these equations provide the means to determine from the EM wave propagation characteristics, the properties of the host material, i.e. ε , σ and μ , can be determined. In practice, modelling propagation and scattering is difficult, but can be simplified under certain conditions. Here we will describe modelling approaches by first imposing the constraint that propagation occurs along a transmission line or waveguide. Propagation is then classed as transverse EM mode or TEM propagation. We will define the transverse plane by the x and y Cartesian coordinates and the longitudinal coordinate by z . Then in the case of a guided electromagnetic wave propagating with zero loss, the z component of both \mathbf{E} and \mathbf{H} is zero. Considering just the x - z plane:

$$\frac{\partial^2 E_x}{\partial z^2} = \mu \varepsilon \frac{\partial^2 E_x}{\partial t^2} \quad (3)$$

Solutions of Eq. 3, the one-dimensional wave equation, lie on characteristic curves [1] defined by:

$$\frac{dz}{dt} = \frac{1}{\sqrt{\mu \varepsilon}} \quad (4)$$

The propagation velocity v defined by Eq. 4 is independent of the rate of change of E with respect to time, and hence independent of the frequency of the propagating signal or of the slope of a voltage pulse or its frequency spectrum as represented by a Fourier series.

The zero loss case is quite restrictive, but if the medium has a small loss so that $\sigma \ll \omega \varepsilon$ where ω is the angular frequency, the conduction current contribution in Eq. 2 may be ignored, and Eq. 3 adequately describes the wave propagation. The complex permittivity that incorporates the loss component ε'' is $\varepsilon' - j\varepsilon''$, and may be replaced by $\varepsilon' - j\sigma/\omega$. The propagation velocity in the low-loss case is then adequately represented by a truncated series expansion (e.g. from [2]):

$$v \approx \left[\mu \varepsilon \left(1 + \frac{\sigma^2}{8\varepsilon^2 \omega^2} \right) \right]^{-0.5} \quad (5)$$

In considering loss in the axial direction, even when considering the skin effect resistance of a parallel waveguide, the real or loss component (resistance) within the conductor is small compared with the imaginary or inductive component (i.e. $R \ll \omega L$ where R and L are respectively the distributed resistance and inductance per unit length of the waveguide). The consequence of the small real loss in the axial direction is negligible contribution to the integral of $E dl$ by the axial magnetic field and negligible contribution to the transverse magnetic field by the displacement current. The negligible axial components of \mathbf{E} and \mathbf{H} enable the use of Eq. 4 to describe the TEM propagation with zero or low-loss permittivity, and quasi-static analysis to represent the transverse fields.

If we impose two more conditions, that the material is non-magnetic, i.e. has permeability μ_0 (the lossless, free space value), and ε_r is a function of the transverse coordinates (x, y) only, i.e. ε_r has no z dependence, then Maxwell's first equation $\nabla \cdot \mathbf{D} = \rho$ becomes;

$$\nabla \cdot (\varepsilon_0 \varepsilon_r(x, y) \nabla \Phi) = -\rho \quad (6)$$

The domain of Eq. 6 may include the source and sink of electric flux, the waveguide. In general, Eq. 6 cannot be solved analytically but needs to be converted to a numerical form. Two approaches for solving Eq. 6 are differential equation methods such as the finite difference (FD) and finite element methods (FEM), and integral equation techniques. Numerical approaches involve a discretisation stage where the differential or integral transforms are represented by discrete steps or summations. Both forms are equivalent since from Eq. 6:

$$\nabla \Phi = \frac{-1}{\varepsilon(x, y)} \mathbf{L}^{-1} \rho(x, y) \quad (7)$$

where \mathbf{L}^{-1} is the inverse of differential operator ∇ . Assuming \mathbf{L}^{-1} is adequately represented by an integral with kernel \mathbf{G} , we obtain the elementary integral form:

$$\begin{aligned} -\nabla \Phi(r) &= \mathbf{E}(r) \\ &= \frac{1}{\varepsilon(x, y)} \int \int \int_{\text{region}} \mathbf{G}(r) \rho(x, y) dr \end{aligned} \quad (8)$$

where r is a 3-space dimension, and $\mathbf{G}(r)$ is a Green's function that incorporates the boundary conditions of the problem.

3 Differential Equation Methods

The two dominant differential equation (DE) methods are finite difference (FD) and finite element (FEM) methods. FD methods have been used for a range of potential field problems, and use a rectangular grid to discretise the region of interest. An approximate solution to the potential function is defined at each mesh node by

expressing the potential function as a difference equation, expressed in terms of the potential at the neighbouring nodes.

The Taylor representation of the derivatives is commonly truncated so that terms in h^2 and higher (where h is the mesh size) are ignored. The FD scheme generates a system of equations with a sparse, banded coefficient matrix, since the potential at each mesh node is represented in terms of the nearest neighbours.

FEMs represent the potential within an element, in turn defined by a basis function over the area of the element, in terms of the potential values at the vertices of the elements. Triangular elements are commonly used, and enable better modelling of complex shapes than the rectangular grids of the FD method, and easily accommodate increased accuracy by the addition of nodes at the mid points of each boundary segment. Further, the mesh size may be altered over the problem domain to provide higher resolution in areas of particular interest.

These DE methods are fast and applicable to two dimensions, but the boundary conditions must be included, so they often require the calculations to extend beyond the region of interest, to cover the entire domain of the problem. This generally means calculating to a boundary selected so that the influence of the boundary field on the result is small, and then imposing boundary conditions (typically Dirichlet boundary conditions where the electric field potential $\Phi(r) = 0$). Although the DE methods result in the construction of a large matrix spanning the problem domain, the matrix is banded thus allowing the use of sparse matrix solution methods.

4 Integral Equation Methods

IE methods generally produce smaller field matrices than the DE methods, but are full matrices. In addition, IE methods used with inhomogeneous dielectric distributions require a volume integral equation approach and hence calculation in 3-D, even if there is invariance in one direction.

Equation 8 is a Fredholm integral equation of the first kind, and inherently ill-posed since any steep gradient in the homogeneous term ($\mathbf{E}(r)$ in this case) can only arise through near singular values of the kernel $\mathbf{G}(r)$ or the function $\rho(x,y)$. Here, $\mathbf{G}(r)$ has so-called $1/r$ or self-term singularities where the field and source points coincide so that $r = 0$. Methods for solving Eq. 8 circumvent the ill-posedness by the use of techniques such as the Galerkin method (identical weighting and basis functions) that approximate the ill-posed form [3], in a similar manner to the approximate or discrete representations of the DE methods

IE methods have been employed for 3-D EM modelling because of their better numerical efficiency, although FD methods are have had a resurgence following the use of staggered grids [4], and the flexible topology of the FEM allows it to accurately model complex shapes. Nevertheless, IE methods continue to be used for 3-D EM modelling for homogeneous problems and when the Green's functions may be easily calculated. In general, IE methods have advantages of better handling of boundary conditions (they are defined within the IE) so only the anomalous or non

free-space region needs to be solved. Hence they tend to be efficient for modelling areas bounded by free space. The numerical efficiency tends to be worse however, particularly for large numbers of cells or elements.

Interestingly, even when the electrical properties of the medium are altered, only diagonal elements of the IE matrix are affected; the off-diagonal elements only depend on the geometry of the anomalous region. These features have a significant benefit where the modelling forms part of tomographic inversion, and where the forward calculation is repeated using the same physical properties of the region, but with an altered impressed field.

Solving IEs numerically follows a discretisation and summation procedure, and the most common technique is the moment method (MM), as described by Harrington [5].

5 An Example Integral Equation Solution

Here we describe an example IE solution, and discuss some issues relating to problem formulation and accuracy. Specifically, we will describe a model of the field of a parallel transmission line (PTL) within an arbitrary inhomogeneous permittivity distribution.

Initially, we define a region in which a solution is sought, and ascribe particular EM properties, although the solution that will be derived may also apply to more general regions. The region is a Hilbert space (a linear vector space with inner product), defined by 3-D Cartesian coordinates (x, y, z) represented by r . Throughout the region and beyond, the permeability is the lossless, free space value μ_0 , and outside the region, ε is the lossless, free space value ε_0 . The region includes an impressed electric field that is fully defined by the quasi-static EM field of a PTL that is enclosed within the region (i.e. there is no other macroscopic electric field either from an external source or an internal permanent polarization field). For the immediate solution, we further constrain the properties of the region and the PTL to have zero or very small loss. Hence:

1. the PTL wires have zero or near-zero resistance (true for typical wires)
2. ε within the region is lossless, or at least $\tan \delta$ is small (true for many composite materials).

Given an impressed electric field $\mathbf{E}_i(r)$ in the region defined above with arbitrary $\varepsilon(r)$, the total resultant electric field distribution $\mathbf{E}(r)$ is:

$$\mathbf{E}(r) = \mathbf{E}_i(r) + \mathbf{E}_P(r) \tag{9}$$

$\mathbf{E}_P(r)$ is the polarization field that is generated by those parts of the region with non-zero susceptibility and in response to $\mathbf{E}_i(r)$. Hence, where the polarization of the region is defined by $\mathbf{P}(r)$ and since:

$$\mathbf{E}(r) = \frac{\mathbf{P}(r)}{\varepsilon_0\chi(r)} \quad (10)$$

then

$$\frac{\mathbf{P}(r)}{\varepsilon_0\chi(r)} = \mathbf{E}_i(r) + \mathbf{E}_P(r) \quad (11)$$

or

$$\begin{aligned} -\mathbf{E}_i(r) &= \mathbf{E}_P(r) - \frac{\mathbf{P}(r)}{\varepsilon_0\chi(r)} \\ &= K(\mathbf{P}) \end{aligned} \quad (12)$$

Here, K is a linear operator acting on the polarization \mathbf{P} , \mathbf{E}_i the external impressed field and $\chi(r)$ is the electric susceptibility ($\varepsilon_r(r) - 1$). The polarization region may now be discretised to enable calculating the matrix of polarization vectors $\mathbf{P}(r)$:

$$[K][\mathbf{P}] = -[\mathbf{E}_i] \quad (13)$$

so that

$$[\mathbf{P}] = -[K]^{-1}[\mathbf{E}_i] \quad (14)$$

To extract the vector of electric field strengths we utilize Eq. 10 so that:

$$[E] = \left[\frac{\mathbf{P}(r)}{\varepsilon_0\chi(r)} \right] \quad (15)$$

Hence if \mathbf{P} , \mathbf{E}_i and K are known, the field can be calculated.

The polarization of a discretised zone or cell within a dielectric material may be represented by a dipole at its geometric centre, quantified by the dipole moment $\mathbf{P} = \hat{r}ql$ where q is the dipole charge magnitude and l the separation distance (Fig. 1).

The total potential Φ due to a dipole resulting from the superposition of the field from one charge of the dipole, and the Taylor expansion of the opposite charge, a distance l from the first is:

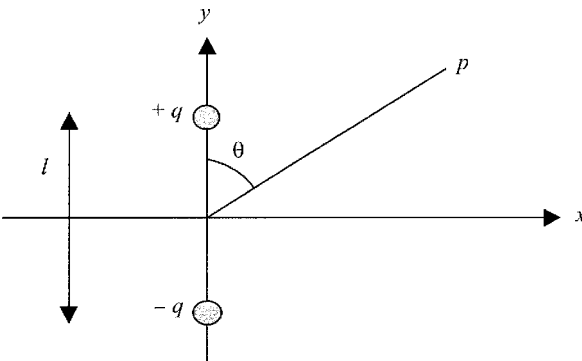


Fig. 1 Cartesian geometry for calculation of the field of a dipole

$$\Phi = -\frac{\partial}{\partial z} \left(\frac{q}{4\pi\epsilon_0 r} \right) l + \frac{1}{2!} \frac{\partial^2}{\partial z^2} \left(\frac{q}{4\pi\epsilon_0 r} \right) l^2 + \dots \quad (16)$$

Then since

$$\frac{\partial}{\partial z} \left(\frac{1}{r} \right) = \frac{-z}{r^3} \quad (17)$$

and

$$\frac{\partial^2}{\partial z^2} \left(\frac{1}{r} \right) = \frac{-1}{r^3} \left(1 - \frac{3z^2}{r^2} \right) \quad (18)$$

hence

$$\Phi = \frac{ql}{4\pi\epsilon_0 r^3} \left(z - \frac{l}{2} + \frac{3lz^2}{2r^2} \right) + \text{higher terms} \quad (19)$$

and since $z = r \cos \theta$ and $\mathbf{P} \cos \theta = \mathbf{P} \cdot \hat{\mathbf{r}}$:

$$\Phi = \frac{\mathbf{P} \cdot \hat{\mathbf{r}}}{4\pi\epsilon_0 r^2} \left(1 - \frac{l}{2r \cos \theta} + \frac{3l \cos \theta}{2r} + \dots \right) \quad (20)$$

$l \ll r$ so that all but the first term vanishes.

In most dielectric materials, there is no net polarization until it is imposed by an external or impressed field. The MM may be considered as the summation in each cell, of the electric field contributions due to the polarization in all other cells. Cubes are a convenient cell shape for a Cartesian coordinate system, but cells with trapezoidal or hexagonal cross-sections are viable but less convenient alternatives.

Consider the influence of a portion (U) of polarized material on the potential at point f . The total polarization of U can be represented by a dipole at the centre of the region, and with dipole moment \mathbf{P} (Fig. 2).

From electrostatic theory, the potential at f is:

$$\begin{aligned} \Phi_f &= \frac{1}{4\pi\epsilon_0} \frac{\mathbf{P} \cdot \hat{\mathbf{r}}}{r^2} \\ &= \frac{1}{4\pi\epsilon_0} \frac{\mathbf{P} \cdot \mathbf{r}}{r^3} \end{aligned} \quad (21)$$

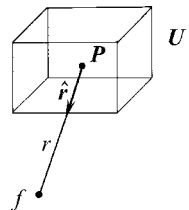


Fig. 2 Geometry of a region of polarization within a polarized material

where $\hat{\mathbf{r}}$ is a unit vector pointing from the centre of U to f . The electric field is $\mathbf{E}_f = -\nabla\Phi_f$ and using the superposition theorem, the total field at s due to contributions from all field points is:

$$\mathbf{E}_f = \int_{\text{all } U} -\nabla \cdot \frac{1}{4\pi\epsilon_0} \frac{\mathbf{P} \cdot \mathbf{r}}{r^3} d\mathbf{v} \quad (22)$$

Since the region is a linear vector space (was previously defined as a Hilbert space):

$$\mathbf{E}_f = -\frac{1}{4\pi\epsilon_0} \nabla \int_{\text{all } U} \frac{\mathbf{P} \cdot \mathbf{r}}{r^3} d\mathbf{v} \quad (23)$$

The region may then be discretised to obtain the required \mathbf{E}_P of Eq. 12.

$$\begin{aligned} \mathbf{E}_P &= \mathbf{E}_f \\ &= -\frac{1}{4\pi\epsilon_0} \nabla \sum_{\text{all } U} \frac{\mathbf{P} \cdot \mathbf{r}}{r^3} \Delta\mathbf{v} \end{aligned} \quad (24)$$

where $\Delta\mathbf{v}$ is the volume of each subregion or cell, U . Since

$$\begin{aligned} \hat{\mathbf{r}} &= \frac{\mathbf{r}}{r} \\ &= \frac{\mathbf{x} + \mathbf{y} + \mathbf{z}}{r} \end{aligned} \quad (25)$$

it follows that:

$$\nabla \frac{\mathbf{P} \cdot \mathbf{r}}{r^3} = \nabla \frac{\mathbf{P} \cdot (\mathbf{x} + \mathbf{y} + \mathbf{z})}{r^3} \quad (26)$$

and since $\frac{dr}{dx} = \frac{x}{r}$ and $\mathbf{x} = x\hat{\mathbf{x}}$, the partial derivatives of Φ are:

$$\Phi_x = \frac{\mathbf{P}}{4\pi\epsilon_0 r^5} \cdot (\hat{\mathbf{x}}(r^2 - 3x^2) - \hat{\mathbf{y}}3xy - \hat{\mathbf{z}}3xz) \quad (27)$$

$$\Phi_y = \frac{\mathbf{P}}{4\pi\epsilon_0 r^5} \cdot (\hat{\mathbf{x}}3yx - \hat{\mathbf{y}}(r^2 - 3y^2) - \hat{\mathbf{z}}3yz) \quad (28)$$

$$\Phi_z = \frac{\mathbf{P}}{4\pi\epsilon_0 r^5} \cdot (\hat{\mathbf{x}}3zx - \hat{\mathbf{y}}3zy - \hat{\mathbf{z}}(r^2 - 3z^2)) \quad (29)$$

Since for Cartesian coordinates, $\nabla\Phi = \hat{\mathbf{x}}\Phi_x + \hat{\mathbf{y}}\Phi_y + \hat{\mathbf{z}}\Phi_z$, the electric field at f due to polarization at a source s , in dyadic form is:

$$\mathbf{E}_f = \frac{-\mathbf{P}}{4\pi\epsilon_0 r^5} \cdot \begin{bmatrix} \hat{\mathbf{x}}\hat{\mathbf{x}}(r^2 - 3x^2) - \hat{\mathbf{x}}\hat{\mathbf{y}}3xy - \hat{\mathbf{x}}\hat{\mathbf{z}}3xz \\ -\hat{\mathbf{y}}\hat{\mathbf{x}}3yx + \hat{\mathbf{y}}\hat{\mathbf{y}}(r^2 - 3y^2) - \hat{\mathbf{y}}\hat{\mathbf{z}}3yz \\ -\hat{\mathbf{z}}\hat{\mathbf{x}}3zx - \hat{\mathbf{z}}\hat{\mathbf{y}}3zy + \hat{\mathbf{z}}\hat{\mathbf{z}}(r^2 - 3z^2) \end{bmatrix} \quad (30)$$

In calculating this expression, r is the distance between a field point f and source point s , defined as $r = \sqrt{x_{fs}^2 + y_{fs}^2 + z_{fs}^2}$, where x_{fs} , y_{fs} and z_{fs} are the relative distances between the source and field points in each direction of the 3-D Cartesian space. Hence the total field at f due to all polarizations in U is:

$$\mathbf{E}_f = \sum_{\text{all } U} \left(\frac{-\mathbf{P}}{4\pi\epsilon_0 r^5} \cdot \begin{bmatrix} \hat{\mathbf{x}}\hat{\mathbf{x}}(r^2 - 3x^2) - \hat{\mathbf{x}}\hat{\mathbf{y}}3xy - \hat{\mathbf{x}}\hat{\mathbf{z}}3xz \\ -\hat{\mathbf{y}}\hat{\mathbf{x}}3yx + \hat{\mathbf{y}}\hat{\mathbf{y}}(r^2 - 3y^2) - \hat{\mathbf{y}}\hat{\mathbf{z}}3yz \\ -\hat{\mathbf{z}}\hat{\mathbf{x}}3zx - \hat{\mathbf{z}}\hat{\mathbf{y}}3zy + \hat{\mathbf{z}}\hat{\mathbf{z}}(r^2 - 3z^2) \end{bmatrix} \right) \quad (31)$$

Matrix K in Eq. 14 contains a set of components for each field point. Each set consists of contributions from a source point representing a region of polarization, U . This results in an $F \times S$ by $F \times S$ matrix of sets of entries, where F and S are respectively the number of field and source points. For the 3-D Cartesian space, each set comprises nine components representing the three coordinate directions of the field, in response to the three coordinates of the source field, nine elements in total.

Then given the impressed field from the PTL (see [6] for example), the resultant surrounding field distribution may be calculated.

6 Basis Functions

Conventionally, pulse or delta basis functions are used for IE methods. They are simpler to implement and have provided satisfactory solutions. However for certain applications, increased accuracy has also been reported [7] by the use of higher order basis functions that prevent the sharp field discontinuity at cell boundaries.

Linear or higher order basis functions provide for a smoother transition of the field across cell boundaries and a corresponding reduction in modelling errors (they provide a continuous field across cell boundaries, but discontinuous first derivatives). Further, linear basis functions, for example through increased accuracy and solution stability/convergence, allow a coarser discretisation to be employed. Countering the above advantages of 2nd, 3rd or higher order basis functions is the increased complexity and calculation time of the scattering matrix. The field matrix requires significant resources to calculate, and in general, reducing the number of cells n is very beneficial, even at the cost of increased calculation time for each cell (doubling calculation time for each cell is typically equivalent to a reduction in n of 20%).

Pulse or delta basis functions provide a one to one correspondence between the polarization in each cell and the corresponding integral operator. If linear or higher order basis functions are used, then the polarisation P in Eq. 14 is represented by lines in each dimension, and these are continuous throughout the discretised region, unlike the representation using pulse basis functions.

Consider the MM applied to determining the polarization in the anomalous region within a 2-D space defined by x, y . Then:

$$L(\mathbf{P}(x,y)) = \mathbf{E}_i(x,y) \quad (32)$$

where in this case L is the linear operator defined by Eq. 7 and $\mathbf{P}(x,y)$ the unknown polarization. Then a series solution is:

$$\mathbf{P}(x,y) = \sum_n \boldsymbol{\alpha}_n(x,y) P_n \quad (33)$$

where $\boldsymbol{\alpha}_n$ are unknown constants representing the polarization at each point n in the x - y plane, and P_n are basis functions. \mathbf{P} may be expressed in terms of delta or pulse basis functions:

$$\mathbf{P} = \sum_n \delta_n [\hat{\mathbf{x}}(a_n(x-x_n) + b_n(y-y_n)) + \hat{\mathbf{y}}(c_n(x-x_n) + d_n(y-y_n))] \quad (34)$$

where δ is the Dirac delta function, n is an index representing the cell number, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are unit vectors that represent the coordinate directions of the polarization, $a_n \dots d_n$ are unknown coefficients, and x_n and y_n are the coordinates of the cell centroid. From Eq. 32:

$$\sum_n P_n L(\boldsymbol{\alpha}_n(x,y)) = \mathbf{E}_i(x,y) \quad (35)$$

Weighting functions, w_1, w_2, \dots that lie within the range of L are used to define an inner product for $L(\boldsymbol{\alpha}_n)$, $\langle w_m, L(\boldsymbol{\alpha}_n) \rangle$, and for the inhomogeneous term \mathbf{E}_i , $\langle w_m, \mathbf{E}_i \rangle$.

If the Galerkin method of choosing w_m equal to P_n is employed then we may formulate as:

$$\sum_n \delta_n L(\delta_n(\boldsymbol{\alpha}_n(x,y))) = \delta_n(\mathbf{E}_i(x,y)) \quad (36)$$

In matrix form, Eq. 36 may be written:

$$[l_{mn}][\boldsymbol{\alpha}_n] = [\mathbf{E}_i(x,y)] \quad (37)$$

so that:

$$[\boldsymbol{\alpha}_n] = [l_{nm}^{-1}][\mathbf{E}_i(x,y)] \quad (38)$$

If linear or higher order basis functions are employed, the one to one correspondence of the Galerkin approach is replaced by an interdependency. A simple example of this approach has been described [7] using linear basis functions in a MM solution for electric current in a 1-D cylinder excited by a plane wave:

$$\mathbf{I}(x_m < x < x_{m+1}) = \mathbf{I}(x_m) + \frac{(x-x_m)[\mathbf{I}(x_{m+1}) - \mathbf{I}(x_m)]}{(x_{m+1} - x_m)} \quad (39)$$

When similarly applied to determining the polarization, there are additional unknowns so that \mathbf{P} is:

$$\mathbf{P} = \sum_n \delta_n \left[\hat{\mathbf{x}}(a_n + b_n(x-x_n) + c_n(y-y_n)) + \hat{\mathbf{y}}(d_n + e_n(x-x_n) + f_n(y-y_n)) \right] \quad (40)$$

and the delta function is retained to define the range of each $(x - x_n)$. Hence there are six unknowns for each cell with linear basis functions compared with four in the case of delta functions. In the latter case, the physical interpretation is that the four unknowns represent the electric field in each of the two coordinate directions, in response to the two coordinate directions of the polarization in a source cell. For linear basis functions, the polarization in the source cell is represented by a plane, in turn defined by three coefficients for each of the two components (x and y) of the polarization in a source cell. However in this case, Eq. 40 possesses redundant information. Consider instead a so called ‘rooftop’ function where the same number of data as used in the pulse basis functions is used to provide a linear interpolation between cell midpoints, as used by Lu and Chew [8] for example. For a particular cell n , it may be shown that the value of the function P over the length of the cell is:

$$P = 3/4xp_n + 1/8xp_{n-1} + 1/8xp_{n+1} \quad (41)$$

where x is the cell length and p_n the value of P at the centre of the cell. This may be extended to 2-D so that:

$$P = 3/4xp_{n,m} + 1/16xp_{n-1,m} + 1/16xp_{n+1,m} + 1/16xp_{n,m-1} + 1/16xp_{n,m+1} \quad (42)$$

This forms a weighted, 2-D spatial average of cells in the $x - y$ plane, and may be calculated using a convolution, which in discrete form represents polynomial multiplication.

When pulse basis functions are used, the field is recovered from the polarisation using $\mathbf{E} = \mathbf{P}/\epsilon_0\chi$. However, the use of linear or higher order basis functions applied to \mathbf{P} , needs to be reflected in χ to obtain appropriate values of \mathbf{E} , suggesting that the Galerkin approach should be retained.

7 Cell Shape and Size

Cell shape and size affect model accuracy, particularly at boundaries. Cubic cells are convenient to use with MM solutions and Cartesian coordinates, but are less well suited than polyhedra to fit smooth and curved boundaries. Tsai et al. [9] used triangular patches to model the surface charge term and cubes for volume integration applied to a MM solution to the EM scattering problem for an arbitrary dielectric. Good agreement was obtained with the Mie [10] analytical solution for a dielectric sphere. Note that there is little point in cell subdivision to improve edge accuracy, unless the edge is also altered in shape (such as by the addition of intermediate cells) to better reflect the shape of the modelled object.

Guo et al. [11] compared the results from numerical modelling with the Mie result for scattering by a dielectric sphere. They concluded that field errors are mostly due to the geometric difference between the dielectric sphere and its discretised

model, and that the error due to volume mismatch dominated over the shape mismatch error. This was observed to be consistent with the variation of field error with discretisation. They tried border smoothing by setting the boundary cells at half the border ε_r , but this was unsuccessful in reducing the oscillatory nature of field error with discretisation. Consequently, they advised smoothing the geometry rather than ε_r .

Cell shape and size are somewhat interdependent in that an inappropriate shape, such as the cube that is most appropriate for Cartesian coordinates, may be compensated for to some extent by decreasing cell size.

8 Wavelength Considerations

An important consideration when formulating the model of a physical system is the size of the scatterer and its sub sections or cells in relation to the wavelength of the illuminating source. Glisson [12] noted that IE methods fail when the wavelength is small compared with the size of scatterer, and that failure also occurs at internal resonance frequencies. Guo et al. [11] and others have stated that the cell dimensions must be less than $\lambda/4$, but Zhou and Shafai [13] claim that for the EM scattering problem and using a MM, the cell size should be no more than $\lambda/10$ to ensure good far-field prediction, a notion supported by Peterson and Klock [14].

9 Convergence and Execution Considerations

Given an appropriate selection of basis functions and cell size and shape for a MM IE solution, a method may fail or perform poorly due to an insufficient number of cells to adequately represent the physical model, in turn demanding larger matrices. Since EM modelling is a computationally intensive process, methods are commonly sought to enable the size of the problem to be reduced. We have already discussed selection of basis functions and cell size and shape to optimise the geometry and minimise the number of discrete cells required to define the space that is being modelled. One general drawback of IE methods is that, as in this case, volume integration is required, which inherently makes the field matrix large. Reduction of problem size may be used in a volume IE problem if there is no variation in one dimension. For transverse EM propagation for example, a quasi-static solution may suffice and reductions in dimensionality may be applied (e.g. [6]) with quite dramatic improvements in matrix size and execution speed.

A slightly different approach is used for geophysical EM modelling of earth structures which frequently uses 2-D models where the excitation source is an axially-invariant line charge [15]. Point source or dipole excitations are more amenable to subsurface investigations using bore holes, and the combination of the 3-D source and 2-D scatterer is commonly referred to as a quasi 2-D or 2.5-D

problem. Consequently, the geophysical 2.5-D problem remains essentially 2-D, but accounts for a 3-D source.

10 Proximity Compensation

Implicit in the above formulation of the point matching method described by Eq. 31 is the $1/r^3$ dependence for calculation of the far field contribution from the dipole at the centre of each cell. The point matching or pulse basis function approach assumes an adequate representation of polarisation by a single dipole at the centre of each cell. This approximation is valid for large r , that is for distant cells. When r is small, errors become more significant due to the $1/r^3$ dependency. A more accurate approach is to calculate the polarisation by integration of the uniform distribution of dipoles over the volume of the cell. Rather than reformulating the entire model to reduce modelling errors induced by this assumption, the error may be compensated for (e.g. [6]).

11 Conclusion

In this Chapter, a background to EM modelling for sensors that use EM fields to determine material properties, velocity or position. Both DE and IE methods have been outlined, and an example IE solution using the MM for the field of a transmission line has been described, along with a discussion of simplifications and factors that affect how the model is constructed, its accuracy and execution speed, such as cell size and shape.

References

1. Williams, W. E., 1980. *Partial Differential Equations*, P. 70 (Oxford: Oxford University Press)
2. Ramo, S., Whinnery, J. R., Van Duzer, T., 1993. *Fields and Waves in Communication Electronics*, P. 285 (New York: Wiley)
3. Tsukerman, I., 1997. Stability of the moment method in electromagnetic problems. *IEEE Transactions on Magnetics* 33:1402–1405
4. Weidelt, P. 1995. Three-dimensional conductivity models: implications of dielectric anisotropy. *An International Symposium on Three-Dimensional Electromagnetics*, Schlumberger-Doll Research, Ridgefield, CT
5. Harrington, R. F., 1968. *Field Computation by Moment Methods* (R E Kreiger)
6. Woodhead, I. M., Buchan, G. D., Platt, I. G., Christie, J. H., 2007. Improved electric field modelling for TDR, *Meas. Sci. Technol.* 18 1110-1117 doi:10.1088/0957-0233/18/4/020
7. Qian, W., Boerner, D. E., 1995. Basis functions in 1D EM integral equation modelling *Three-Dimensional Electromagnetics* Ridgefield, CT

8. Lu, C. C., Chew, W. C., 2000. A coupled surface-volume integral equation approach for the calculation of electromagnetic scattering from composite metallic and material targets. *IEEE Transactions on Antennas and Propagation* 48: 1866–1868
9. Tsai, C.-T., Massoudi, H., Durney, C. H., Iskander, M., 1986. A procedure for calculating fields inside arbitrarily shaped, inhomogeneous dielectric bodies using linear basis functions with the moment method. *IEEE Transactions on Microwave Theory and Techniques* MTT-34:1131–1139
10. Mei, K. K., Van Bladel, J. G., 1963. Scattering by perfectly-conducting rectangular cylinders. *IEEE Transactions on Antennas and Propagation* AP11:185–192
11. Guo, T. C., Guo, W. W., Oguz, H. N., 1993. A technique for three-dimensional dosimetry and scattering computation of vector electromagnetic fields. *IEEE transactions on magnetics* 29:1636–1641
12. Glisson, A. W., 1989. Recent advances in frequency domain techniques for electromagnetic scattering problems. *IEEE Transactions on Magnetics* 25:2867–2871
13. Zhou, R., Shafai, L., 1997. Study on the convergence of volume integral equation method. *Antennas and Propagation Society International Symposium 1997* 3:1830–1833
14. Peterson, A. F., Klock, P. W., 1988. An improved MFIE formulation for TE-wave scattering from lossy, inhomogeneous dielectric cylinders. *IEEE Transactions on Antennas and Propagation* 36:45–49
15. Torres-Verdin, C., Habashy, T. M., 1994. Rapid 2.5-dimensional forward modeling and inversion via a new nonlinear scattering approximation. *Radio Science* 29:1050–1079

Dielectric Characterization of Biological Tissues: Constraints Related to Ex Vivo Measurements

Mustapha Nadi

Abstract Electrical Impedance Spectroscopy (EIS) has been previously reported as a technique for non-invasive assessment of electromagnetic tissue properties. In the frequency range up to 10 MHz, current conduction through tissue is mainly determined by the tissue structure, i.e. the extra- and intra-cellular compartments and the insulating cell membranes. Therefore, changes in the extra- and intra-cellular fluid volumes are reflected in the impedance spectra. E.I.S. systems include electrodes for the measurement of the impedance. Different electrodes configurations are used to measure bioelectric phenomenon for both macroscopic and microscopic approaches. Electrodes for macroscopic characterization are used for bio impedance measurement of a great biological tissue sample or organ. In this paper, we briefly review and discuss metrological aspects relating to electrical characterization of biological tissues based on the difficulty to compare between different author's results. Experimental results obtained on different kinds of biological tissues (blood and bone) are presented and discussed as examples according to the influencing constraints specific to their physiological nature.

Keywords Bioimpedance · dielectric properties · blood · bone · microelectrodes · biological sensor · electromagnetic dosimetry · impedance spectroscopy

1 Introduction

Interest in the electromagnetic (EM) properties of biological tissues began about more than one century ago. The first measured quantities were the resistance and the capacitance of ex vivo samples [1]. Interactions between electromagnetic field and biological tissue are applied in many therapeutic and diagnosis methods. Biomedical apparatus based on EM fields increased in the last decades while development of devices radiating electromagnetic field, like mobile phone, led to the questions

Mustapha Nadi
Nancy University, L.I.E.N., BP 239 Faculty of Sciences and Techniques,
54506 Vandoeuvre les Nancy, e-mail: mustapha.nadi@lien.uhp-nancy.fr

of their possible biological effects. These so called EM bio effects have led to models for EM dosimetry simulation. These simulations are strongly depending on the knowledge of the dielectric properties of the human body. The study of these interactions requires the definition of the source on one hand and the characterization of the target that is the biological tissue on the other. Indeed, the attempt to evaluate the induced currents *in situ* and, thus, the field distribution in the tissue depends on its dielectric characteristics. The electromagnetic properties of biological tissues are thus fundamental parameters that are obviously necessary for any research in these domains. Electrical conductivity and relative permittivity may be deduced from the measured impedance of the biological tissue of interest as an average value for the considered sample of tissue.

The use of electrical impedance spectroscopy (EIS) is one of the no-destructive, low-invasive and most promising techniques for biological tissues characterization. This so called bio impedance is now a well known tool for characterizing different physiological quantities like fat content [2]. Other applications of bio impedance were and are still developed in many clinical or domestic fields [3, 4].

Many techniques and methods for measuring the bio impedance exist and are primarily motivated by the frequency band of interest [1, 5]. Electrodes for microscopic characterization were recently developed for biological cell or very small cells aggregate. From the main constraints are the control of the interface between the sensor and the biological matter and the calibration process. As example of the parameters influencing the measures, the temperature is one whose effects on the variations of the conductivity and permittivity have been studied by many authors. This is not the only crucial factor and other parameters had to be taken into account for reliability of the experimental results.

Several authors have worked on dielectric characterization of biological matter and bio impedance; one of the most involved in this field, following Fricke and other pioneers [2, 4] has been Herman Schwan* [6, 7] at the University of Pennsylvania. Up to now, the values provided by the literature are not always reliable since these data are sparse and disseminated over the frequency range with great variations even for a same biological organ. Some authors have proposed a review of the existing electric properties of biological material. An example is the compendium published by Geddes [8] or the more exhaustive compilation by Gabriel et al. [9]. These data are the most used by many research teams. The differences and variations due to biological species or organ under test as well as to the metrological difficulties related to the interface between the sensor and the biological tissue. Providing and maintaining a data base of the existing published values faces a methodological problem because comparative studies are not always well-documented.

In this paper, the major constraints related to metrological aspects of bio impedance spectroscopy measurements are summarized. Metrological specificity for *ex-vivo* is presented according to the medium under test and the influencing parameters. To illustrate the previous outlined constraints, electric properties of bone and blood will serve as examples by comparing between different author's results. Experimental results obtained on human and animals *ex vivo* samples are presented.

*(1915–2005)

The differences are discussed based on the experimental conditions, their natures and other constraints summarized below.

2 Metrological Aspects of Bioimpedance Spectroscopy

Electrical Impedance Spectroscopy (EIS) is a well-established technique for dielectric tissue characterisation [2, 5, 10]. Basically, in its most classical form, the electrical bio impedance measuring technique consists in measuring the voltage induced by a current injected across the sample under test. For a reasonable current density, the behaviour is linear and thus the Ohm's law is valid (Fig. 1). For the frequency range below ten Megahertz, the sample is placed between two electrodes forming a loss capacitor. The impedance is measured by the two points or four points techniques [1, 2, 11]. For frequencies higher than about 10 MHz the impedance is deduced from measuring the reflection coefficient in a coaxial line or wave guide loaded by the sample under test [2, 4, 5, 12].

The measures on biological tissue are particular for two reasons: firstly because of variations between the samples under test, on the other hand because of the diversity of influencing factors and the non-availability of references for the sensor calibration [1, 5, 13].

The dielectric behaviour of biological tissues is dependent on their nature and on the frequency of interest. They are deformable, heterogeneous, anisotropic and may be solid or liquid. Spectroscopic measurement permits to carry out their values according to each intended application or goal. At microscopic level, the linearity linked to the dispersion of tissue dielectric properties becomes also a limiting factor. Figure 2 shows an example of a basic electrical model proposed to modelize intra and extra cellular medium and the membrane capacitance [2, 4, 10].

Moreover, the measurements taken from a sample are likely to be affected by multiple influencing factors such as the contact with the membrane, the temperature at which the measurement is performed, the pressure on the sample, the post mortem duration or the structure of the tissue.

Bio impedance measurements need to distinguish the influencing factors related to biological aspects and those related to metrology and instrumentation. Precautions taking account of these aspects are required to ensure metrological reliability of the results.

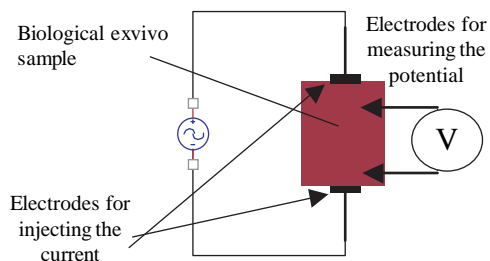
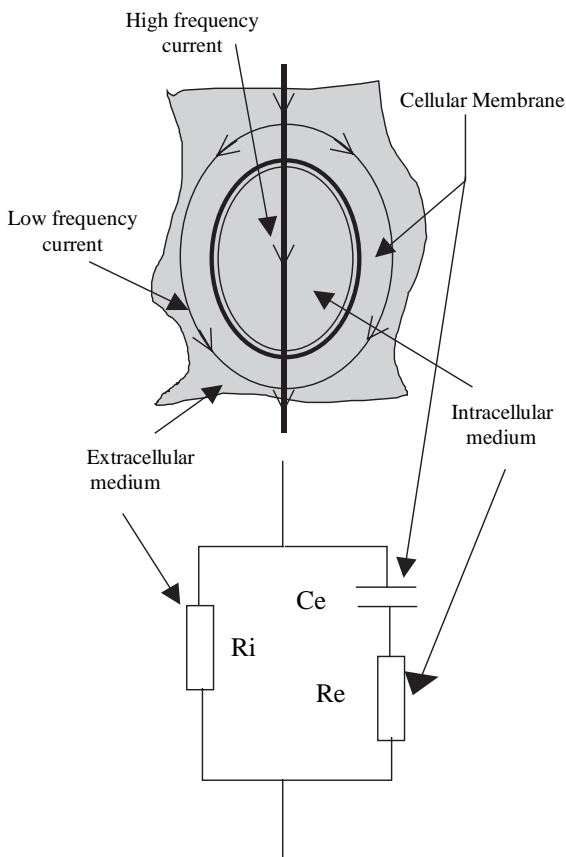


Fig. 1 Basic principle of bio impedance measurement

Fig. 2 Circuit model for biological cell electrical properties



3 Measurement Constraints Related to the Bioimpedance Sensor

The frequency, the current density, the material of electrodes, the geometry of the sensor, the temperature are among the main physical and instrumental factors that influence the measure. Systematic error are mainly due to two main sources : electrode polarization and parasitic inductance due to the connections which appears at the lower and higher ends of the frequency range, respectively. Another intrinsic constraint is the In the low frequency band, the electrode/medium interface is of primary importance because of the polarization effect [14]. This is true for both macroscopic and microscopic electrodes. This complex phenomenon is not easy to control in particular for sensors based on microelectrodes. The polarization impedance [13, 15] is an important factor governing the measures that depends on the well known problem of determining the cell factor [2, 16].

Difficulty to calibrate the instrumentation and the sensor [1, 2, 12, 13]. In fact, there are no references reliable to measures on biological medium. The validation of

the sensor and the measurement system is then determined by the characterization of standard materials like saline water [1, 2, 16, 17].

4 Measurement Constraints Related to the Biological Tissue Sample

The characterization of a biological medium depends on its nature but also on the conditions of its preservation. Measurements may be made *in situ*, however tests on *ex vivo* samples are the most commonly used [2, 4]. For *ex vivo* measures on a sample, its physical state (solid or liquid), its geometry], are all affecting the sensor/medium interface.

The significance of the measures faces anisotropy and non-homogeneity of the sample under test. For *ex vivo* measurement, other factors such as the post-mortem time, the pathological or not nature (cancerous/healthy) of the sample [4, 8] the use or no of anti-coagulant agents or other preservative drug, oxidation of the surface following excision must be overcome or at least their influence estimated and taken into account. As regard to the accessibility to the site under investigation and to the previous comments, the needs for well-documented measures are necessary to compare between different data and a protocol must be provided. The measurement of tissue dielectric properties may be constrained by other factors, related to the sort of tissue. Other aspects like non linearity or the order-disorder behavior occurring at the cell scale are still investigated [18]. These numerous influencing factors are too difficult to be managed during the experiences and must be documented as much as possible. That was not so obvious in the past and this is one among the reasons why comparative studies remain critical even between samples from similar animals or organs.

In case of microscopic measurements, in addition to the above factors impacting the measures, connection to cells or biological molecules to an *ad hoc* sensor necessitates modelization and control of all the previously quoted factors. Among them, the access to and the manipulation of the cell or molecule whose size is generally in the range 5–20 microns induce an electrochemical interaction occurring at the sensor/membrane interface that controls the exchange of ions and molecules between the extra cellular and intracellular media (Fig. 2). In the frequency range up to 10 MHz, current conduction through tissue is mainly determined by the tissue structure, i.e. the extra- and intra-cellular compartments and the insulating cell membranes. Therefore, changes in the extra- and intra-cellular fluid volumes are reflected in the impedance spectra [9, 19]. This microscopic EIS was recently developed by numerous teams and is probably one of the most promising tools for biosensors and biochips applications. An example of challenge faced by the bioimpedance research groups is the determination of a correlation between the electrical properties of one cell (microscopic properties) versus an aggregate of cells (macroscopic properties). This new way for bioimpedance applications will not be more discussed here.

5 Examples of Ex Vivo Results

The measurement of tissue electrical properties, *in vivo*, is complicated and was previously reviewed by many authors. In this paper we consider only macroscopic characterization that consists on impedance measurement of a great biological tissue sample or organ. Electrodes are designed according to the nature of the sample and the frequency investigated band.

Effect of influencing factors on the experimental results obtained on human blood are presented below. They are compared to previous datas obtained other authors. A second example will focus on the influence of the nature of the biological tissue at the results obtained on bovine femoral bone slices and the effect of the anisotropy on their electric properties.

6 Measurements on Human Blood

Blood is one biological medium that was investigated by many research groups [13, 17, 20, 21, 22]. This medium is not too difficult to manipulate experimentally and the difficulties are related to its preservation by anticoagulant agent and the influence of physical agents like temperature. It could be considered in a first reasonable approach as non anisotropic however its biological composition, in particular its haematocrit, influence the measures [23].

We have developed an experimental set up based on the V/I method extended to high frequency using a material analyser (HP4191A) (see Fig. 3) that operates between 1 MHz and 1,8 GHz [24]. A homemade coaxial cell allowing to contain small blood sample at a constant temperature was modeled and designed. Dielectric properties are deduced from the bioimpedance measures and the cell factor of the sensor [12, 25, 26].

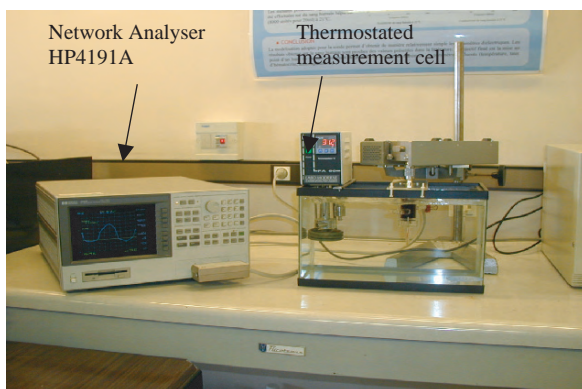


Fig. 3 Experimental set up for blood dielectric characterization

7 Ex Vivo Results

We have selected only two examples from the literature [27, 28] and we have compared them to the results obtained by our group [23, 29]. This is due to the fact that it is not always possible to find results concerning human blood done in the same conditions. The main difference between these authors is that not all the operating frequency range (1 MHz-1GHz) was investigated by them. However their data were documented by giving details like the temperature (24 & 25°C). Bianco gave the haematocrit value (41%) which is not the case for Wei. Relative permittivity and electric conductivity versus frequency are presented on Figs 4 and 5.

7.1 Discussion

The haematocrits are close (45 & 46) for [27] and [29]. Temperatures are similar (24 & 25°C) only for [27] and [28]. Comparing these results, it is evident and obviously it was shown before in the literature [13, 22, 23, 30, 31] that the temperature is a major influencing factor for the blood conductivity. This phenomena is more pronounced at high frequency. The relative permittivity remains less sensitive at the temperature influence for small variations. Even these results are evident, one can comments on the difficulty to find well documented results done in the same conditions. This difficulty is still present in too much papers even recently published. This questions remains open for data comparison.

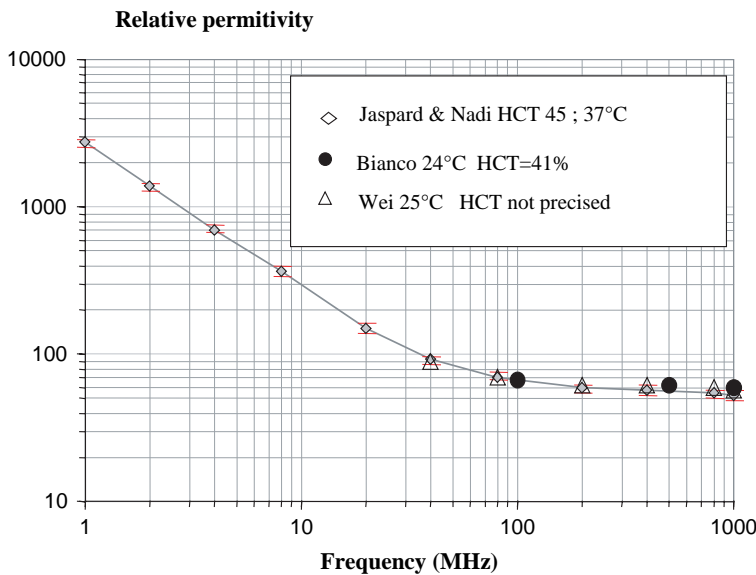


Fig. 4 Relative permittivity of human blood versus frequency [27, 28, 29]

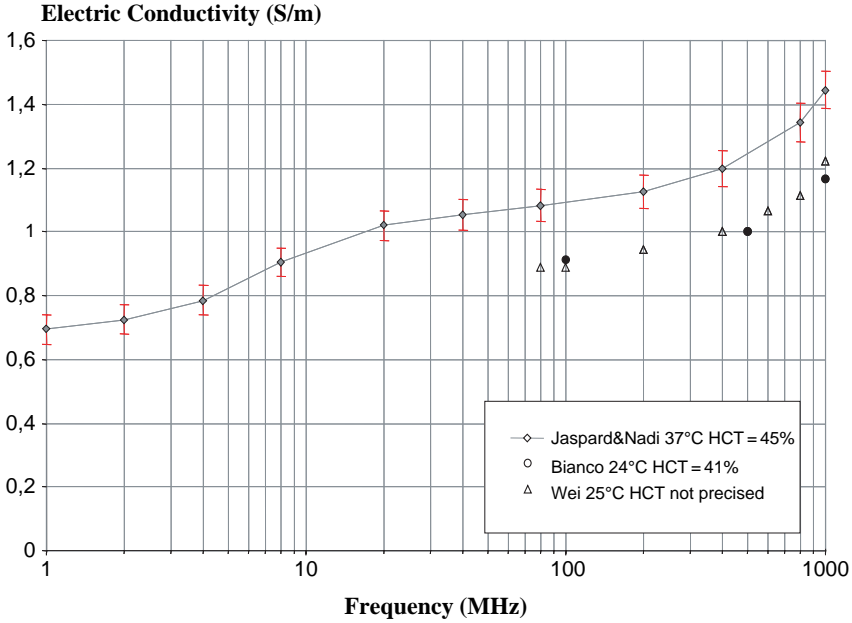


Fig. 5 Electric conductivity of human blood versus frequency [27, 28, 29]

7.2 Modelization Approach for Dielectric Characterization

One possible way of investigation is the use of biophysical models based on Maxwell-Fricke model or other theories [19, 32, 33, 34]. For blood, a possible model as summarized in [29] is based on the following equation

$$\sigma_{GR}^* = \frac{\sigma_{int}^* + \frac{2d}{R}(\sigma_{int}^* - \sigma_{mb}^*)}{1 + \frac{d}{R} \frac{\sigma_{int}^* - \sigma_{mb}^*}{\sigma_{mb}^*}}$$

The electrical behaviour of a red corpuscle may be considered in a simplified approach by a spherical homogeneous particle surrounded by a membrane (Fig. 2).

σ_{GR}^* , σ_{int}^* et σ_{mb}^* are the respective conductivities of the red corpuscle, the intra cellular medium and the membrane. R is the radius of the spheroid and d the thickness of the membrane. η is the shape factor ($\eta = 2$ for a sphere). This is a simplified equation when the assumption $d \ll R$ is valid. More complex models are available in the literature [22, 34, 35, 36].

8 Simulation

We present two examples (Figs 6 and 7) for both dielectric permittivity and electric conductivity simulated by the simplified Maxwell-Fricke model compared to experimental datas on animals blood done for different values of the haematocrit [37]. Measurements were done on blood for beef and sheep at a thermostated temperature of 37°C and different haematocrit. Five different mixtures were prepared

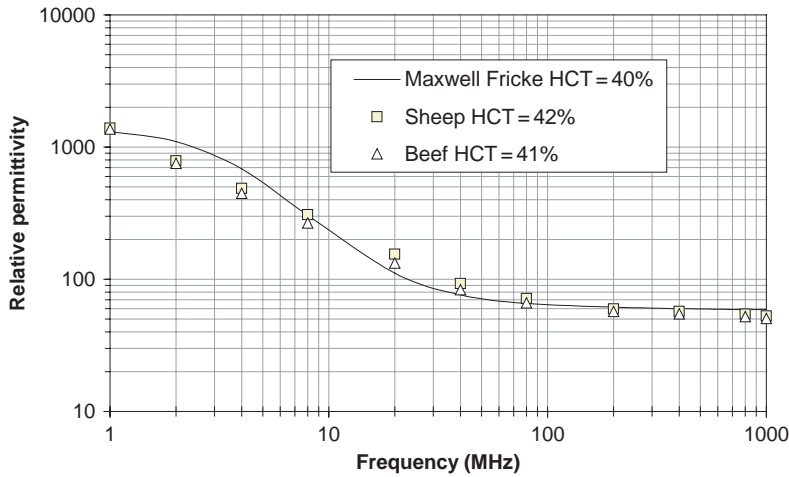


Fig. 6 Comparison between animal (sheep and beef) permittivity and Maxwell-Fricke model at 37°C

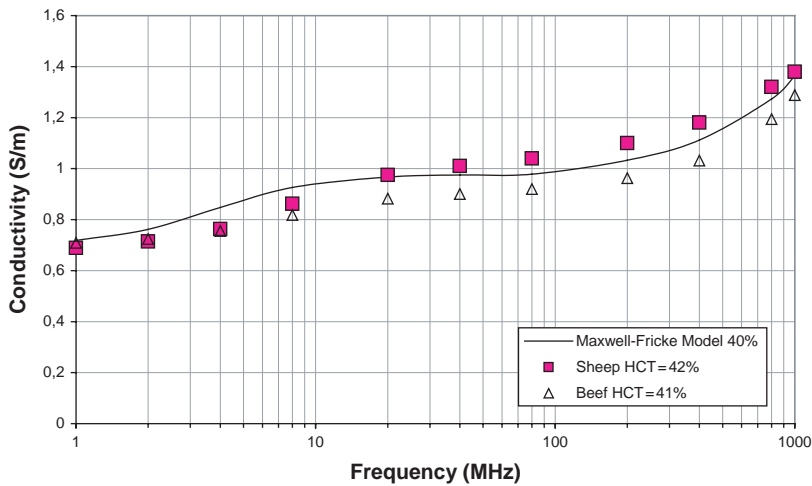


Fig. 7 Comparison between animal (sheep and beef) permittivity and Maxwell-Fricke model at 37°C

according to the procedure described in [29]. For clarity of the results, only one set of similar haematocrit values are compared: Beef (HCT = 41), Sheep (42%) and the Maxwell-Fricke model (HCT = 40%).

Simulation was performed using datas taken from [18] for σ_{mb}^* , [35] and [36] σ_{GR}^* and σ_{int}^* . The dielectric parameter (relative permittivity, ϵ') was deduced from [38]. To use these datas, approximations were necessary and averages from the references extended to beef and sheep. We assume these approximations. The main objective being to check the validity of the Maxwell-Fricke as a first approach for modeling the behaviour of the electric properties as function of the haematocrit.

8.1 Discussion

The Maxwell-Fricke model gives results similar to the experimental datas obtained from sheep and beef at a constant temperature between 1 MHz and 1 GHz. The most significant differences appear at low frequency (1–10 MHz) for the permittivity and at high frequency for the conductivity. This comparison must be analyzed in the light of the great errors classically obtained when measures are done on biological matter. This approach gives a simple and efficient complementary tool for experimental dielectric characterization of blood versus the haematocrit. Investigation of dielectric permittivity and electric conductivity as function of the haematocrit is given in details in [29].

9 Measurements on Anisotropic Medium

Some biological materials, such as bone and skeletal muscle, are distinctly anisotropic. Therefore, when referring to published conductivity and permittivity values, one needs to check the orientation of the electrodes relative to the major axis of the tissue (e.g., longitudinal, transversal, or a combination of both). Electrical anisotropy is related to the physiological demands made on the tissue to insure biomechanical functions. This is the case for bones and muscles. This anisotropic nature lead to a practical problem when measuring the electrical properties of such biological materials. The meaning of the measures is governed by the direction of the applied electric field.

Bone tissues are very particular because their anisotropy is not negligible compared to blood or muscle tissue [4, 8, 39]. We have done measurements on animal ex vivo samples excised from a bovine femur after the soft tissues were removed [40, 41]. Diaphysis was then separated from the rest of the bone and immersed in a saline solution (conductivity 0.005 S/m) and stored in a fridge below 10°C. Five days later, the diaphysis was cutted in 20 slices, 8 to 9 mm thick, with a hack-saw under a constant flow of water to avoid dehydration and thermal damage. Five

regularly spaced slices were then selected and three cubic samples excised in each of them. Sites of excisions were carefully marked (Fig. 8) and samples orientation defined toward the bone microstructure (Fig. 9).

Fig. 8 Three sites of excision per slice

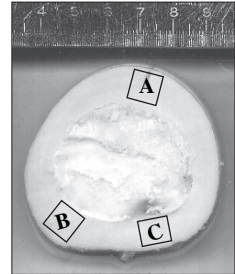


Fig. 9 Samples orientation

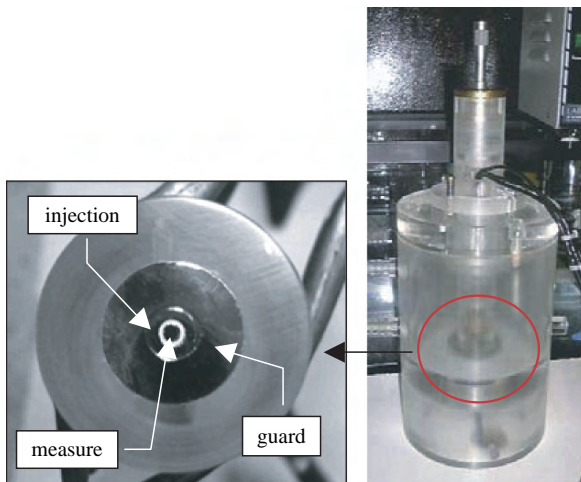
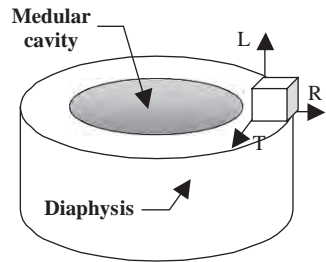


Fig. 10 A set of electrodes and the thermostated measurement cell

9.1 Electrodes Configuration

Since measurements are performed at low frequencies, a four electrodes technique was used.

Each of the two sets of electrodes is constituted of three concentric and circular electrodes (Fig. 10). The inner one (2 mm diameter) is the voltage measurement electrode. This electrode is surrounded by those for injecting the current (5 mm diameter). The outer electrode acts as a guarding electrode. The instrumentation and the procedure were previously described in [40, 41].

10 Ex Vivo Results

Fifteen samples were characterised at $37.5 \pm 0.3^\circ\text{C}$ along there three axis (longitudinal, radial and transversal) and at five frequencies per decade in the range 100 Hz–1 MHz. Each point of measurement was repeated ten times before calculation of mean and standard deviation. In order to validate the measurement cell, the mean of the results obtained on these 15 samples was calculated and compared to results published by other authors [4, 42].

10.1 Relation Between the Permittivity and the Conductivity

A singular behaviour was observed for both the conductivity and permittivity of a slice near the femoral head. A correlation between these two parameters was suspected such as already been highlighted by De Mercato et al. [43]. To test this hypothesis, the longitudinal permittivity of 15 samples was plotted as a function of the longitudinal conductivity of the same samples. Only summarized results of permittivity versus conductivity at 5 frequencies and their determination coefficient are given here.

11 Discussion on Bone Anisotropy Measurements at Low Frequency

The distribution of samples excision sites allows observing if variations of conductivity and/or permittivity are occurring along the bone. The longitudinal conductivity is clearly depending on the location along diaphysis. From the midway of the diaphysis to the proximal head, conductivity is decreasing. However, near the proximal head an important conductivity variation occurs. This higher variation is similarly observed on the longitudinal permittivity. The same variations are

observed when measuring electrical parameters on the radial and transversal directions. These variations could be associated with bone microstructure changes as proportions between the different kinds of osteones.

The conditions under which our relationship has been observed are different from those of [43]. Results obtained by our group [40] used the reference conductivity measured at 100 Hz.

For the measures at 100 Hz, 1 kHz, 100 kHz and 1 MHz, only data relating to the longitudinal axis were used (Fig. 11). These curves confirm that there is no correlation between permittivity and conductivity. The curve obtained at 10 kHz on the other hand uses measurements along the three axes. In this case a significant correlation. However, it is also remarkable that the data are divided into two distinct and distant clouds, which needs a deeper analysis.

These results are inconsistent with those published by De Mercatto. The coefficients of determination shown on the graph very clearly refute the hypothesis of an effective correlation between the permittivity and the conductivity. No statistically significant correlation between these two parameters seems to exist. It is surprising that in [43] an effective significant correlation was observed compared to our results. We explain this by the fact that De Mercato results are valid in a narrow frequency band but not if the frequency band covers several decades.

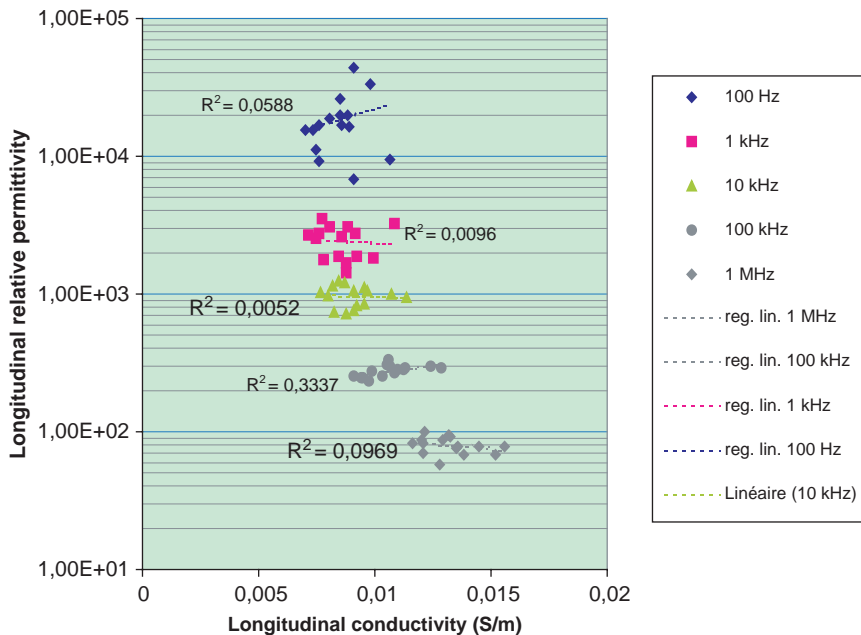


Fig. 11 Coefficients of determination R^2 for permittivity versus conductivity of an anisotropic medium

12 Conclusion

Electrical biological tissues characterization suffers from the lack of precisions and experimental results comparison is difficult to assess. When one is building database for electromagnetic dosimetry simulations, it is not obvious to classify the existing experimental results of dielectric permittivity and electric conductivity. That was previously shown in the well known compilation of existing datas.

Bio impedance measurements depend on the frequency, the temperature, the geometry, the nature of the sample, and specific physical parameters. The definition of the instrumentation is governed by the design of the sensor or the electrodes and by the sample under test. One crucial problem is the interface between the sensor and the biological tissue. Measurements on samples need rigorous preparation in order to alter the influencing parameters as much as possible. This is also necessary to enable the comparison of results from different measurement campaigns. The validation of the dielectric measurement on biological tissues results is also problematic because there are no standard and because of the obvious variations from one sample to another. Thus biological and metrological constraints must be managed or at least estimated when not measurable.

The comparison with or the use of the results of previous works must be done with caution. Most of the data are sparse and often questionable due to the lack of details concerning the validation of the measuring chain or the influencing parameters both metrological and biological.

Using microelectrodes system in the future will open the possibility to monitor cell movement electrically in a biochip laboratory or to characterize a biological cell evolution by the variations of its dielectric properties. Understanding their electromagnetic behaviour, the non linear phenomena or analysing electromagnetic properties of an isolated cell versus an aggregate of cells are a few examples of benefits that bio impedance spectroscopy applications will take from miniaturization in the near future

Acknowledgements We thank M. Roth Patrice (Laboratoire d'Instrumentation Electronique de Nancy) for his technical help and the French Ministère de la Recherche et de l'Enseignement Supérieur for his support.

References

1. Schwan H.P., Determination of biological impedances - in Physical techniques in biological research, Academic press (1963).
2. Rigaud B., Morucci J.P, and Chauveau N., Bioelectrical impedance techniques in medicine, Impedance spectrometry, Critical reviews in biomedical engineering ed. Bourne J.R., pp. 257–351 (1996).
3. Pethig R. and Kell D.B., The passive electrical properties of biological systems: their significance in physiology, biophysics and biotechnology, Phys. Med. Biol. vol. 32(8), pp. 933–970, 1987.

4. Foster K.R. and Schwan H.P., Dielectric properties of tissues, in *Handbook of Biological Effects of Electromagnetic Fields* (2ème edition) Ed :Polk C. et Postow E.,CRC Press 27–102 (1996).
5. Ackmann J.J. and Seitz M.A., Methods of complex impedance measurements in biologic tissue, *CRC Crit. Rev. Biomed. Eng.*, vol. 11 pp. 281–311, 1984.
6. Foster K.R., HERMAN P. SCHWAN: A Scientist and Pioneer in Biomedical Engineering, *Ann. Rev. Biomed. Eng.* vol. 4, pp. 1–27, August 2002.
7. Foster K.R., “In Memorium : Herman P. Schwan [1915–2005]”, *BioMedical Engineering On-Line*, (4:21), (2005), doi:10.1186/1475-925X-4-21.
8. Geddes L.A. and Baker L.E., The specific resistance of biological material—a compendium of data for the biomedical engineer and physiologist, *Med. biol. Eng.*, vol. 5 pp. 271–293, 1967.
9. Gabriel S., Lau R.W., and Gabriel C., The dielectric properties of biological tissues: III Parametrics models for dielectric spectrum of tissues, *Phys. Med. Biol.*, vol. 41, pp 2271–2293, 1996.
10. Schwan H.P., Electrical properties of tissues and cell suspension, *Adv. biol. med. physiol.*, vol. 5, pp. 147–209, 1957.
11. Geddes L.A., *Electrodes and the Measurement of Bioelectric Events*, Wiley-Interscience/Wiley, New York (1972).
12. Misra D., Chhabra M., Epstein B., Mirotznik M., and Foster K., Noninvasive electrical characterization of materials at microwaves frequencies using an open ended coaxial line: Test of an improved calibration technique, *IEEE Trans. Micr. Theo. Tech.*, vol. 38(1), pp. 8–14, 1990.
13. Geddes L. and Sadler C., The specific resistance of blood at body temperature—Medical and Biological Engineering, pp. 336–339, May 1973.
14. Schwan H.P., Alternating current electrode polarisation, *Biophysik*, vol. 3, pp. 181–201, 1966.
15. Mc Adams E.T. and Jossinet, J. Electrode-electrolyte impedance and polarisation, *Innov. Tech. Biol. Med.*, vol. 12, pp. 11–20, 1991.
16. Olthuis W., Streekstra W., and Bergveld P., Theoretical and experimental determination of cell constants of planar-interdigitated electrolyte conductivity sensors, *Sens. and Actuators (B 24-25)* pp. 52–256, 1995.
17. Cook H.F., A comparison of the dielectric behaviour of pure water and human blood at microwave frequencies, *Br. J. Appl. Phys.*, vol. 3, pp. 249–255, 1952.
18. Bistolfi F., *Biostructures and radiation order disorder*, Ed Minerva Medica, Torino, (1991).
19. Fricke H., A mathematical treatment of the electric conductivity and capacity of dispersive systems II. The capacity of a suspension of conducting spheroids surrounded by a non conducting membrane for a current of low frequency, *Phys. Rev.*, vol. 26, p. 678, 1925 (cité dans [FOST 96]).
20. Lu Y., Yu J., and Ren Y., Dielectric properties of human red blood cell in suspension at radiofrequencies. *Bioelectromagnetics*, vol. 15, pp. 589–591, 1994.
21. Sakamoto K., and Kanai H., Electrical characteristics of flowing blood, *IEEE Trans. Biom. Eng.*, vol. BME 26(12), pp. 686–694, 1979.
22. Schwan H.P., *Electrical Properties of Blood and its Constituents: Alternating Current Spectroscopy.*, Blut, vol. 46, pp. 185–197,1983.
23. Hill D.W. and Thomson F.D. The effect of haematocrit on the resistivity of human blood at 37°C and 100 kHz., *Med. Biol. Eng.*, pp 182–186, 1975.
24. Hewlett P., New technologies for wide impedance range measurements to 1.8 GHz, Hewlett Packard Product Note 4291–1.
25. Bussey H.E., Dielectric measurements in a shielded open circuit coaxial line, *IEEE Trans. Instr. Meas.*, vol. IM29 (2), pp 120–124,1980.
26. Jaspard F. and Nadi M., Open ended coaxial line for electrical characterization of human blood – 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 25–28 October 2001, Istanbul, Turkey.
27. Bianco B., Drago G.P., Marchesi M., Martini C., Mela S., and Ridella S., Measurements of complex dielectric constant of human sera and erythrocytes, *IEEE Trans. Instr. Meas.*, vol. 28(4), pp 290–295, 1979 (cité dans [NEEL 83]).

28. Wei Y. – Private communication quoted in [2].
29. Jaspard F. and Nadi M., Dielectric properties of blood: an investigation of haematocrit dependence, *Physiol. Meas.*, vol. 24 (17), pp. 137–147, 2003.
30. Alison J.M., Sheppard R.J., Dielectric properties of human blood at microwave frequencies, *Phys. Med. Biol.*, vol. 38, pp. 971–978, 1993.
31. Cook H.F., Dielectric behaviour of human blood at microwave frequencies, *Nature*, vol. 168, pp. 247–248, 1951.
32. Fricke H., A mathematical treatment of the electric conductivity and capacity of dispersive systems I. The electric conductivity of a suspension of homogeneous spheroids, *Phys. Rev.*, vol. 24, p 575, 1924.
33. Fricke H. and Curtis, H.J., The electric impedance of hemolyzed suspension of mammalian erythrocytes, *J. Gen. Physiol.*, vol. 18 (6), p. 821, 1935.
34. Neelakantaswamy P.S., Apsar K., Rajaratnam A., and Das N., A dielectric model of the human blood – *Biomedizinische Technik*, vol. 28, pp. 18–22, 1983.
35. Pauly H. and Schwan H.P., Dielectric properties of ion mobility in erythrocytes, *Biophysica*. vol. 6, pp. 621–639, 1966.
36. Ulgen Y. and Sezdi M., Estimating the resistivity of the interior fluid of red blood cells using Fricke's equation – *Proceedings 19th International Conference IEEE/EMBS Chicago USA*, pp. 2112–2115, 1997.
37. Cha K., Faris R.G., Brown E.F., Wilmore D.W., An electronic method for rapid measurement of haematocrit in blood samples, *Physiol. Meas.*, vol. 15, pp.129–137, 1994.
38. Mouneimne Y., *Spectroscopie électrique en milieu biologique. Etude du globule rouge. Etude de l'hydratation des protéines*, Thèse, Université Claude Bernard de Lyon, 1986.
39. Gabriel S, Lau R.W., and Gabriel C., The dielectric properties of biological tissues: II Measurements in the frequency range 10 Hz to 20 Ghz, *Phys. Med. Biol.*, vol. 41, pp. 2251–2269, 1996.
40. Chateaux J.F., and Nadi M., *Mesure des propriétés diélectriques de l'os compact dans la bande 100 Hz-1 MHz : Prise en compte de l'anisotropie*, *Revue Internationale de Génie Electrique*, vol. 9 (4–5), pp. 405–416, 2006.
41. Chateaux J-F, *Conception et réalisation d'une cellule de caractérisation des tissus biologiques par spectroscopie de bioimpédance dans la gamme fréquentielle [100 Hz – 1 MHz]. Application aux tissus osseux – Prise en compte de l'anisotropie'*, PhD Thesis, Nancy University, (2000)
42. Kosterich J.D., Foster K.R., and Pollack S.R., Dielectric permittivity and electrical conductivity of fluid saturated bone, *IEEE Trans. Biomed. Eng.*, vol. BME-30 (2), pp. 81–86, 1983.
43. De Mercato G. and Sanchez F.G., Dielectric properties of fluid saturated bone: a comparison between diaphysis and epiphysis, *Medical & Biological Engineering & Computing*, Vol. 26, pp. 313–316, 1988.

Estimation of Property of Sheep Skin to Modify the Tanning Process Using Interdigital Sensors

V. Kasturi and S. C. Mukhopadhyay

Abstract In this paper we are reporting a sensing system to improve the tanning process of sheep skin to produce better quality leather. The dielectric properties of the sheep skin are measured using a sensor system based on an interdigital sensor. Once the skin is converted into leather the process cannot be reversed. Over-treatment of the sheep skin can damage the quality of leather or inadequate tanning may not yield the desired level of quality, so it is important to know the appropriate tanning process required for the skin. Using a non-destructive testing method based on an interdigital sensor, it was attempted to correlate the dielectric property read by the sensor with the looseness value of the skin. At the same time, the testing system must be a low-cost, reliable and fast. The objective of the reported work is to develop a non-destructive low-cost system to identify the looseness of the skin and modify the process to maximize the profit and improve the quality of the finished skin.

1 Introduction

Sensors are extensively used in every aspect of daily human life. Sensors are classified into different groups depending on their use in certain fields and also according to the parameter that they measure or detect. For example, chemical sensors respond to the change in the concentration of a chemical or recognition of a chemical substance [1]. Biosensors respond to the micro-organisms that either stick to it or grow on the surface of the sensors [2]. In the process of detecting or responding to certain factors, the sensors produce either a current or voltage signal. These signals often need to be conditioned before processing. The processing can be efficiently done using a digital data acquisition system. Sensors which respond effectively even at relatively low frequencies are particularly helpful in the development of a low cost sensing systems [3].

V. Kasturi
School of Engineering and Advanced Technology, Massey University, Palmerston North,
New Zealand

S.C. Mukhopadhyay
Massey University, Palmerston North, New Zealand

Much research efforts have been concentrated in developing a non-destructive and non-invasive testing system. In a non-destructive testing (NDT) method, the product under test is not destroyed or damaged, also the test doesn't alter the physical and/or chemical composition of the product. NDT technique is used in various fields like power stations, metal industry, petrochemical industry, transportation, food industry, medical sciences, civil engineering, aircraft inspection and pipe inspection. Its also used for inspection of printed circuit boards [4, 5, 6], inspection of saxophone reeds [7], dairy products [8], estimation of near-surface material properties [9, 10]. Interdigital sensors are used to read the moisture diffusion process in power transformers [11], for qualitative and quantitative gas analysis [12], for measurement of moisture diffusion in power cables [13], for investigating the fluid flow [14] and for materials evaluation [15]. In this paper we report about the inspection of sheep skin using interdigital sensors to modify the tanning process using non-destructive testing to maintain the high quality and standard. Even the sensors are glad-wrapped to avoid the direct contact between the sensors and the skin.

2 Motivation

Leather in New Zealand mostly comes from the animals that are slaughtered at the meat processing industries and hence the skins that are tanned mostly are of sheep or cattle [16]. When an animal dies, its skin loses the property of toughness, flexibility and waterproof nature. The process of retaining the above three properties results in leather and is called tanning. Leather industry is New Zealand's seventh largest industry. Tanned leather is reasonably waterproof and is suitable for making shoes, bags, furniture, jackets and other wearable items. Sheep skins are converted to leather using an 8-step process which involves unhairing, liming, deliming and bating, pickling, tanning, neutralizing and dyeing, drying and finally finishing [16]. The process is lengthy and time consuming. Excessive tanning or not tanning up to the desired level will result in low quality leather. Farmers are paid according to the quality of leather produced from the skins. Farmers, all over the world are looking for ways of improving income from the farm. Sheep farmers are especially concerned with the improvement of the quality, and thus the value, of sheep pelts.

Once the tanning is done the process cannot be reversed. The aim of this project was to identify the skin properties before the tanning process is done to determine the desired level of tanning. Researchers have employed interdigital sensors to measure various material properties [17, 18, 19]. Use of interdigital sensors makes the measurement process fast and suitable for online operations. An added advantage is that it is also non-destructive and non-intrusive; hence the samples tested are not destroyed.

In this work we investigate the use of planar interdigital sensors to measure the dielectric property of the skin before it is treated. The sensors are used to calculate the looseness value of the skin so that the tanning process can be suitably adapted to produce more flexible and tougher leather. This will help in improving the quality as well as profits from the pelts.

3 Operating Principle of Interdigital Sensors

The operating principle of Interdigital sensor is same as that of a parallel plate capacitor [20, 21, 22, 23, 24, 25]. The relationship between the sensor and the capacitor can be seen in Fig. 1, how the transition takes place from a capacitor to the sensor [3]. There is an electric field created between the positive and negative electrode (instantaneous polarity) which are shown in Fig. 1 (a) and (b) respectively. When a material is placed on the sensor, the electric field passes through the material under test which can be observed in Fig. 1 (c). The dielectric properties of the material as well as the geometry of the material under test affect the capacitance and conductance between the two electrodes. The variance in the electric field can be used to determine the properties of the material depending upon the application.

The Interdigital sensor gets the name from its positive and negative electrodes being arranged in the shape of interlocking comb. It is seen in Fig. 2, one side of the electrodes are driven by an AC voltage source and the other set of electrodes is connected to ground. An electric field is formed between the driven and ground electrodes. The electrodes of the interdigital sensor are coplanar, so the measured capacitance will have a very low signal-to-noise ratio. Depending upon the requirement the electrode pattern can be repeated multiple times to achieve a stronger signal.

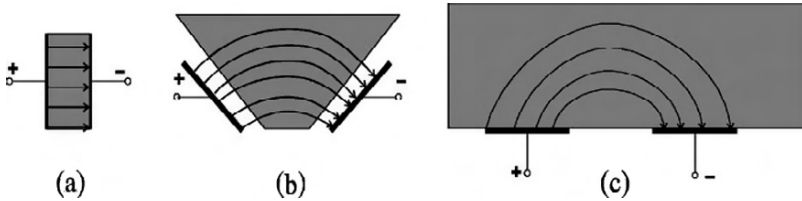


Fig. 1 Operating principle of an Interdigital sensor [20]

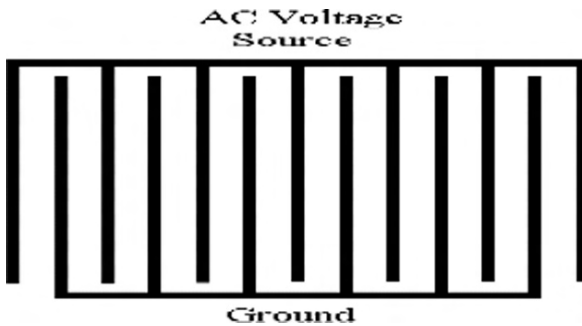


Fig. 2 Interdigital sensor structure [20]

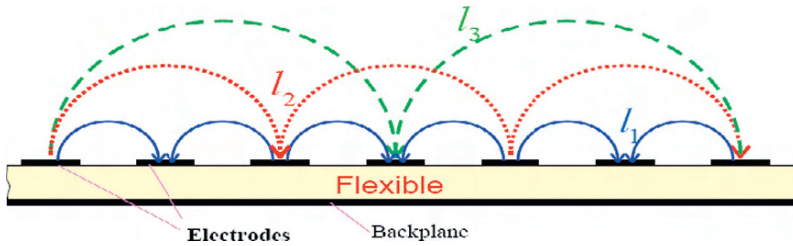


Fig. 3 Electric field formed between two electrodes for different pitch [20]

The flow of electric field lines between electrodes for varying lengths, in between electrodes are shown in Fig. 3. The electric field lines corresponding to minimum separation distance between the positive and negative electrodes is ' l_1 ' and to that for the maximum separation distance is ' l_3 '. So, depending upon the requirement the desired extent of electric field can be achieved by varying the length between the electrodes and the strength of the signal can be controlled by controlling the electrode pattern.

The length between the two adjacent electrodes of same type is referred to as spatial wavelength (λ) and usually the penetration depth is one third of the spatial wavelength [20]. The spatial wavelength is 1 mm as shown in Fig. 4, being the penetration depth is little and the penetration depth increases with increase in the spatial wavelength. D is driving electrode or the AC voltage source electrode and S is sensing electrode or the ground electrode.

As discussed earlier in the introduction, interdigital sensors can be employed in various applications depending upon the requirements. It could be used to measure the density of the material as shown in Fig. 5 (a), the distance between the material under test and sensor could be measured with the help of varying excitation fields as in Fig. 5 (b). It is also possible to identify the non-uniform or unevenly shaped materials using the interdigital sensors as shown in Fig. 5 (c) and they are also very good moisture sensors, as shown in Fig. 5 (d).

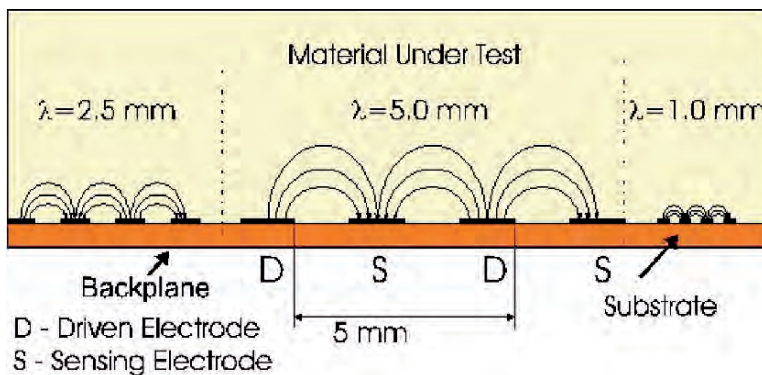


Fig. 4 Penetration depths for varying spatial lengths between the electrodes [15]

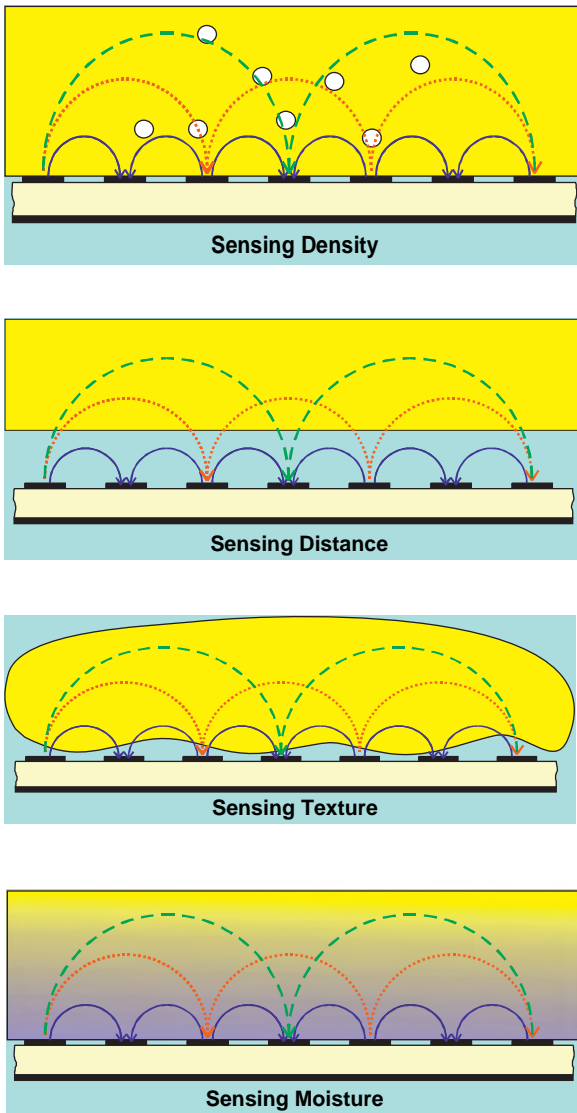


Fig. 5 Sensing various properties of materials [26]. (a) Sensing the material density [26], (b) Measure the distance between sensor and the material [26], (c) Track the structure of the material under test [26], (d) Sensing the moisture [26]

A few fabricated interdigital sensors have been presented in the Fig. 6, with varying pitch lengths, sizes and varying physical structure.

Figure 7, shows how the sheep skin is placed on the interdigital sensor so that it covers most of the sensor area. For an effective measurement, the skin should

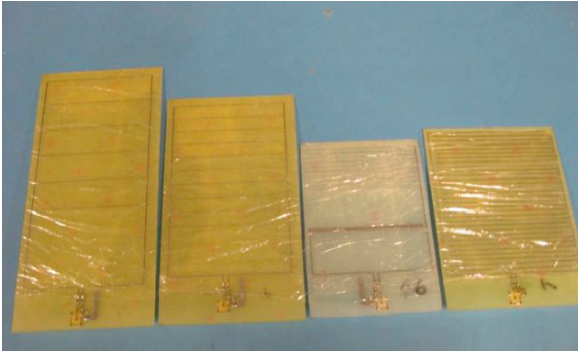


Fig. 6 Interdigital Sensors with varying pitch length and size



Fig. 7 The inspection of skin using interdigital sensor

cover the maximum area of the sensor. However, it should not physically touch the electrodes at one end where the input signal is provided as it would affect the results.

4 Processing of Sheep Skin

The complete tanning is a 8 step process [16]. The process can be described in short:

Step 1 – Unhairing

The animal skins are steeped in an alkali solution that breaks down the structure of the hair at its weakest point (the root) and so removes the hair.

Step 2 – Liming

The hairless skin is immersed in a solution of alkali and sulphide to complete the removal of the hair and to alter the properties of the skin protein (collagen). The collagen becomes chemically modified and swells, leaving a more open structure.

Step 3 – Deliming and Bating

The skin structure is then opened further by treatment with enzymes, and further unwanted material is removed.

Step 4 – Pickling

The skins are then treated with acid to preserve them for up to two years.

Step 5 – Tanning

This is the most chemically complex step. During tanning, the skin structure is stabilised in its open form by replacing some of the collagen with complex ions of chromium. Depending on the compounds used, the colour and texture of the leather changes. When leather has been tanned it is able to ‘breathe’ and to withstand 100°C boiling water, as well as being much more flexible than an untreated dead skin.

Step 6 – Neutralising, Dyeing and Fat Liquoring

The leather is then treated with alkali to neutralise it and so prevent deterioration, and then dyed. This involves fixing a variety of compounds onto the chromium, as that is the most reactive site present. Once the leather is dyed, it is treated with reactive oils that attach themselves to the fibrous structure, improving suppleness and flexibility.

Step 7 – Drying

Water is removed from the leather, and its chemical properties stabilised.

Step 8 – Finishing

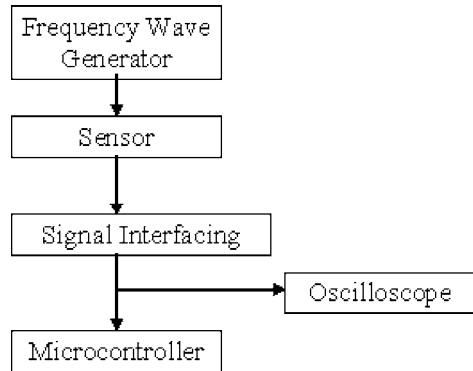
A surface coating is applied to ensure an even colour and texture, and to improve its ability to wear. Suede leather is also buffed at this point to give it its distinctive finish.

After this 8 step process, the leather is used to produce different types of goods or sent overseas for further treatment [16].

5 Experimental Setup

The idea of measuring the sensor response to each skin was to be achieved by exciting the driving electrode using a frequency generator. The signal at sensing electrode is then converted into a dc voltage so that it could be converted into a digital unit using one of the ADC’s of C8051F020 microcontroller. The block diagram is shown in Fig. 8. The signal was also being monitored by using an oscilloscope before feeding it to the microcontroller.

Fig. 8 Block diagram of the experimental set-up



6 Preliminary Experimental Results

First experiments were done to check the sensitivity and nature of the sensors towards materials. For this water, air and commercially available butter and cheese were chosen. A sinusoidal voltage signal with amplitude of 10 V peak to peak and a varying frequency from 1 KHz to 10 KHz was supplied to excite the sensors and the output voltage measured. The experimental set-up is shown in Fig. 9a.

Two sensors with varying pitch lengths and areas are considered for the experiment and were called sensor 1 and 2 to differentiate the results. The sensor 1 is 15 cm in length and 12 cm in breadth, both the source and ground electrode have three fingers each with a pitch length of 3cm between each of the fingers, the width of the electrode is 0.5 mm. The sensor 2 is 13.5 cm in length and 12 cm in breadth, both the source and ground electrode have five fingers each with a pitch length of 1.5 cm between each of the fingers with the same electrode width. With reference to Fig. 9b, the output voltage proportional to the current through the sensor is used as a parameter.

The current is measured by measuring the voltage across the resistance, R , which is connected in series with the sensor. So we have,

$$V_R = I * R$$

where,

V_R is the voltage across R

R is the series resistance

I is the current drawn by the sensor.

Now,

$$I = V/Z$$

where, Z is the impedance of the sensor along with R

$$Z = (X^2 + R^2)^{1/2} \approx X \text{ as } X \gg R$$

$$V_R = R * V/X$$

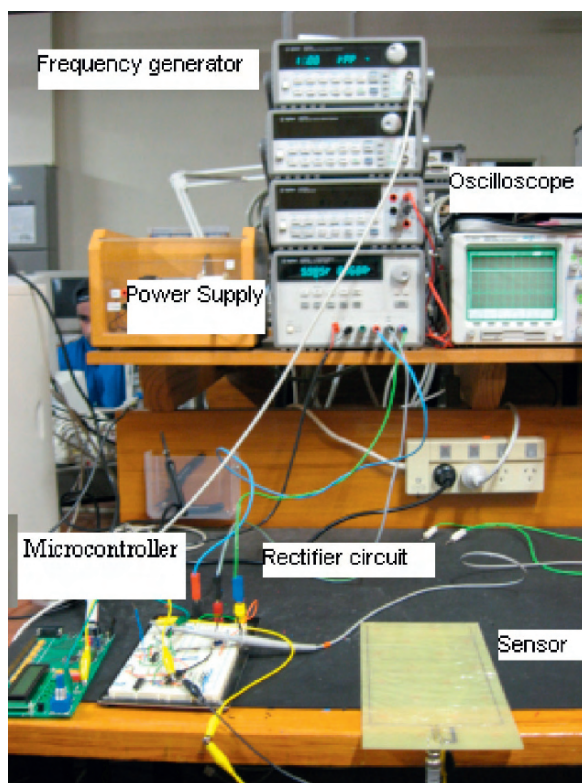


Fig. 9a The experimental set-up

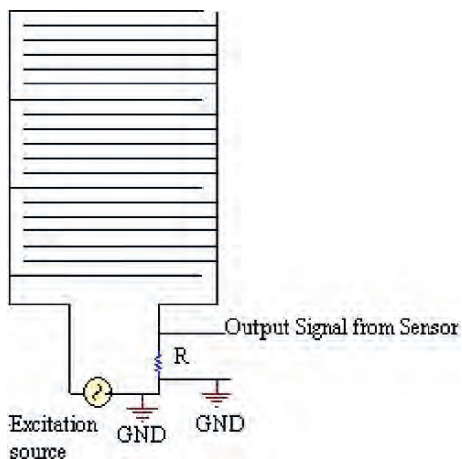


Fig. 9b The sensor, excitation and output signal

$$\begin{aligned}
 |V_R| &= RV\omega C \text{ as } X = 1/(\omega C) \\
 &= Rv2\pi f\epsilon_0\epsilon_r A/d \\
 &= K(\epsilon_r f)
 \end{aligned}$$

where $K = RV2\pi\epsilon_0 A/d$; K is constant for a fabricated sensor;

So, $V_R \propto f$, $V_R \propto \epsilon_r$. The voltage across R is proportional to both frequency and relative permittivity.

The output voltages of both the sensors 1 and 2 for water, air, butter and cheese at different frequencies are plotted in the Figs. 10 and 11 respectively. It can be seen that the sensors have different output values for the same frequencies. However, the nature of the response is similar – the output increases fairly linearly with frequency. The difference in output values for the two sensors can be attributed to the varying pitch lengths and areas. The readings for water were taken by holding the water in

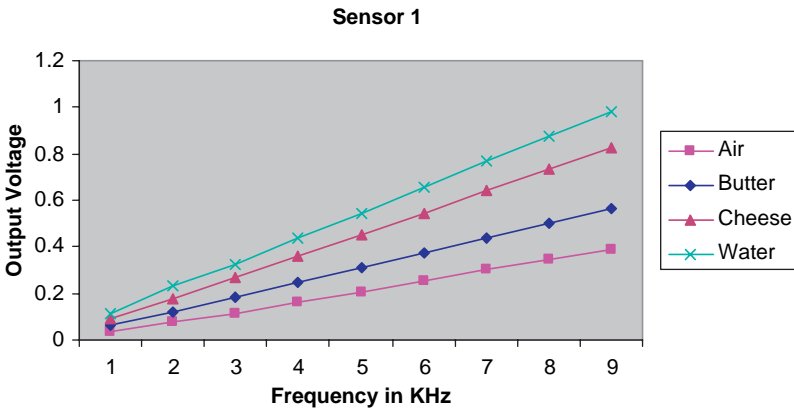


Fig. 10 Output voltage of sensor 1 for air, butter, cheese and water

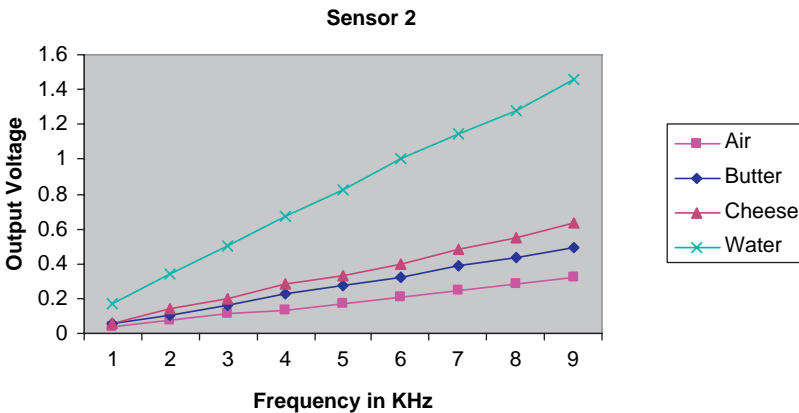


Fig. 11 Output voltage of sensor 2 for air, butter, cheese and water

a plastic bag and placing it over the glad-wrapped sensor. Butter and cheese blocks were also packed in glad wrap.

It can be observed that the output values for butter and cheese are between those of air and water because their relative permittivity is more than air but less than water. Measures are taken to ensure that there is no moisture content on the sensors or the material under test as this would affect the output of the sensor. To avoid this, sensors are wrapped in glad wrap which also helps to eliminate the direct contact of the materials with the sensors.

7 Interfacing and Development of an Embedded Controller Based Sensing System

The voltage signal across the series resistor of sensor is captured and measured by a data acquisition system comprising a precision rectifier circuit and C8051F020 microcontroller for the ADC conversion. The sensors are driven by a 10V p-p sinusoidal waveform. Voltage to the microcontroller’s ADC input is restricted to 3.3 V. Various sensors of different sizes (area) and pitch lengths were used for the estimation of skin properties. One of the purposes of this work is to design a low cost measurement system, readings were taken at relatively low frequencies, from 1 KHz up to 10 KHz. Some of the sensors had very low output voltage at low frequencies which rendered them unsuitable for rectification using silicon diodes. Hence a full-wave precision rectifier circuit, which does not use any diodes was designed and built using operational amplifiers. The rectifier circuit is shown in Fig. 12.

IC1 is operated with a ±9V bipolar supply to ensure that both the positive and the negative halves of the sinusoidal input voltage signal are restored. The output signal from the sensor is passed through the IC1 buffer to also avoid loading problems and is then passed on to the precision full-wave rectifier circuit. The rectifier circuit functions as follows: when the sensor output voltage $V_s > 0$, then IC2 output is half of the circuit input voltage (i.e. $V_s/2$), and IC3 operates as a subtractor, whose output equals the input voltage (i.e. V_s). The waveforms at different stages of the

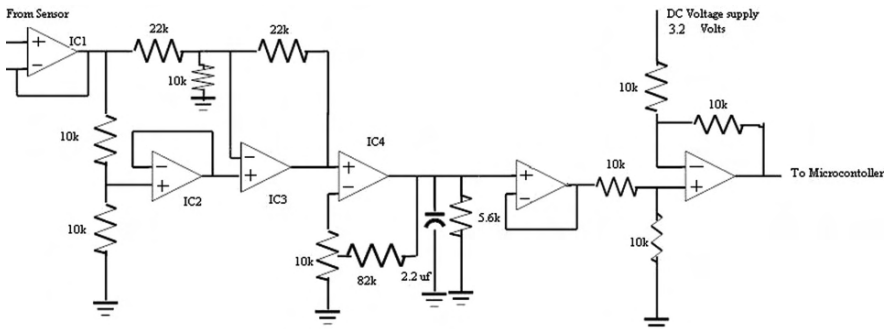


Fig. 12 The interfacing of sensor signal to micro-controller

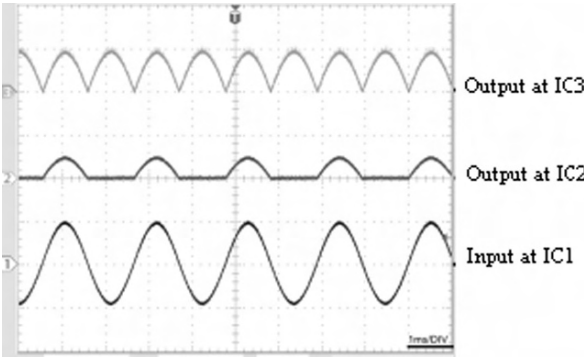


Fig. 13 Voltage waveforms at different stages in the precision rectification circuit [27]

precision rectifier circuit are shown in Fig. 13. The rectified signal is then passed on to IC4 which implements a gain of about 8. The output of IC4 is passed through an RC circuit. The DC signal across the capacitor is passed through a buffer to avoid loading problems. It is then fed to a differential amplifier to get the minimal output voltage and then on to the C8051F020 microcontroller for ADC conversion.

In the case of a specific sensor whose output voltage is relatively higher, the voltage signal after the gain is supplied to the positive terminal of a differential amplifier. A separate DC voltage supply less than the positive terminal voltage is supplied to the negative terminal and measures are taken to ensure that $V+$ is always greater than $V-$ by changing the resistors as required. Now, the final output voltage is supplied to the microcontroller to get converted into digital mode.

The microcontroller’s ADC has a 12-bit resolution. The converted digital data is displayed on the LCD display of the expansion board. First, a few experiments were done to verify if the output of the ADC follows the same pattern of values as

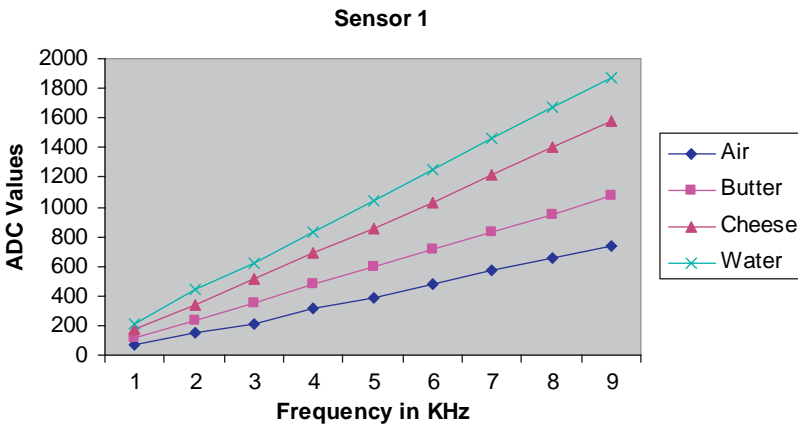


Fig. 14 Response of the sensor 1 for air, butter, cheese and water

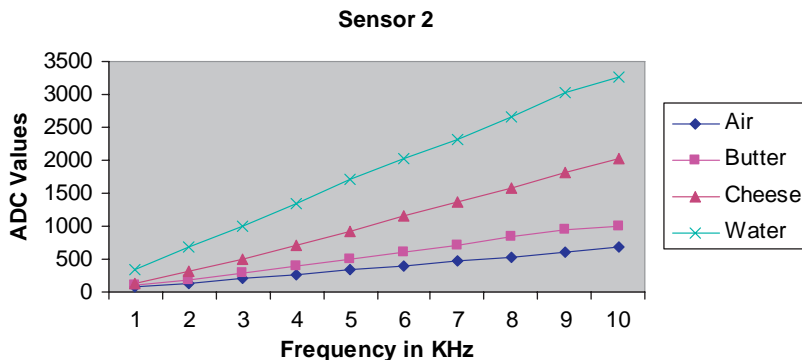


Fig. 15 Response of the sensor 2 for air, butter, cheese and water

the sensor output voltage. Experiments were done using two interdigital sensors for water, air, cheese and butter. The results are shown in Figs. 14 and 15. Even though both the sensors have different output values, they respond in similar fashion at different frequencies.

8 Experiments with Sheep Skins and Results

Experiments were done to determine sensor output at different locations of a sheep skin to establish a relationship between the qualities of a sheep skin based on looseness values and sensor output. Three groups of sheep skins were selected for doing the experiments and each group has six samples. One group of skins was treated using standard process, the second group is treated with excessive concentration of enzyme and the last group of skins were left in alkali for 48 hours; the standard process is for only 12 hours.

A sheep skin is labelled into five zones or locations as shown in Fig. 16 and the output voltage of sensors at each of these positions are measured. The Sensor used to make the measurements on samples is shown in Fig. 17. The sensor was excited using a 10 V sinusoidal signal with a frequency of 10 KHz. The experimental set up is shown in Fig. 18.

The values of looseness for each skin were provided by LASRA (Leather and Shoe Research Association), New Zealand. The looseness of each skin for different groups was evaluated by two people and is shown in Figs. 19, 20 and 21. Looseness is defined on a scale of 1 to 6, if the value is 3 or below 3 it means that the skin is tight and is considered as good quality leather whereas 4 and above 4, the skin is considered as loose and inferior quality leather.

In the figures below, you can see that each skin could be distinctive from the other irrespective of their same chemical treatment. Each of the individuals mostly had a different looseness value for the same skin, which depends on their expertise.



Fig. 16 Sheep skin labelled into five zones

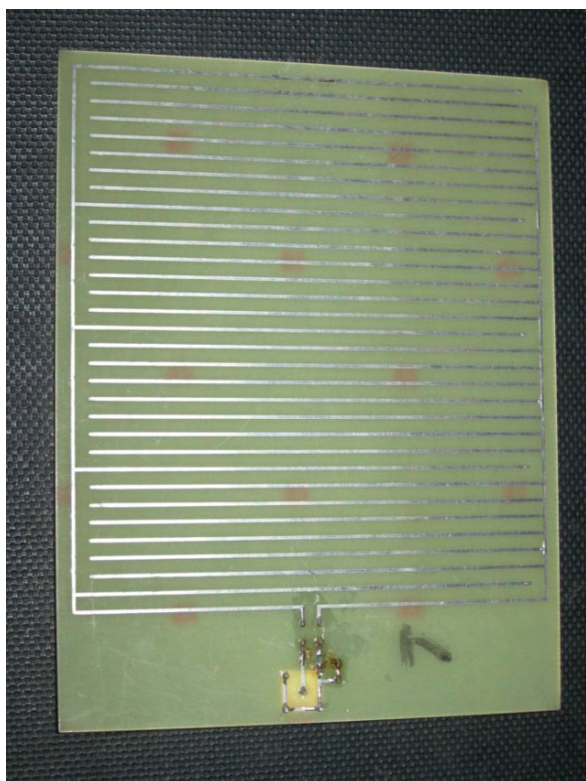


Fig. 17 Sensor used for taking measurements across the sheep skins

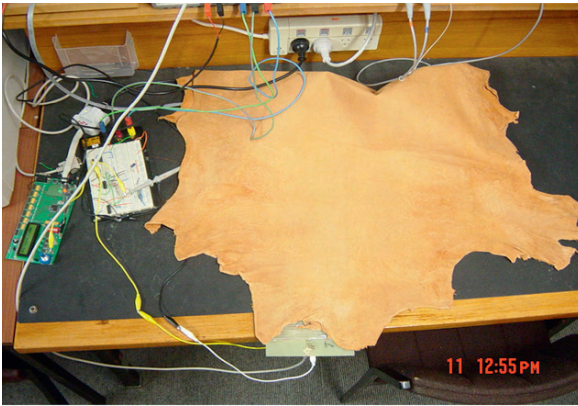


Fig. 18 Experimental set-up for measuring ADC value across the skin

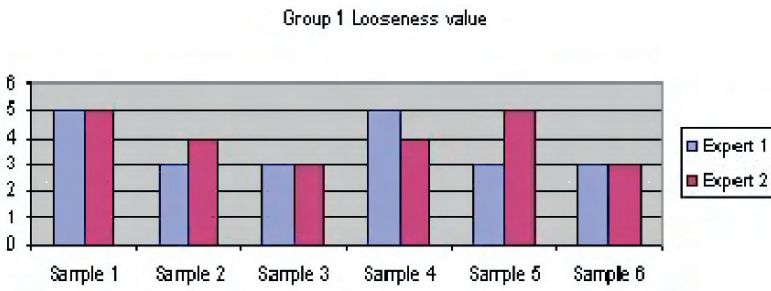


Fig. 19 Looseness values for group 1, estimated by two experts

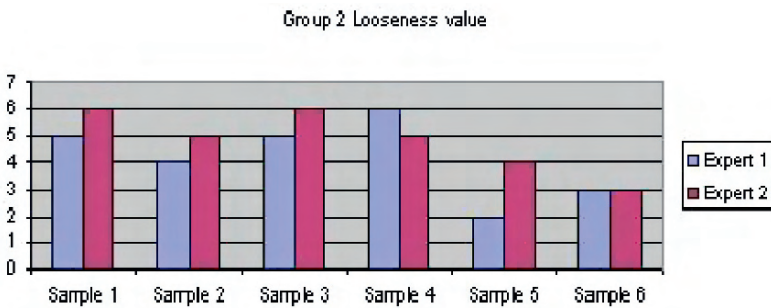


Fig. 20 Looseness values for group 2, estimated by two experts

The looseness values in the above figures were measured at positions 4 and 5, the sensor output voltage at the same positions 4 and 5 are recorded and compared with the looseness values in Figs. 22–27. In Fig. 22, we compare looseness values with the sensor output voltage at position 4 for the samples of group 1. By omitting

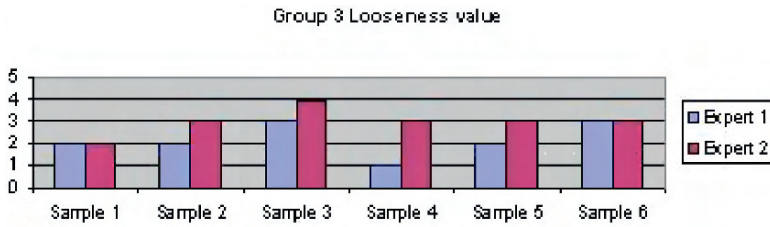


Fig. 21 Looseness values for group 3, estimated by two experts

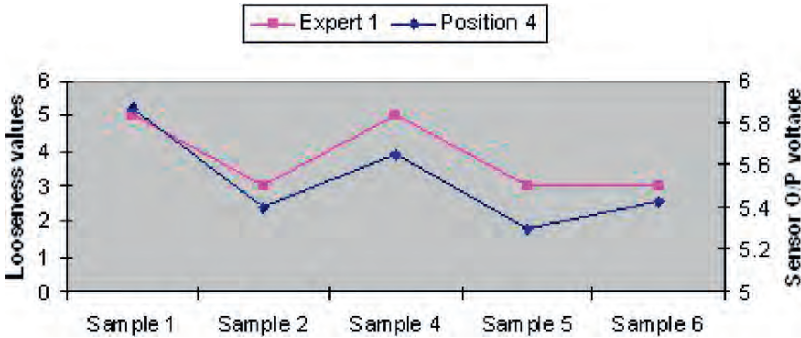


Fig. 22 Comparison between looseness value and sensor output voltage for group 1 at position 4

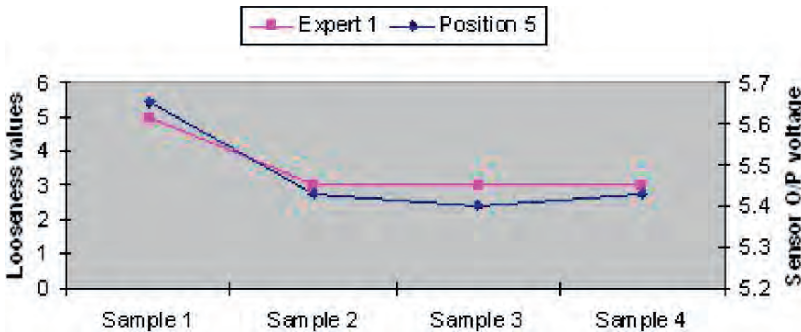


Fig. 23 Comparison between looseness value and sensor output voltage for group 1 at position 5

Sample 3, we can see that sensor output voltage also drops and rises along with looseness values of the skin. In Fig. 22, we compare looseness values with sensor output voltage at position 5 for samples of group 1 and consider only the samples 1, 2, 5 and 6 as it gives a better relation. The looseness value may not co-relate with the sensor output voltage for some samples due to human error or traces of fat left behind on the skin or lift off or the skin not able to cover the cross-section of the sensor.

In Fig. 24, we compare looseness values with the sensor output voltage at position 4 for the samples of group 2. We can see that for samples 1,2,3 and 4 looseness and

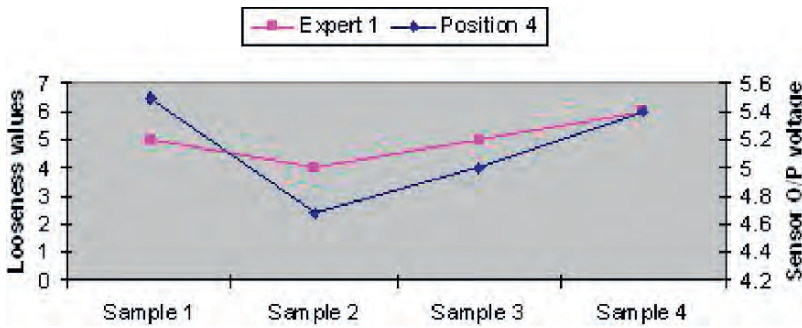


Fig. 24 Comparison between looseness value and sensor output voltage for group 2 at position 4

voltage can be correlated. As the looseness value drops from sample 1 to 2 so does the sensor output voltage. In Fig. 25, we compare looseness with sensor output voltage at position 5 for the samples of group 2. We consider samples 1, 4, 5 and 6 there is a heavy drop in looseness value from sample 4 to sample 5 but there is only a little voltage drop which could be again due to human error or the skin properties.

In the Fig. 26, we compare the looseness values with sensor output voltage at position 4 for samples of group 3. We consider the samples 1,3,5 and 6. We can see

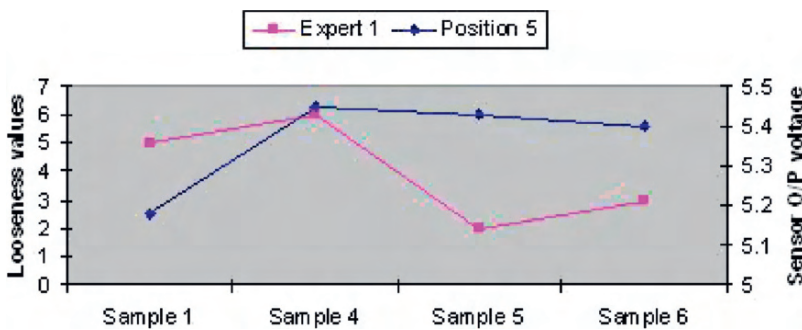


Fig. 25 Comparison between looseness value and sensor output voltage for group 2 at position 5

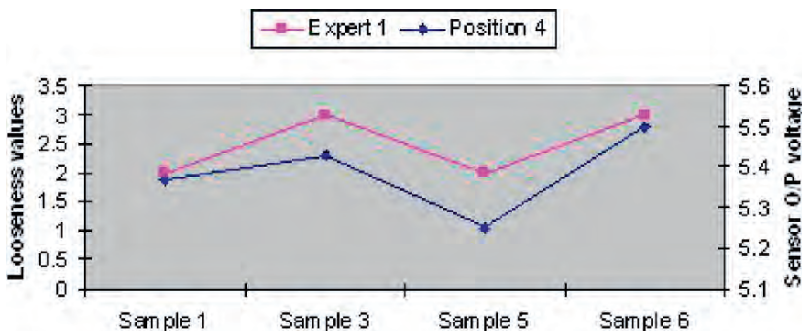


Fig. 26 Comparison between looseness value and sensor output voltage for group 3 at position 4

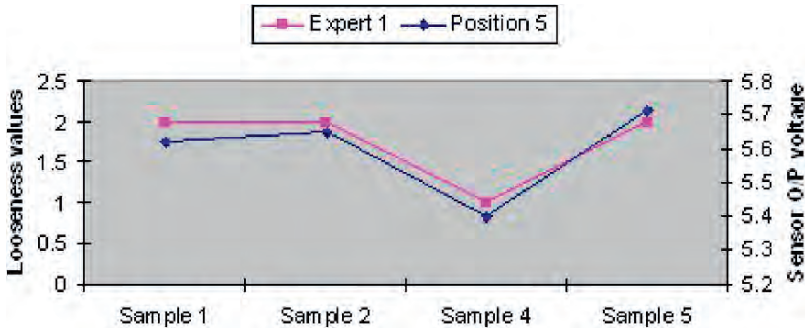


Fig. 27 Comparison between looseness value and sensor output voltage for group 3 at position 5

that the voltage changes along with the looseness values across the samples. In the Fig. 27, we compare the looseness values with sensor output voltage at position 5 for samples of group 3. Sample 1 and sample 2 has got same looseness value but the sensor output voltage changes slightly for sample 1 and sample 2 but after that it changes along with the looseness value for samples 2, 4 and 5. Mostly position 4 has better values corresponding to the looseness values than position 5 across the samples of group 1, 2 and 3.

9 Conclusion

A planar interdigital sensor with an effective data acquisition system has been designed. The sensors responded distinctively to air, water, cheese and butter. Sensors have different values for different materials so they are effective in measuring dielectric properties of various materials and as both the sensors are of same interdigital structure but with varying pitch lengths, their values may differ but follow the same trend. The same set-up was used to measure the di-electric property of pelts and establish a relationship between the dielectric property and looseness of the skin. From the above results it is clear that sensor output voltage drops and rises along with the looseness values for the samples in each group. It is seen that we could co-relate looseness with the di-electric property of the skins. More experiments are being done to establish a relationship between the dielectric property of sheepskin and their looseness.

References

1. Newman, J.D., "Chemical sensor analysis", IEE Colloquium on Materials Characterisation – How Can We Do It? What Can It Tell Us, Dec 4 1997, pp 6/1–6/3.
2. Wang, J., "Electrochemical nucleic acid biosensors", *Analytica Chimica Acta* V. 469 (1), 26 September 2002, pp. 63–71.

3. Mukhopadhyay, S.C., "Sensing and instrumentation for a low cost intelligent sensing system", SICE-ICASE International Joint Conference Oct 2006, 1075–1080.
4. Yamada, S., Katou, M., Iwahara, M., Dawson, F.P., "Eddy current testing probe composed on planar coils", IEEE Transactions on Magnetics 1995, 31, 3185–3187.
5. Yamada, S., Katou, M., Iwahara, M., Dawson, F.P., "Defect images by planar ECT probe of Meander-Mesh coils", IEEE Transactions on Magnetics 1996, 32, 4956–4958.
6. Yamada, S., Fujiki, H., Iwahara, M., Mukhopadhyay, S.C., Dawson, F.P., "Investigation of printed wiring board testing by using planar coil type ECT probe", IEEE Transactions on Magnetics 1997, 33, 3376–3378.
7. Mukhopadhyay, S.C., Woolley, J.D.M., Sen Gupta, G., "Inspection of saxophone reeds employing a novel planar electromagnetic sensing technique", Proceedings of 2005 International Instrumentation and Measurement Technology Conference. 2005, IEEE Catalog Number 05CH37627C, ISBN 0-7803-8880-1, 209–213.
8. Mukhopadhyay, S.C., Gooneratne, C.P., Demidenko, S., Sen Gupta, G., "Low cost sensing system for dairy products quality monitoring", Proceedings of 2005 International Instrumentation and Measurement Technology Conference. 2005, IEEE Catalog Number 05CH37627C, ISBN 0-7803-8880-1, 244–249.
9. Mukhopadhyay, S.C., Yamada, S., Iwahara, M., "Investigation of near-surface material properties using planar type meander coil", JSAEM Studies on Applied Electromagnetics and Mechanics 2001, 11, 61–69.
10. Mukhopadhyay, S.C., Yamada, S., Iwahara, M., "Evaluation of near-surface material properties using planar mesh type coils with post-processing from neural network model.", International journal on Electromagnetic Nondestructive Evaluation 2002, 23, 181–188.
11. Mamishev, A.V., Du, Y., Lesieutre, B.C., Zahn, M., "Measurement of Moisture Spatial Profiles in Transformer Pressboard", Electrical Insulation and Dielectric Phenomena, 1998. Annual Report. Conference on, p.: 323–326, 1, 1998.
12. Niebling, G., Schlachter, A., "Qualitative and quantitative gas analysis with non-linear interdigital sensor and artificial neural networks", Sensors and Actuators B 1995, 26–27, 289–292.
13. Thomas, Z.M., Zahn, M., Yang, W., "Sensors for measurement of moisture diffusion in power cables with oil-impregnated paper", Journal of Physics: Conference Series (2007), 76, 012006.
14. Da Silva, M.J., Sühnel, T., Schleicher, E., Vaibar, R., Lucas, D., Hampel, U., "Planar array sensor for high-speed component distribution imaging in fluid flow applications", Sensors 2007, 7, 2430–2445.
15. Mamishev, A.V., Du, Y., Zahn, M., "Interdigital Frequency-Wavelength Dielectrometry Sensor Design and Parameter Estimation Algorithms for Non-Destructive Materials Evaluation", http://lees.mit.edu/lees/old_files/full/faculty/Zahn/Publications/zahn3_france.pdf
16. <http://www.nzic.org.nz/ChemProcesses/animal/5C.pdf>
17. Sundara-Rajan, K., Byrd II, L., Mamishev, A.V., "Moisture content estimation in paper pulp using fringing field impedance spectroscopy" IEEE Sensors Journal 2003, 4, 378–383.
18. Toda, K., Komatsu, Y., Oguni, S., Hashiguchi, S., Sanemesa, I., "Planar gas sensor combined with interdigitated array electrodes" Analytical Sciences 1999, 15, 87–89.
19. Timmer, B.H., Sparreboom, W., Olthuis, W., Bergveld, P., van den Berg, A., "Planar interdigitated conductivity sensors for low electrolyte concentrations" Sensors and Actuators B (1997), 43, 211–216.
20. Mamishev, A., Sundara-Rajan, K., Yang, F., Du, Y., Zahn, M., "Interdigital sensors and transducers" Proceedings of the IEEE 2004, 92, 808–845.
21. Fratticcioli, E., Dionigi, M., Sorrentino, R., "A planar resonant sensor for the complex permittivity characterization of materials" IEEE MIT-S Digest 2002, 647–649.
22. Toda, K., Komatsu, Y., Oguni, S., Hashiguchi, S., Sanemesa, I., "Planar gas sensor combined with interdigitated array electrodes" Analytical Sciences 1999, 15, 87–89.
23. Timmer, B.H., Sparreboom, W., Olthuis, W., Bergveld, P., van den Berg, A., "Planar interdigitated conductivity sensors for low electrolyte concentrations" Proceedings of SeSens 2001, 878–883.

24. Sundara-Rajan, K., "Estimation of moisture content in paper pulp containing calcium carbonate using fringing field impedance spectroscopy" *Appita Journal* 2004, 413–419.
25. Sundara-Rajan, K., Byrd II, L., Mamishev, A.V., "Moisture content estimation in paper pulp using fringing field impedance spectroscopy" *IEEE Sensors Journal* 2003, 4, 378–383.
26. <http://www.official.kishore-sr.com/publications/Talks/ifpac04.ppt#7>
27. <http://www.edn.com/contents/images/090105di.pdf>

Part II
Fiber Optic/Optical Fibre Sensors

Fiber Bragg Gratings Evanescent Wave Sensors: A View Back and Recent Advancements

Andrea Cusano, Antonello Cutolo and Michele Giordano

Abstract Among the large number of fiber optic sensors configurations, Fiber Bragg Grating (FBG) based sensors, more than any other particular sensor type, have become widely known and popular within and out the photonics community and seen a rise in their utilization and commercial growth. The most relevant milestones of their technological evolution in thirty years from the discovery of Kenneth Hill in 1978 are overviewed. It also reviews the advances in the area of FBGs evanescent wave sensors as valuable technological platforms for chemical and biological applications. Emphasis will be placed on principles of operation, technological developments and overall performances discussing perspectives and challenges that lie ahead.

1 Introduction

The fiber optics field has undergone a tremendous growth and advancement over the past 40 years. Initially conceived as a medium to carry light and images for medical endoscopic applications, optical fibers were later proposed in the mid 1960s as an adequate information-carrying medium for telecommunication applications. The outstanding success of this concept is embodied in the millions of miles of telecommunication fiber that have spanned the earth, the seas, and utterly transformed the means by which we communicate. This has all been documented with awe over the past several decades. Among the reasons why optical fibers are such an attractive

Andrea Cusano

Optoelectronic Division, Engineering Department, University of Sannio, C.so Garibaldi 107 82100 Benevento, Italy, e-mail: a.cusano@unisannio.it

Antonello Cutolo

Optoelectronic Division, Engineering Department, University of Sannio, C.so Garibaldi 107 82100 Benevento, Italy

Michele Giordano

Institute for Composite and Biomedical Materials, National Research Council (IMCB-CNR), P.le Enrico Fermi 1, 80055 Portici, Italy

are their low loss, high bandwidth, EMI immunity, small size, lightweight, safety, relatively low cost, low maintenance, etc.

As optical fibers cemented their position in the telecommunications industry and its technology and commercial markets matured, parallel efforts were carried out by a number of different groups around the world to exploit some of the key fiber features and utilize them in sensing applications [1, 2]. Initially, fiber sensors were lab curiosities and simple proof-of-concept demonstrations. However, more and more, optical fibers are making an impact and serious commercial inroads in other fields besides communications such as in industrial sensing, bio-medical laser delivery systems, military gyro sensors, as well as automotive lighting & control—to name just a few—and spanned applications as diverse as oil well downhole pressure sensors to intra-aortic catheters. This transition has taken the better part of 20 years and reached the point where fiber sensors enjoy increased acceptance as well as a widespread use for structural sensing and monitoring applications in civil engineering, aerospace, marine, oil & gas, composites, smart structures, bio-medical devices, electric power industry and many others [3, 4, 5, 6]. Optical fiber sensor operation and instrumentation have become well understood and developed. And a variety of commercial discrete sensors based on Fabry-Perot (FP) cavities and fiber Bragg gratings (FBGs), as well as distributed sensors based on Raman and Brillouin scattering methods, are readily available along with pertinent interrogation instruments. Among all of these, FBG based sensors—more than any other particular sensor type—have become widely known, researched and popular within and out the photonics community and seen a rise in their utilization and commercial growth. Given the capability of FBGs to measure a multitude of parameters such as strain, temperature, pressure, chemical and biological agents and many others coupled with their flexibility of design to be used as single point or multi-point sensing arrays and their relative low cost, make them ideal devices to be adopted for a multitude of different sensing applications and implemented in different fields and industries.

This chapter reviews the major milestones of their technological evolution during the 30 years from the discovery of Kenneth Hill in 1978. It also reviews the main advances in the area of FBGs evanescent wave sensors limiting the attention on short period FBGs and considering literature from 1996 to present. Emphasis will be placed on principles of operation, technological developments and overall performances. Also, few applications in chemical and biological applications are also described discussing perspectives and challenges that lie ahead.

2 Fiber Bragg Gratings Hystory

Figure 1 illustrates the significant milestones and timeline evolution of the FBG industry over the past 30 years.

The formation of permanent grating was first demonstrated by Hill et al. [7]. They excited a germania-doped optical fiber with intense argon-ion laser radiation at 488 nm and observed that after several minutes the intensity of reflected light

FBG Technology Evolution:
Major Milestones

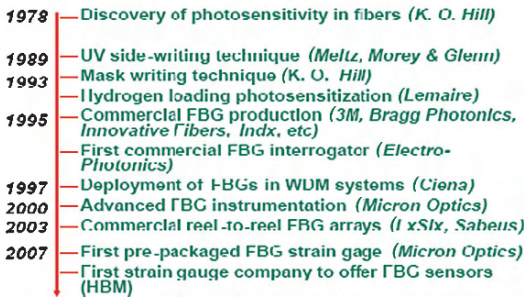


Fig. 1 FBG Technology Evolution
(Source: A. Mendez MCH Engineering, LLC)

increased until eventually almost all the light was reflected from the fiber. The growth in back reflected light was explained in terms of non linear effect called photosensitivity, which permits the index of refraction in the core of the fiber to be increased by exposure to intense laser radiation. In this early experiment, a fiber Bragg grating was formed when a small amount of the laser light reflected back from the end of the optical fiber interferes with the exciting laser light to establish a standing wave pattern. ‘Photosensitivity’ causes the index of refraction to be increased to a much greater extent at position where constructive interference results in a maximum of laser intensity. As the strength of the grating (proportional to the depth of its index modulations) increases the intensity of the back-reflected light increases until it saturates near 100%.

Although photosensitivity appeared to be an ideal means for fabricating these early ‘Hill gratings’ in optical fibers, their usefulness was extremely limited because they only reflected at wavelengths in the visible close to the wavelength of the writing light, were spread along the optical fiber with varying strength and took a long time to produce. These limitations were overcome 10 years later by Meltz et al. [8] who recognized from the work of Lam and Garside [9] that photosensitivity was a two photon-process that could be made more efficient if it were a one-photon process corresponding to the germania oxygen vacancy defect band, at a wavelength of 245 nm (i.e. 5 eV) [10]. In the experiment of Meltz (1989) the fiber was irradiated from the side with two intersecting coherent ultraviolet laser beams of wavelength 244 nm, (see Fig. 2), which corresponds to one half of the 488 nm, the wavelength of the blue argon laser line. The transverse holographic method worked since fiber cladding is transparent to UV light, whereas fiber core is highly absorbing of this radiation. The principal advantage with regard grating fabrication is related to the fact that spatial period of photo-induced perturbation depends on intersecting angle between the two interfering beams. This permits a versatile and efficient fabrication of custom Bragg gratings operating at much longer wavelengths than the writing wavelength as shown in Fig. 3.

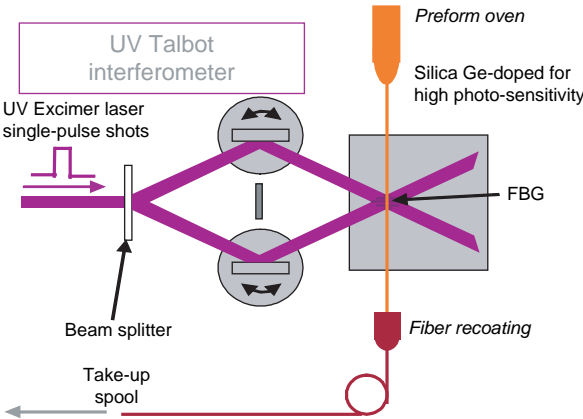


Fig. 2 Schematic of interferometric configuration used by Meltz in 1989 (Source: IPHT Jena)

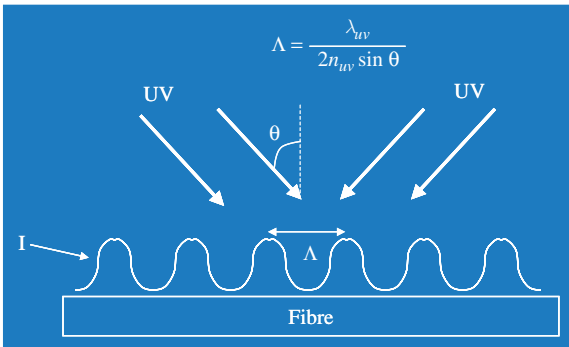


Fig. 3 Grating pitch depends on intersecting angle between UV beams

The periodic perturbation of the core index of refraction gives rise to successive coherent scattering for a narrowband band of the incident light.

The grating thus effectively acts as a stop-band filter, reflecting light with wavelengths close to the Bragg wavelength, and transmitting wavelengths sufficiently different from resonance condition. Each reflection from a peak in the index perturbation is in phase with the reflection from the next peak when the wavelength of the light corresponds to the Bragg wavelength as shown in Fig. 4.

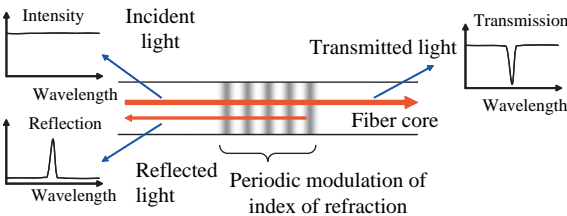


Fig. 4 Principle of operation of FBGs

Theoretical formulations based on coupled mode theory [11] have been developed to analyse fiber grating spectra by Erdogan et associates [12].

Successively, a variety of different Continuous Wave and pulsed lasers with wavelengths ranging from the visible to the vacuum UV have been used to write gratings in optical fiber. In practice, krypton-fluoride (KrF) and Argon fluoride (ArF) excimer lasers that generate (10 ns) pulses at wavelengths of 148 and 193 nm, respectively, are used most frequently to produce FBGs. The exposure required to produce a FBG is typically a few minutes with laser intensities of 100 to 1000 mJ/cm² and pulse rates of 50 to 75 s⁻¹. Under these conditions, the change in the core index of refraction is between 10⁻⁵ and 10⁻³ in germanium doped single-mode optical fiber. Techniques such as hydrogen loading proposed by Lemaire in 1993 can be used to enhance the optical fiber's photosensitivity prior to laser irradiation [13]. Hydrogen diffusion makes the core more susceptible to UV laser radiation. Changes in refractive index of the order of 10⁻² have been achieved by this means.

Successively, the transverse holographic method of writing fiber Bragg gratings has largely been superseded by the phase mask technique in 1993 [14]. Phase mask is a thin slab of silica glass into which is etched (using photolithographic techniques) a one-dimensional square wave periodic surface relief structure as shown in Fig. 5. Since this material is transparent to UV laser radiation the primary effect of the phase mask is to diffract the light into the 0, +1 and -1 diffraction orders. Careful control of the depth of the corrugations in the phase mask suppresses zero-order diffraction, leaving the +/- 1 diffracted beams to interfere and produce the periodic pattern of intense laser radiation needed to photoimprint a Bragg grating in the core of an optical fiber. If Λ_{mask} is the phase mask period, the photoimprinted index grating is $\Lambda_{\text{mask}}/2$. Note that grating period is independent on the writing radiation wavelength. Although, the usual practice bring the optical fiber almost into contact with phase mask, Othonos in 1995 demonstrated the improvements in the spatial coherence of the laser writing relaxed the need for such close contact [15].

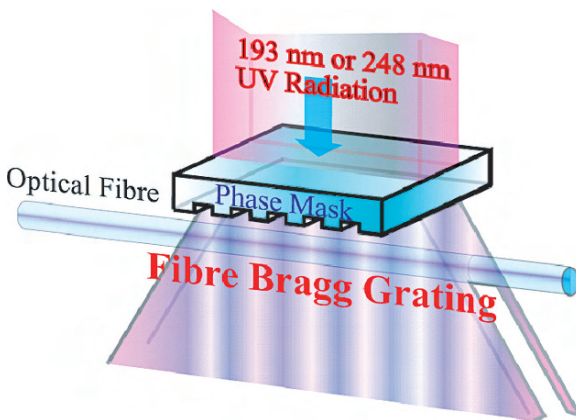


Fig. 5 Diffracted UV beams from phase mask

The phase mask technique greatly simplifies the manufacture of FBGs through easier alignment, reduced stability requirements on the photoimprinting apparatus, and lower coherence demands on the laser beam. It also permit the use of cheaper UV excimer laser source and tends to consistently yield high performance gratings. The prospect of manufacturing high performance gratings at low cost is critical to the large scale implementation of this technology for sensing applications.

The main drawback associated to this approach relies to the need of separate phase mask for each grating with a different operating wavelength. On the other hand, it results very flexible since it can be used to fabricate gratings with controlled spectral responses characteristics.

As a consequence of technological assessment, in the mid 1990s many research groups have been engaged in the study and realization of new grating devices through more complex refractive index modulation profiles. Examples include apodized FBGs, chirped FBGs, tilted FBGs, phase shift FBGs and long period fiber gratings [16, 17, 18, 19, 20, 21, 22].

Although the theory and use of FBGs dates back to the late 1980s, the commercial transition did not happen until the mid-1990s and was subsequently strongly driven by communications needs and the ramping up of the telecommunications 'bubble', which saw a tremendous explosion on the number of companies and research groups engaged with the design, fabrication, packaging and use of gratings.

First companies to produce commercial FBGs were 3M, Photonics and Bragg Photonics in 1995. At the same time, Innovative Fibers was founded by Benoit Lavigne and Bernard Malo in 1995 and was a leader in the design and manufacture of FBG based components for the fiber optics industry including gain flattening filters, 50 Ghz and 100 Ghz DWDM filters and 980 nm and 1480 nm pump laser stabilizers. Successively, in 1997, Ciena Corp a manufacturer of WDM devices, became the largest public start-up company in corporate history and with first year earnings of ~\$200 million had the fastest revenue track ever.

Soon after the telecommunications bubble collapse, there was a significant shift by many players in the industry from communications to sensing applications. At the time, this was a prudent and strategic move on the part of FBG manufacturers to keep exploiting the technical and manufacturing infrastructure they had available and ride the telecomm crisis until a possible comeback. Nevertheless, the sensing sector benefited tremendously from this shift and resulted in an increase in activity and demand of FBG-based sensors.

As FBGs made the transition, from optical communications devices to sensing elements in the 1990s, the bulk of the sensing applications centered on discrete, single-point sensing of specific parameters—such as strain and temperature—using sensors based on embedded or packaged gratings. These early gratings were typically written using phase masks or side exposure interferometric techniques. These fabrication methods initially relied heavily on manual skills and labor, severely limiting many of the features and performance of the gratings in terms of production capacity, repeatability, mechanical strength, as well as number and quantity of FBGs written on a continuous fiber. Furthermore, during the boom years of the

telecommunication industry in the late 1990s, it was possible to absorb the cost of the low yield from such manufacturing.

Due to this increasing interest in FBG sensing technology, many research studies were devoted to the conception of optoelectronic unit able to demodulate FBGs based sensors. As matter of fact, the first optoelectronic unit able to interrogate FBGs sensors was developed 1996 by ElectroPhotonics corporate solutions and was based on the edge filtering concept [6, 23].

However, the sensor industry is much more cost sensitive, demanding multiple sensing points and greater mechanical strength. Such requirements also call for the capability to fabricate an array of multiple FBGs at different locations along a same length of optical fiber. Such needs are being addressed by more sophisticated, on-line, reel-to-reel fabrication processes and systems that allow the writing of complex FBG arrays along a continuous single fiber spool. Since 2000, more than twenty companies have been active in FBGs sensor market. In 2000, Micron Optics was able to launch on the market the first line of advanced FBG interrogators, while in 2003, LxSix and Sabeus launched first reel to reel production of FBGs arrays. Finally, in May 2007 HBM—the world's largest supplier of strain sensing systems—began offering optical strain gages and interrogators based on FBG technology!! This is the first time that a conventional foil strain gage manufacturer has adopted and embraced FBGs. A broad and hard commercial pull should be expected from this initiative.

3 Fiber Bragg Gratings as Sensors

As described in the previous section and with reference to Fig. 4, the fiber optic intracore grating relies on the narrowband reflection from a region of periodic variation in the core index of refraction of a single mode optical fiber [24]. The central wavelength of the reflected Bragg signal is generally called Bragg wavelength and is linearly dependent upon the product of the effective index of refraction of the fundamental mode and the grating pitch: $\lambda_B = 2n_{\text{eff}}\Lambda$. This means that changes in strain or temperature to which the optical fiber is subjected linearly shift the Bragg wavelength leading to a wavelength encoded measurements that is self referencing [25, 26, 27]. Furthermore, intrinsic wavelength encoding also provides a convenient and simple method for serial sensor multiplexing (see Fig. 6)[6].

On the bases of this principle, a large number of solutions based on the Bragg gratings have been proposed in the last decades, for strain, temperature, acoustic waves, ultrasound measurements as well as pressure and magnetic fields. Extensive and excellent overviews of FBG sensing have been provided in past literature, just to cite a few of them [25, 26, 27]. These potentialities combined with FBGs flexibility of design to be used as single point or multi-point sensing arrays and their relative low cost, also make of FBGs ideal devices to be adopted for a multitude of different sensing applications and implemented in different fields and industries. FBG-based sensors have been indeed developed for a wide variety of mechanical

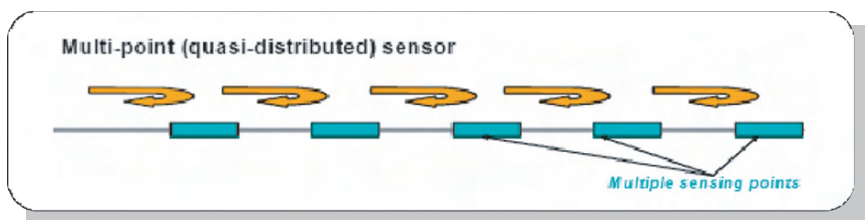


Fig. 6 Multipoint sensors

sensing applications including monitoring of civil structures (highways, bridges, buildings, dams, etc.), smart manufacturing and non-destructive testing (composites, laminates, etc.), remote sensing (oil wells, power cables, pipelines, space stations, etc.), smart structures (airplane wings, ship hulls, buildings, sports equipment, etc.), underwater applications (hydrophones), transportation (rail monitoring, train tracking, carriage safety) as well as traditional strain, pressure and temperature sensing [6, 25, 26, 27, 28] (see Fig. 7).

The present FBG sensor market is primarily composed of 3 key segments: 1) sensing devices, 2) instrumentation, and 3) system integration & installation services [28]. The sensing devices segment is composed of bare FBGs for sensing applications, packaged FBG sensors and FBG arrays. The instrumentation market segment is composed of FBG interrogating instruments and related ancillary components such as multiplexers, switches, data acquisition systems, software and graphical user interfaces. Finally, the third segment is mostly covering services—rather than products—and entails all project management and engineering aspects related


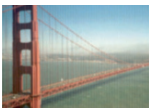




	<p>Oil & Gas</p> <ul style="list-style-type: none"> - Reservoir monitoring - Downhole P/T sensing - Seismic arrays 		<p>Civil</p> <ul style="list-style-type: none"> - Bridges - Dams - Roads - Tunnels - Land slides
	<p>Energy Industry</p> <ul style="list-style-type: none"> - Power plants - Boilers & Steam turbines - Power cables - Turbines - Refineries 		<p>Transportation</p> <ul style="list-style-type: none"> - Rail monitoring - Weight in motion - Carriage safety
	<p>Aerospace</p> <ul style="list-style-type: none"> - Jet engines - Rocket & propulsion systems - Fuselages 		
	<p>Underwater</p> <ul style="list-style-type: none"> - Leaks in subsea pipeline monitoring - Flood detection - Hydrophone 		

Fig. 7 FBGs Applications
(Source: A. Mendez MCH Engineering, LLC)

to implementing sensing solutions and system installations such as design, planning, system integration, customer training, service and on-site installation.

The present worldwide volume demand for bare and packaged FBG sensors is estimated to be greater than 10,000 pieces per year. The worldwide volume demand for FBG arrays is estimated at several 100s to 1,000s arrays per year. The combined present global market size of this segment is estimated to be in the range of \$15M to \$35M USD a year, with an annual growth rate of 15% to 25%. The instrumentation market has been growing steadily over the past three years, in part due to a variety of new fiber sensing projects and installations throughout Asia. Furthermore, as more people get interested in the use and application of FBGs, there will always be a need for interrogating equipment for R&D work, measurements, testing and qualification, as well as actual sensor interrogation. The global volume for FBG interrogating instruments is estimated at several hundred units a year, with an annual growth rate of 20% to 30%. The total market size is estimated to be in excess of \$50M USD.

Regardless of the application of FBG type, two common factors remain: reliability and packaging. Further strides and engineering is still needed to provide effective and suitable packaging to meet the needs of each sensing application. As observed in the telecommunications industry, standards can become a barrier but more often than not, they help simplify and normalize devices, formats and protocols. At the present time, there is no Bragg grating or FBG sensor standards in place. This has led to a broad variability in available grating designs and specifications offered by commercial vendors, as well as a variation in the performance of FBG-based sensors when used in conjunction with instruments from different vendors. Furthermore, there is a tendency to make specific FBGs or FBG arrays from scratch which results in tedious and time consuming up-front phases of any sensing related project, not to mention making the components unnecessarily more expensive. In general, custom products are always more expensive and difficult to manufacture than standardized ones. Hence, sensor interrogation systems need to be standardized as well. Several groups in North America, Europe and Asia are active in standards for fiber optic sensors including OIDA (Optoelectronic Industry Development Association, in Washington, D.C.), ISIS Canada, the European Union COST 270/299 Committee, and RILEM. For its part, OIDA has made initial attempts at formulating standards for FBGs for sensing applications for WDM and TDM interrogating schemes. However, there is not, as of yet, a formal standard development process or group.

Although FBG-based sensors and, for that matter, fiber sensors in general have attracted commercial interest and developed some lucrative niche markets, there are a number of significant technical hurdles and market barriers to overcome. In general, there is still a pervasive lack of awareness and understanding about the operation and benefits of using fiber optic sensors and fiber gratings. Many customers and end-users still distrust the 'subconscious' fragility of optical fibers. However, by far, the most significant barriers that have prevented a more widespread use and commercial diffusion of FBG sensors are inadequate reliability of some existing products and excessive cost. Reliability is a key feature that needs to be taken very seriously and incorporated in every aspect of the fiber sensing design and production facets. It is the reliability that can make or brake the commercial acceptance

and rapid adoption of a given design or product and the one limiting factor that can slow down the utilization of a given product or technology. Many industries are naturally conservative and adverse to failures—such as the electric power, mining and biomedical industries—such attitudes demand that devices demonstrate proven reliability and a solid record of performance established via prototype testing and field trials. Another significant barrier is the fact that most of the sensor developers and manufacturers only provide one piece of the complete sensing solution puzzle. Customers and end users require, in most cases, complete turn-key solutions that encompass all the necessary sensing components, data telemetry & acquisition systems, as well as all the necessary software and data processing algorithms and, most importantly, the actual sensing system design.

4 FBGs Evanescent Wave Sensors

With numerous life sciences applications in mind, the photonic science and technology have recently been evolving into an important interdisciplinary field—biophotonics [29]. One highly attractive area in this large field, the optical biosensor, is experiencing a new surge in activity, seeking to exploit novel optical structures and bio-coating materials and techniques, driven by the demand for high performance analytical tools capable of detecting and discriminating among large classes of biomolecules. Through rapid advances in this area, the highly bio-sensitive/selective optical sensors appear set to become viable and preferred alternatives to traditional “solution based” assay biosensors for applications in genomics, proteomics and drug discovery research and development, as well as in food industry, homeland security and environmental monitoring applications. Over the last decade, fiber Bragg gratings have provided the basis for families of optical sensors for applications in aerospace, maritime and civil engineering structure monitoring, undersea oil exploration, and many other fields. There has also been an increasing flurry of activity aimed at implementing optical biochemical sensors by exploring the grating’s response to the refractive index of the surrounding medium or of thin specific and functionalised overlays. In fact, besides the direct influence of temperature and strain onto the optical length of the fiber grating, there is the possibility to modify the effective refractive index of the guided mode via the evanescent wave interaction [30]. This feature combined with the intrinsic multi-measurand capability in single point and multi point sensor configuration open the way to develop high performances FBGs evanescent wave sensors as valuable technological platforms for chemical and biological applications.

Among fiber gratings, only long period fiber gratings are readily, intrinsically sensitive to SRI via the coupled light from core to cladding penetrating the surrounding medium. As a core-to-core mode coupling device, the light in an FBG is well screened by the cladding, effectively precluding strong interaction with the surrounding medium. However, FBGs can be SRI-sensitised by tailoring either the grating structure or the host fiber creating the basis for the development of possible

technological platforms for chemical and biological sensing applications. Furthermore, despite high SRI sensitivity in comparison with FBGs, LPFGs exhibit several disadvantages as deployable devices. LPFGs possess much higher temperature and bending cross-sensitivities and, thus, can be severely influenced by their environmental conditions. Another disadvantage of the LPFG is that its spectral response can be measured only in transmission, and often with poor resolution due to the broad (typically tens of nanometres) transmission loss-type resonances. To reduce the bandwidth of the LPFG significantly, the device length has to be increased substantially, compromising its use as a localised or point sensor device.

This section reviews the advances in the area of fiber grating evanescent wave sensors focusing the attention on short period FBGs.

The works reported in literature in the last decade could be generally sub-divided in five different classes, namely:

- Fiber Bragg gratings written in D-shaped optical fibers
- Uniformly ThFBGs (Thinned FBGs)
- Tilted FBGs
- Micro-structured FBGs
- Coated FBGs based on chemo-mechanical stress

4.1 FBGs Written in D-shaped Optical Fibers

The first attempt to use FBGs as chemical transducers via evanescent wave interaction was demonstrated in 1996 by Meltz et al. [31]. Since unlike long-period fiber gratings (LPFG), FBGs are intrinsically insensitive to the surrounding-medium refractive index (SRI), since the light coupling takes place only between well-bound core modes that are well screened from the influence of the SRI by the cladding. To address this issue, the basic idea proposed by Meltz et al. was to use fiber gratings written in D-fibers sensitised to SRI through post processing cladding removal.

As reference, Fig. 8 illustrates the cross section of a D-shaped fiber supplied by the KVH Industries, Inc.. The two primary features of D-fiber are the elliptical core, which maintains polarization, and the proximity of the fiber core to the flat surface above the core. This proximity allows access to light in the core by removal of the cladding above the core. Because of the D shape of the fiber, the mechanical integrity and approximately the same dimensions of the fiber are preserved.

Only a few μm cladding thickness on the flat surface side of the D-shaped fiber allows access to the light field in an FBG core directly, and allows for more effective interactions with the surrounding medium. As the cladding is removed, the modal propagation constants are changed thereby shifting the Bragg wavelength of a grating in the etched region. Experiments were reported with etched elliptical-core D-fiber to demonstrate the effect. The Bragg lines of both the fast

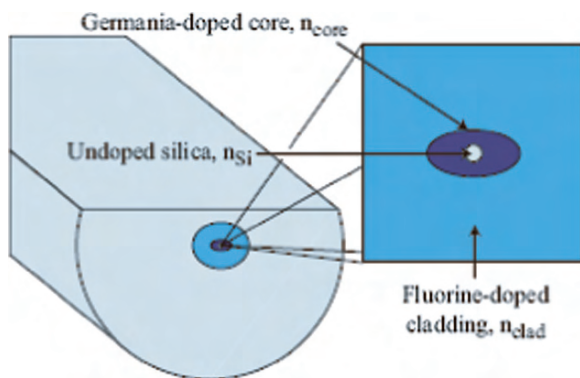


Fig. 8 Schematic of a D-shaped Fiber as supplied by KVH Industries Inc (Source: Brigham Young University)

and slow eigenmodes are blue-shifted when the silica cladding layer is removed and replaced with water or methanol films (lower than the core refractive index). Changes in the fiber birefringence were observed because the perpendicular and parallel modes decay into the cladding at different rates. By using a tunable laser with a narrow band Bragg grating filter or three-grating Fabry-Perot interferometer, refractive index variations of 5 by 10^{-6} could be detected. Temperature compensation methods were also discussed including the use of an isolated reference grating and the simultaneous combination of birefringence and Bragg line wavelength shift measurements.

Successively, planar side polishing [32] was demonstrated as effective technique to develop reliable chemical transducers based on FBGs written in D-shaped fibers with normal birefringence and depressed cladding [33].

In this experiment, standard single mode optical fiber was embedded in a glass block then polished to residual cladding thickness less than $2\ \mu\text{m}$ (see Fig. 9). Finally, Bragg gratings were written in the central part of the polished fiber allowing multipoint sensor array being addressed by wavelength multiplexing of sensors heads in series along a single optical fiber. In follow-up works, the sensing performances of the proposed configuration have been in details numerically and experimentally investigated revealing that [34]:

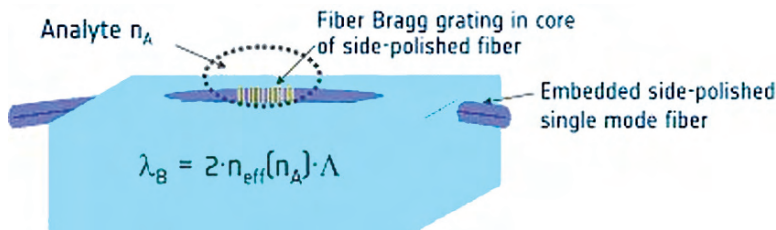


Fig. 9 Schematic of a side polished FBGs chemical sensor (Source: IPHT Jena)

- Due to the asymmetric removal of the cladding, the dependence of the effective index of the fundamental mode on the surroundings is affected by light polarization, in particular, TM polarization exhibits a higher efficiency when compared to the TE polarization state.
- The effective refractive index of the fundamental mode increases non linearly approaching its maximum for SRI values close to the core refractive index (cut off of the fundamental mode)
- Maximum sensitivity is obtained in case of cladding layer completely removed on the flat side of the structure leading to the maximum evanescent wave interaction
- In case of coated structures with low refractive index overlays, sensing performances decrease as the coating is thinned compared with the penetration depth of the evanescent wave
- Due to the dependence of the penetration depth on the optical wavelengths, sensing performances can be improved if longer wavelengths are used.

The typical non linear sensor characteristic is reported in Fig. 10 with the classification of technically relevant substances.

By using a spectral demodulator based on a high resolution mono-chromator with imaging grating and CCD line array also developed at ensuring 0.5 pm wavelength shift resolution, limits of detection of 10^{-5} and 10^{-4} were obtained for SRI close to 1.45 and 1.40, respectively.

In order to increase these values especially for SRI approaching the water refractive index (1.33), a promising approach involving the use of thin high refractive index overlays (HRI) was proposed in 2001 [34]. The main effect of a superstrate with high refractive index is the modification of the mode distribution of the core and cladding modes of the optical fiber leading to a shift of the mode field content towards the overlay. This effects is responsible for the enhancement of the evanescent wave content. If the HRI overlay thickness is of the order of hundreds of nanometers, it was demonstrated that higher sensitivity can be obtained (up to one order of magnitude for SRI close to the core refractive index) allowing also the possibility

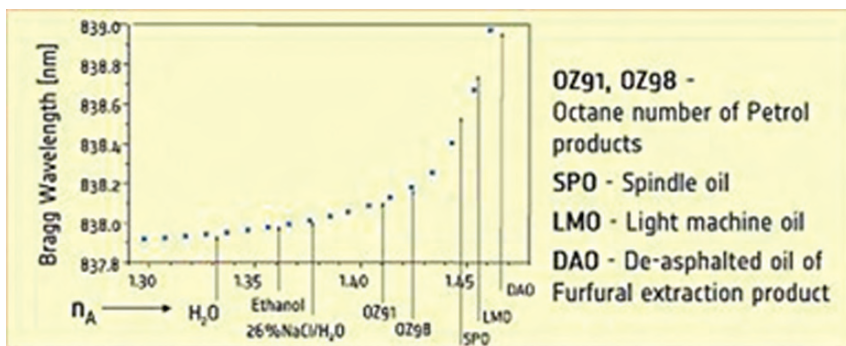


Fig. 10 SRI Sensor characteristic at 838 nm
(Source: IPHT Jena)

to tune the SRI range where the maximum sensitivity occurs. The effect has been recently investigated enabling a full comprehension of the effects of HRI nano-sized overlays on the effective index and mode distribution of core and cladding modes [35, 36, 37, 38].

In order to compensate the SRI detection from thermal changes occurring during practical measurements, the same group proposed three different approaches. The first one relies on the typical solution of using an additional reference grating along the same optical fiber not affected by SRI variations while the second one was based on the use of few mode optical fiber gratings [39] and the coupling with higher order counter propagating modes [34]. In the last case, thermal compensation was obtained by differential wavelength detection, however this method suffers the lower sensitivity of the effective indices of core modes in multimode optical fibers when compared with single mode fibers. Also, considering the higher order modes, a maximum SRI sensitivity of 30 nm/riu (riu: refractive index unit) was obtained compared with the corresponding value of 70 nm/riu in the single mode case at 839 nm [34]. The last method is very interesting and relies on the deposition of overlays supporting SPR (Surface Plasmon Resonance) [40]. An accurate modelling of the coated device and successive experimental validation revealed that TE-polarized wave (within an experimental error of Bragg wavelength shifts <1 pm) is completely shielded from any changes of the refractive index in the analyte. This feature of the SPW enables the ideal separation of the refractometric effect from temperature and strain cross-sensitivities. The FBG reflection from TE mode can be advantageously used as a reference for temperature and strain effects that may occur at the identical measuring site and in the same Bragg wavelength region during the refractometric measurement. Also, the range of refractive indices of the sample at which the SPW (Surface Plasmon Wave) is excited (operation range) may be efficiently tuned by an additional thin high-refractive overlay. As the penetration depth of the SPW into the analyte is limited to <1 μm [40], this new device should exhibit higher sensitivity for thin-film measurements, compared to the conventional evanescent-wave FBG sensing structure [41, 42].

The solid work carried out at IPHT in Jena was also validated in practical applications including on line quality control of petrol products in the bare configuration, in situ monitoring of saline liquids in boreholes with the use of HRI (High Refractive Index) overlay to enhance the sensitivity characteristics in aqueous environments, PH measurements involving polyaniline as specific sensitive coating and hydrogen gas detection by using Palladium as specific opto-chemical sensitive overlay supporting SPR (see Fig. 11) enabling the development of biological sensors based on evanescent wave FBGs [34, 35, 36, 37, 38, 39, 40, 41, 42, 43].

Similar results have been also obtained using FBGs written in D-fibers and sensitised by HF etching adopted for sugar concentration measurements with a sensitivity level of 0.02nm/% [44].

Recently, a novel approach based on surface relies FBGs (SR FBGs) have been demonstrated by Schultz et al. involving RIE etching on photoresist films deposited on the flat side of D-shaped fibers (see Fig. 12) [45]. In this case the grating is formed mainly in the cladding layer close to the elliptical core edge providing a

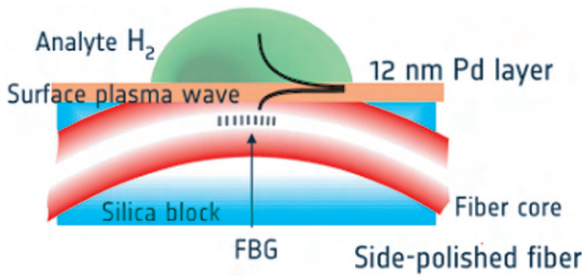


Fig. 11 SPR Sensor based on side-polished FBG
(Source: IPHT Jena)

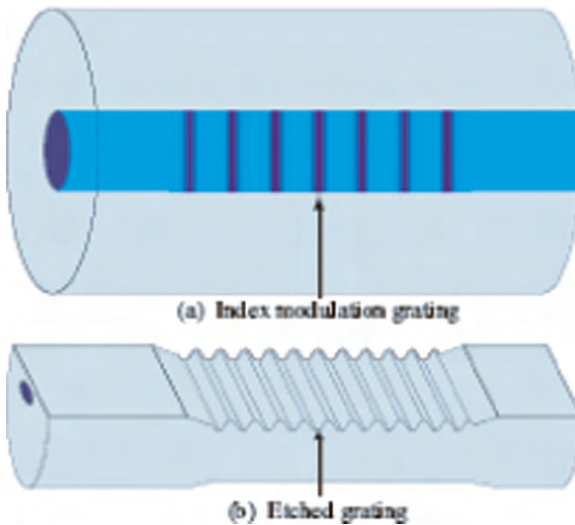


Fig. 12 Illustration of a standard FBG fabricated in a circular fiber (a) and a SR-FBG etched into a D-fiber (b). The gratings are made possible by the periodic perturbation of the structures
(Source: Brigham Young University)

grating refractive index modulation via evanescent wave [46]. The fabrication process includes three basic steps:

- (1) The fiber is placed in a bath of hydrofluoric acid and in situ monitoring of the birefringence between the two polarization states (along the major and the minor axis) in the fiber allows a controlled etch that can be terminated when a predetermined amount of cladding is left above the core [46]
- (2) Photoresist is spun onto the fiber and a grating is patterned into the resist by using a two-beam interference method.
- (3) The grating pattern is transferred into the glass by using a reactive ion etcher and CF_4 plasma.

A schematic of the process is shown in Fig. 13.

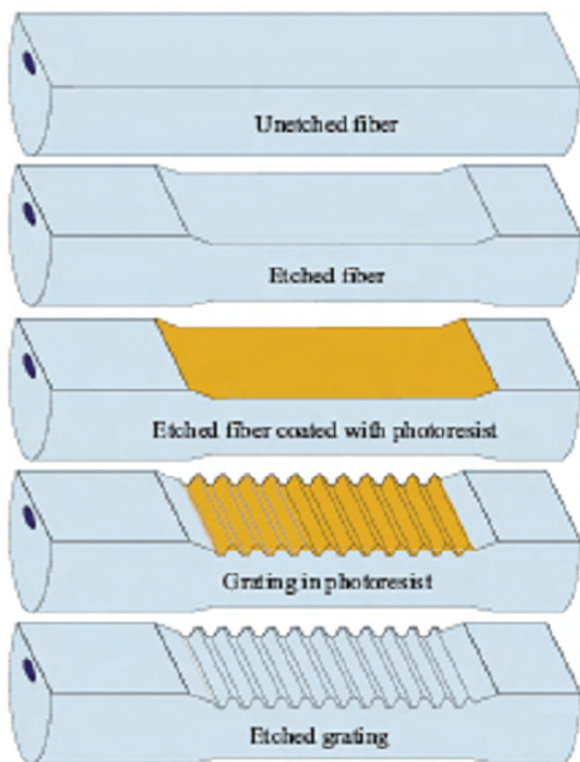


Fig. 13 Diagram showing a D-fiber in the five fabrication stages leading to a SR-FBG (Source: Brigham Young University)

This new class of devices was proposed as volatile organic compounds (VOCs) detector by using Polydimethylsiloxane (PDMS) as polymeric sensitive layer since it has an index of refraction of 1.41 that is lower than that of the optical fiber core, is optically transparent at the operating wavelength (NIR region), and is sensitive only to a specific class of chemicals, namely, VOCs. Also in this case refractive index resolution of the order of 10^{-5} were obtained leading to VOC limit of detection of the order of 4000 ppm [47].

4.2 Thinned FBGs (ThFBGs)

Uniform etched FBGs or ThFBGs were first demonstrated by Asseh et al. [48] (see Fig. 14). In standard single mode optical fibers, the fundamental mode is strongly shielded by the cladding layer avoiding any influence of the surroundings on the guiding properties.

However, if the cladding layer is removed enough, SRI sensitization through evanescent wave interaction occurs. Furthermore, if the circular symmetry is maintained, the modifications of the guiding properties are polarization independent

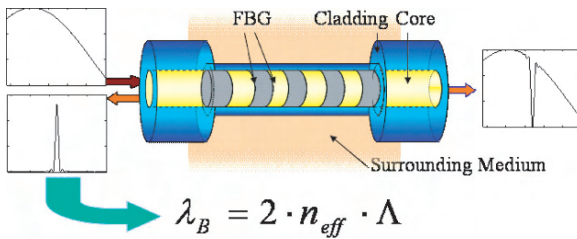


Fig. 14 Schematic of a uniform THFBG

allowing simpler and lower cost equipment for practical sensor interrogation. Similarly to FBGs written in D-shaped fibers, the sensitization through cladding removal depends on the residual fiber diameter and the surrounding refractive index. Figure 15 shows the dependence of the effective refractive index corresponding to the fundamental guided mod on the SRI at 1550 nm for different values of etching depth, full etching case is referred to completely removed cladding layers and D_{Th} represents the residual cladding diameter. As expected, evanescent wave enhancement occurs for SRI close to the core refractive index approaching its maximum in the case of completely removed cladding similarly to FBGs written in D-shaped fibers [34]. In 2003, ThFBGs have been widely investigated by our group to the aim of assessing the fabrication process [49, 50].

Indeed, one of the main disadvantage of ThFBGs relies on the significant weakening of the grating structure especially in cases where maximum sensitivity is required. Proper fabrication procedures and packaging are thus needed to provide reliable SRI refractometers employing this configuration.

In Fig. 16(a), a schematic of the sensor packaging for sensor preparation and testing is shown. A PMMA (poly methyl methacrylate) tube was properly designed and realised to be used for both the etching process and further sensors operation. The fiber was fixed at the two bases using an epoxy based resin (EPON 828 by SHELL)

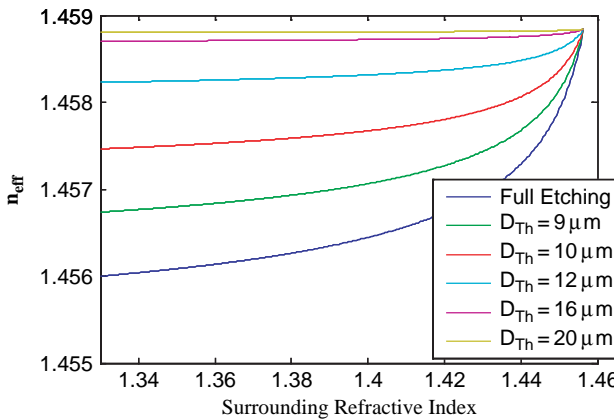


Fig. 15 Effective refractive index behaviour at 1550 nm for different etching depth

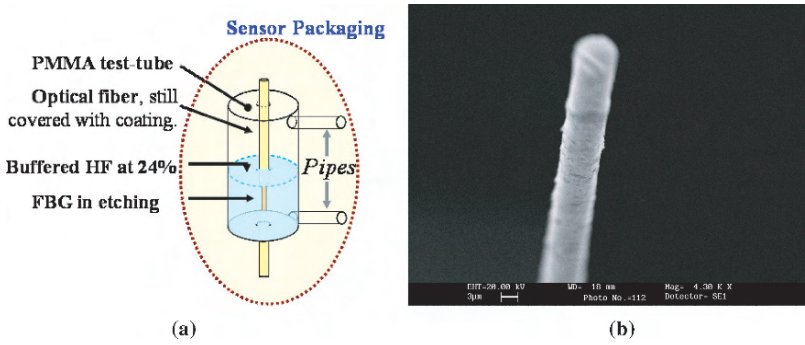


Fig. 16 E(a) Schematic of sensor packaging, (b) SEM image of a full etched ThFBG

and dual functionality pipes were arranged along the test-tube for liquid injection. Teflon layers deposited by spray coating technique were useful to prevent capillarity and undercutting effects. The packaged device was thermo-stated and buffered HF aqueous solution (24%) selected enabling cladding removal at an etching rate of $0.65 \mu\text{m}/\text{min}$ at 24°C (room temperature) [51]. To stop the etching process at the desired depth, the HF solution was removed and the test tube filled with a basic solution (calcium oxide). Figure 16(b) shows the SEM image for a full etched ThFBG. The optoelectronic set-up, used for both fabrication process monitoring and refractive index measurements, is shown in Fig. 17. It comprises a broadband superluminescent diode (2 mW) operating at 1550 nm with 40 nm FWHM (full width half maximum), a directional 3 dB coupler to collect the reflected spectrum from the sensor head and an optical spectrum analyser for spectral measurements. Due to the circular symmetry, sensor interrogation is allowed avoiding expensive polarization components needed in case of D-fiber FBGs. Figure 18(a) shows the experimental behaviour of the Bragg wavelength during a typical etching process.

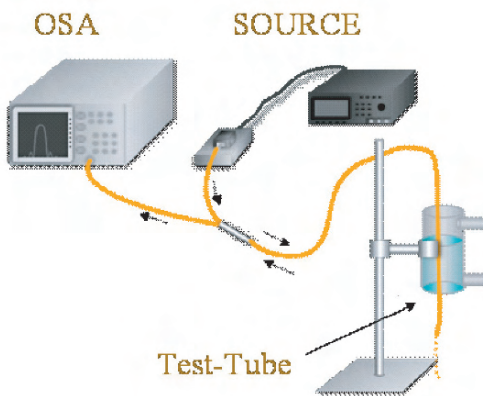


Fig. 17 Schematic of the optoelectronic set up used for spectral measurements

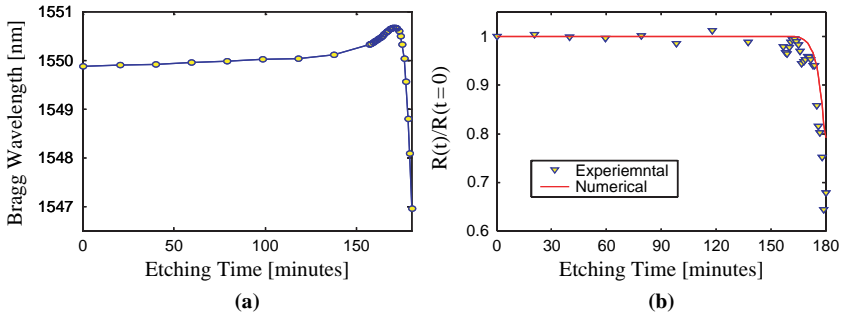


Fig. 18 (a) Bragg wavelength behaviour during etching; (b) Maximum Reflectivity behaviour during etching

For the first 170 minutes, the spectrum slowly moves towards longer wavelengths suggesting that the strain state in the packaged structure is changing during fiber thinning. After 170 min, corresponding to a residual cladding diameter of approximately $20\ \mu\text{m}$, the predominant effect is the evanescent field interaction according to numerical results reported in Fig. 15. Since the refractive index of the 24% HF solution is lower than the core refractive index, a diminution in the effective refractive index occurs leading to a blue shift of the Bragg spectrum.

The wavelength shift of the reflected spectrum is not the unique effect observable during the etching process, also the amplitude of the reflected spectrum decreases depending on the etching depth. Figure 18(b) shows the relative peak reflectivity normalized to the un-etched case measured during the etching process. In case of completely removed cladding, maximum reflectivity demonstrates a maximum diminution of approximately 30% measured with water as external medium. This effect can be explained by considering the mismatch of the numerical aperture between the un-etched and thinned optical fibers depending on the SRI [52]. A typical sensor characteristic is reported in Fig. 19 for a full etched sensor and SRI range 1.33–1.45. Good agreement was obtained when the measured data were compared with numerical results obtained by using the multilayer fiber model in combination with the transfer matrix approach for Bragg reflectance calculation [51].

In case of interrogation units able to discriminate wavelength shifts with a resolution of 1 pm at 1550 nm, refractive index resolutions of $\approx 10^{-5}$ and $\approx 10^{-4}$ are possible for SRI ranging around 1.45 and 1.33, respectively. In addition, diminution in FBG peak amplitude due to the thinning process is not able to influence the system performance for most interrogation units proposed in the literature and commercially available.

Regarding the thermal compensation close to the sensing element when in situ investigations in non-controlled environments are required, a first possibility is provided by adding a standard grating element sensitive only to thermal changes [53]. However, this approach, although efficient for single point monitoring, can not be considered a suitable solution when a large number of spatial locations have to be monitored.

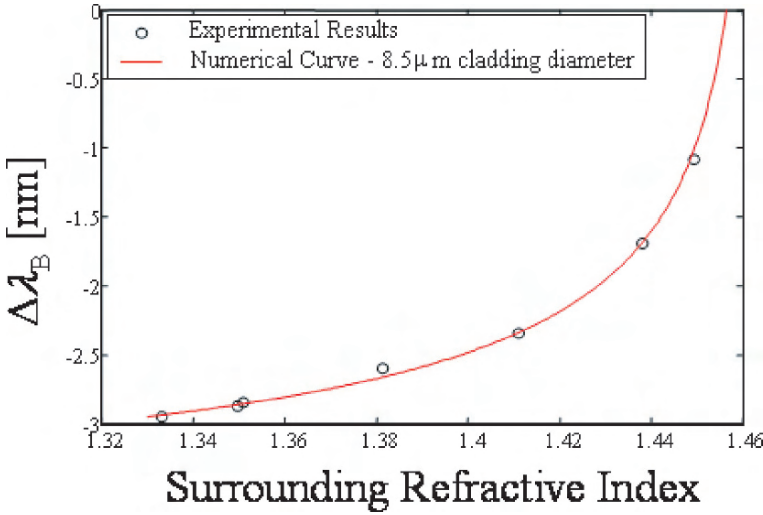


Fig. 19 SRI sensitivity of a full etched ThFBG

A valid alternative using a single sensing element deals the use of non-uniform ThFBG for the simultaneous measurements of refractive index and temperature [54]. The structure relies on a standard grating: in part of the sensing element, the cladding layer is partially or totally removed (see Fig. 20). In these conditions, the etched grating structure would exhibit a spectral response depending on the local temperature and the surrounding refractive index, while the remaining part (un-etched) would respond only to local thermal changes (see Fig. 20(b)) [55]. The main effect of the perturbation is the splitting of the spectral response of the original grating in two peaks located at two distinct wavelengths depending on the etching depth and SRI value.

It is worth noting that thermal sensitivity of the thinned region would depend on the thermo-optic coefficient of the surrounding medium. In case of liquids, the

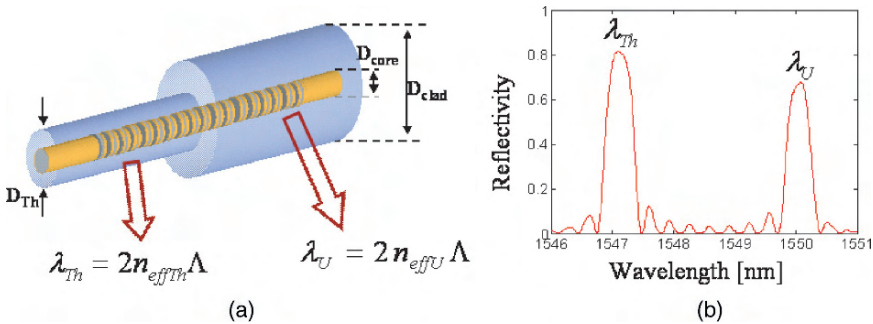
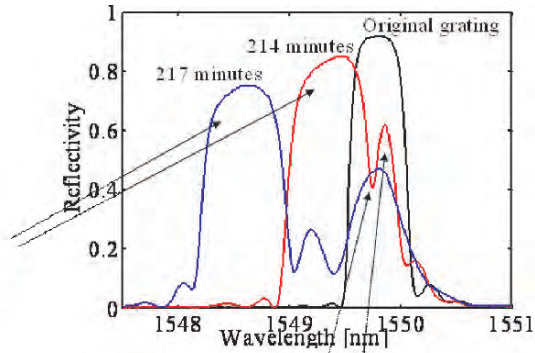


Fig. 20 (a) Schematic of a non-uniform ThFBG for simultaneous refractive index and temperature measurements, (b) Spectra Splitting due to non-uniform fiber thinning

The peak related to the thinned region moves to lower wavelength. This effect is due to the decrease in the effective refractive index.



The peak at the original Bragg wavelength is due to the unperturbed grating region.

Fig. 21 Bragg spectrum evolution during etching of non uniform ThFBGs

thermo-optic coefficient is generally negative leading to a slight lowering of thermal sensitivity in respect to silica cladding.

Figure 21 shows the spectral responses of the device during the etching process. According with the theoretical analysis, during the etching process the Bragg reflected signal splits into two lobes. The first one at longer wavelengths is due to the un-etched grating and matches the wavelength of the Bragg device before etching. The spectral peak at short wavelengths corresponds to the thinned part region and moves demonstrating a blue shift during the etching process due to lowering of the effective refractive index in the thinned region.

Figure 22(a) shows the relative Bragg wavelength shifts of the two sub-peaks due to SRI changes. As expected, SRI insensitivity is obtained with regards the spectrum corresponding to the un-etched grating [55, 56]. Figure 22(b) shows the relative shift

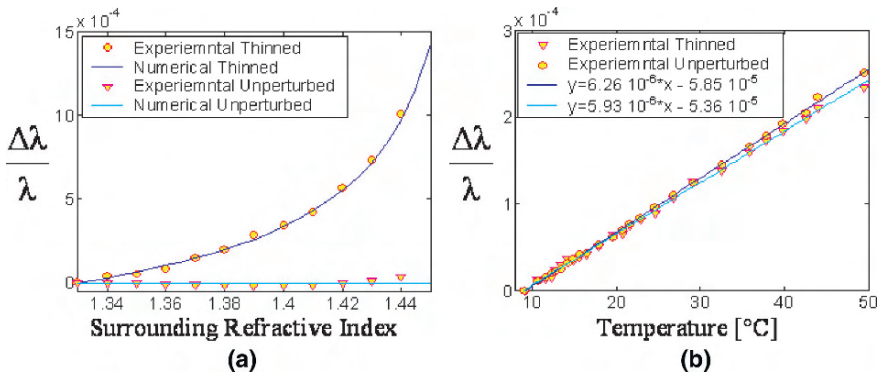


Fig. 22 (a) Relative Bragg wavelength sensitivity to surrounding refractive index, (b) Relative Bragg wavelength thermal sensitivity

in the Bragg wavelengths of the two sub-peaks as function of the local temperature in the range 10–50°C. Linear behaviour was found, demonstrating thermal sensitivities of $6.3 \cdot 10^{-6}/^{\circ}\text{C}$ (not thinned region) and $6 \cdot 10^{-6}/^{\circ}\text{C}$ (thinned region).

A validation of the method was carried out in the case of sugar concentration detection in water environment [56]. Figure 23 shows the relative wavelength shifts of the thinned region before (circle marker) and after (square marker) the thermal compensation together with the thinned grating response for the same sugar concentration in isothermal conditions at 27°C. It is easy to argue how the thermal effects are completely removed allowing self temperature referenced refractive index measurements by using a single sensing element.

Regarding the enhancement of sensor sensitivity, an effective solution is provided by ultra-thinned FBGs whereas the etching process acts also within the fiber core [57]. Numerical and experimental investigation on the sensitivity gain have been carried out by Dagenais et al. revealing that SRI sensitivity can be enhanced up to 1300nm/riu for refractive index close to 1.45 if residual diameter is reduced to 3.4 μ (see Fig. 24 Nevertheless, also an increasing in the optical losses occurs preventing high resolution wavelength shift and the final structure requires particular care in packaging design.

A first demonstration of biosensor concept based on etched core FBGs where single stranded DNA oligonucleotide probes of 20 bases were immobilized on the surface of the fiber grating using relatively common glutaraldehyde chemistry [58]. Hybridization of a complimentary target single strand DNA oligonucleotide was monitored in situ and successfully detected. Figure 25 shows the Bragg wavelength shift of a 5 μm core etched grating during the hybridization of the target.

Further studies on sensitivity enhancement by considering Bragg coupling to higher order modes or cladding mode resonances in etched core ThFBGs have been

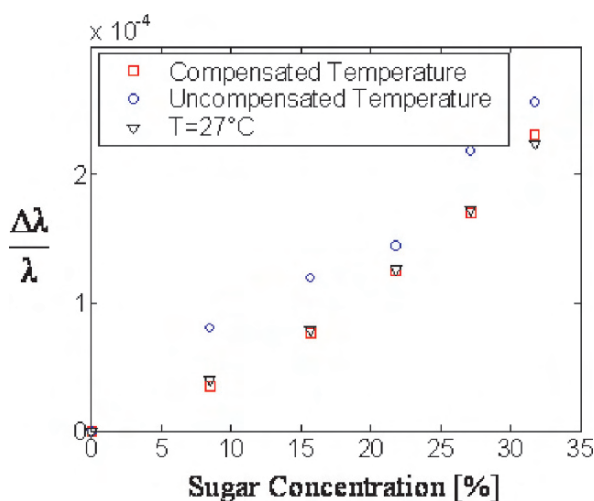


Fig. 23 Refractive index measurements by non uniform ThFBGs with and without temperature self-referencing compensation

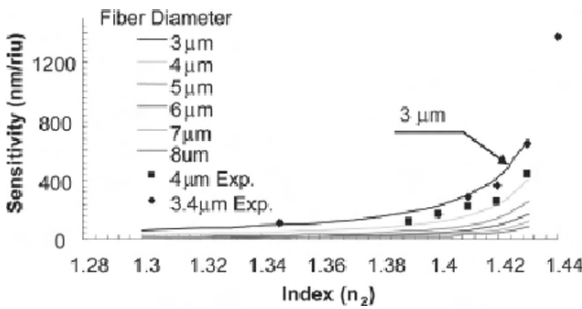


Fig. 24 Sensitivity characteristics in etched core ThFBGS
 (Source: A. N. Chryssis, S. M. Lee, S. B. Lee, S. S. Saini, and M. Dagenais, (2005) ‘High sensitivity evanescent field fiber Bragg grating sensor’ IEEE Photon. Technol. Lett., vol. 17, no. 6, pp. 1253–1255)

also demonstrated [59, 60]. If the SRI results lower than the cladding refractive index, the thinned region allows multimode propagation enabling light coupling to counter propagating high order modes [39]. The sensitivity of the third-order mode can be up to 6.6 times larger than that of the fundamental mode for a 5 micrometer fiber device.

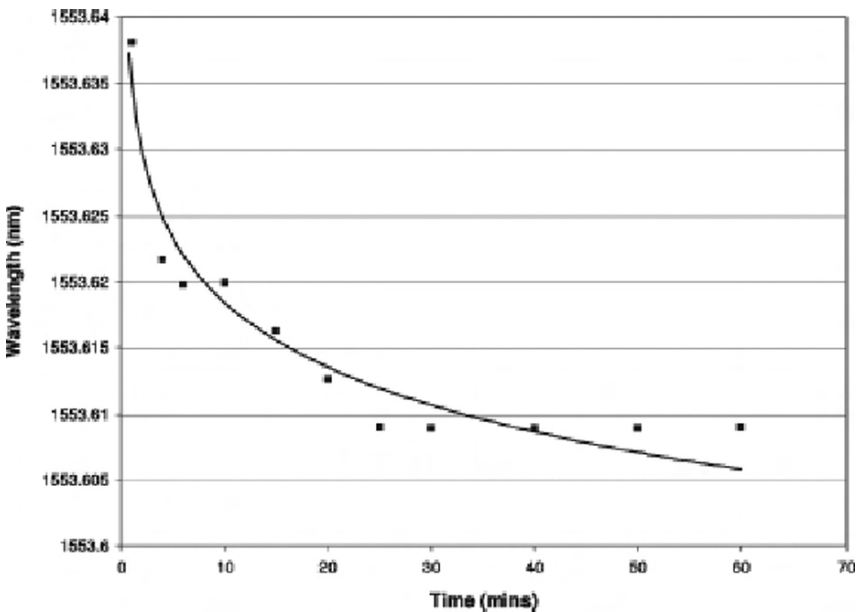


Fig. 25 Shift of wavelength during the hybridization of the target
 (Source: A. N. Chryssis, S. S. Saini, S. M. Lee, Y. Hyunmin, W. E. Bentley, and M. Dagenais, (2005) ‘Detecting hybridization of DNA by highly sensitive evanescent field etched core fiber Bragg grating sensors,’ IEEE J. Sel. Topics Quantum Electron., vol. 11, no. 4, pp. 864–872)

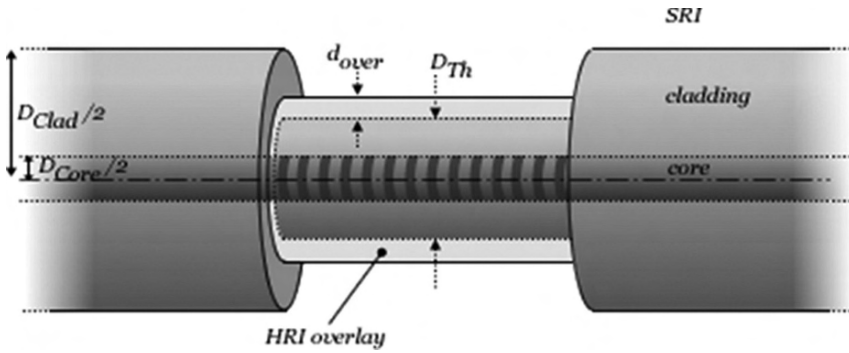


Fig. 26 Schematic of coated ThFBGs

Thus, for chemical and biological sensing where the index of the surrounding medium is less than the core of the fiber, maximum sensitivity to a change of the surrounding index is achieved by monitoring the highest order mode supported by the etched fiber. Also, it was shown that the change of wavelength for both the higher order and the fundamental modes is nearly same when:

- 1) the fiber grating period is altered by physically stretching the fiber or by thermal expansion
- 2) with a change of the core index.

This property can be used to separate a pure index change of the surrounding medium from temperature and stress effects.

A different method to enhance the SRI sensitivity of ThFBGs is the deposition of nano-sized HRI overlays along the thinned region (see Fig. 26) [61, 62]. HRI overlays, indeed, modifying the field distribution of core and cladding modes of the provide a simple method to strongly enhance the sensitivity characteristics towards SRI measurements. Sensitivity enhancement up to one order of magnitude are indeed possible with a suitable choice of the overlay features (see Fig. 27).

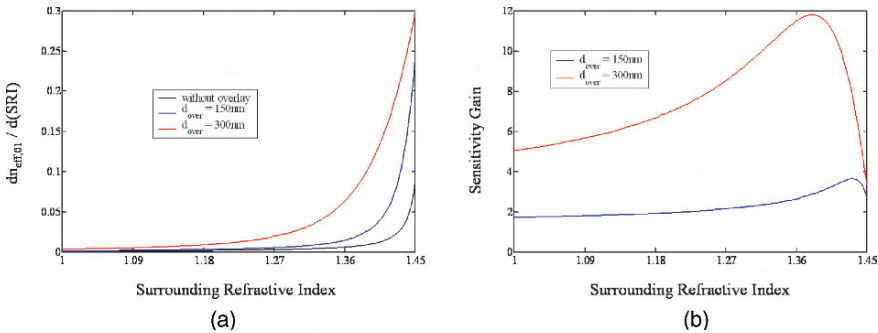


Fig. 27 (a) Sensitivity characteristics in HRI nano-coated ThFBGs (overlay refractive index 1.6), (b) Sensitivity gain

Also, it was proved that the same solution can be also used to enhance the sensing characteristics towards the sensing properties of the thin overlay (thickness and refractive index) useful when chemical or biological sensitive coatings are employed [62].

4.3 Tilted FBGs

A different approach to develop fiber optic refractometers involving FBG technology was proposed 2001 by affont and Fordinand [63]. The idea relies on the use of tilted FBGs (TFBGs) [18] as SRI transducers by using the transmission spectrum changes due to the cladding modes resonances sensitivity to the surroundings. TFBGs belong to the short-period grating family ($0.5 \mu\text{m}$) but their index modulation pattern is blazed by an angle θ with respect to the fiber axis (see Fig. 28) [18]. This asymmetry enables and enhances the coupling to circularly and non-circularly symmetric contra-propagating cladding modes. It reduces in return the energy coupling to the contra-propagating core mode (i.e. the Bragg peak).

Figure 29 gives the transmission spectrum of a 8 mm long $\theta_{\text{ext}} = 16^\circ$ -TFBG photo-written in a standard step-index single-mode optical fiber.

The TFBGs used in this experiment were home-made, photowritten in a standard single-mode step-index fiber (previously 150 bar H2 loaded, for two weeks at room temperature) using an improved Lloyd mirror interferometer. As the external refractive index was changed from $n_{\text{ext}} = 1.0$ to 1.3, the centre wavelength λ_i of the dips experienced a red shift (200 pm) without any significant change of their attenuation. From $n_{\text{ext}} = 1.3$ to 1.43, they drop progressively in addition to their spectral shift, to fit a smooth loss curve. Figures 30(a) and (b) show the transmission spectra obtained for two intermediate values of n_{ext} , i.e. respectively 1.354 and 1.383. When the external refractive index rises and reaches the effective refractive index of the i^{th} cladding mode, this mode becomes weakly guided (due to the decrease of the overlap integral between the fundamental guided mode and the given cladding mode), thereby reducing the amplitude of the coupling coefficient and hence the amplitude of this dip.

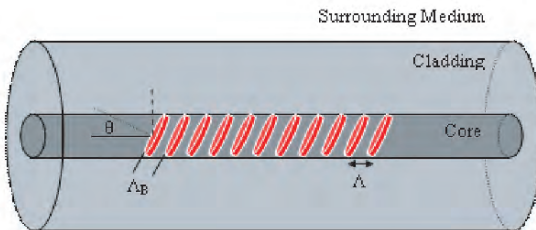


Fig. 28 Schematic of a tilted FBG

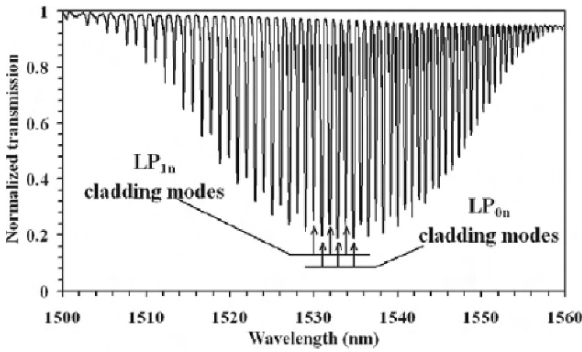


Fig. 29 Experimentally measured transmission spectrum of a 8 mm long $\theta_{\text{ext}} = 16^\circ$ -tilted TFBG surrounded by air

(Source: G Laffont and P Ferdinand, (2001) 'Tilted short-period fibre-Bragg-grating induced coupling to cladding modes fo accurate refractometry' Meas. Sci. Technol. 12 765–770)

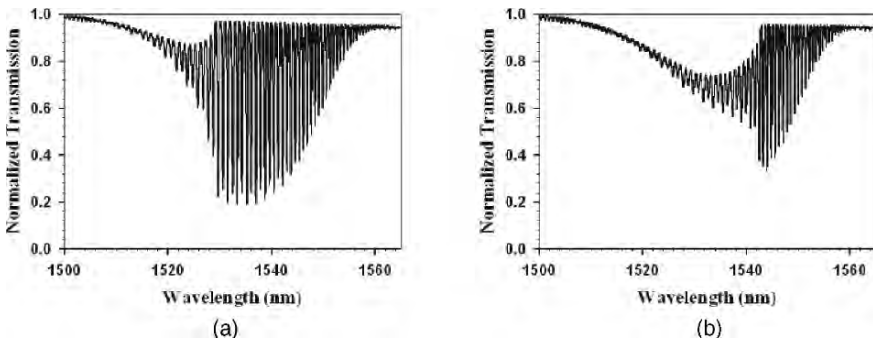


Fig. 30 (a) Measured transmission spectrum of a 8 mm long $\theta_{\text{ext}} = 16^\circ$ -tilted TFBG for an external refractive index value of 1.354, (b) Measured transmission spectrum of a 8 mm long $\theta_{\text{ext}} = 16^\circ$ -tilted TFBG for an external refractive index value of 1.383

(Source: G Laffont and P Ferdinand, (2001) 'Tilted short-period fibre-Bragg-grating induced coupling to cladding modes fo accurate refractometry' Meas. Sci. Technol. 12 765–770)

To take advantage of this phenomenon, an algorithm which determines the lower and upper envelope of the transmission spectrum providing a normalised area A depending on the surrounding refractive index.

Figure 31 shows the results obtained for the five TFBGs with different tilt angle. We can notice that the dynamic can be tailored by properly choosing the tilt angle, also the usual trade-off between range and sensitivity of the refractive index measurement can be clearly inferred. Typical refractive index resolution are of the order of 10^{-4} without the needs of cladding removal typical for SRI sensitization of standard FBGs avoiding fiber weakening. Other difference from previous configurations is the use of the whole transmitted spectrum instead of reflected Bragg wavelength monitoring requiring more complex interrogation units. Finally, it is worth to note that this configuration can be also investigated for SRI higher that the core refractive

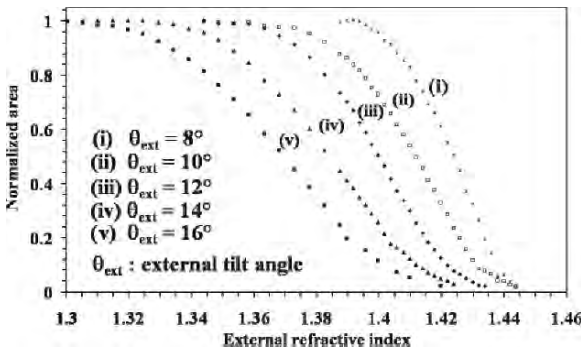


Fig. 31 (a) Evolution of A with the external refractive index for five 8 mm long TFBGs (Source: G Laffont and P Ferdinand, (2001) ‘Tilted short-period fibre-Bragg-grating induced coupling to cladding modes for accurate refractometry’ Meas. Sci. Technol. 12 765–770)

index, which in turn limits the SRI detection for FBGs written in D-shaped fibers and ThFBGs [64]. When employed for high SRI measurements, TFBGs exhibit new resonances in the transmission spectrum differently from cladding removed FBGs where no Bragg reflection is obtained in the same SRI range due to the leaky character of the fundamental mode. These new resonance dips appearing in the transmission spectrum can not be referred to the coupling of light towards guided cladding modes. Indeed, an external refractive index greater than that of the cladding prevents any guided regime in the optical cladding. Actually the core mode is coupled to the attenuated cladding modes [65].

These attenuated cladding modes are formed as a result of the increase in the Fresnel reflection coefficient at the boundary between the fiber cladding and the external medium. Because of the interferences between successive reflections, there exist only a discrete number of such modes.

Recently, sensitivity improvement have been proposed by HF etching in multimode TFBGs or by adopting different demodulation technique [66, 67]. Also, the deposition of HRI overlay provide an useful tool to tune the sensitivity regions providing also an enhancement of the SRI wavelength shift associated to cladding modes coupling [68, 69].

4.4 Micro-Structured FBGs

A novel configuration based on photonic bandgap engineering in FBGs was proposed in 2005 [70]. The approach relies on the use of microstructured FBGs (MSFBGs), a schematic of the basic structure is shown in Fig. 32(a). It consists in a localized stripping of the cladding layer with radial symmetry along the grating structure. The parameters of the structure can be resumed as follows: the etching length L_{Th} , thinned diameter D_{Th} , the unperturbed grating regions lengths on both

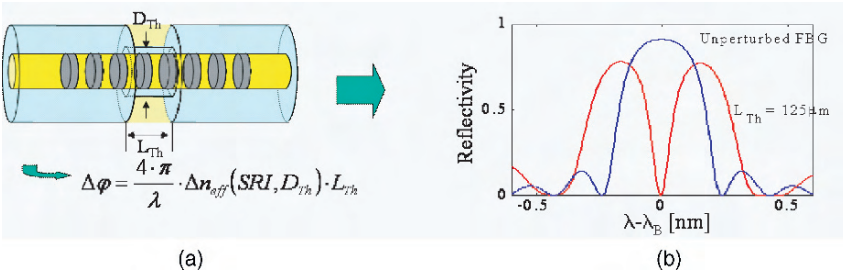


Fig. 32 (a) Schematic of a MSFBG (b) Reflected Spectrum from a 125 μm MSFBG compared with a standard Bragg spectrum

sides of the perturbation L_1 and L_2 respectively, and the cladding and core fiber diameters D_{clad} and D_{co} , respectively.

With regard to Fig. 32(b), the introduction of the defect along the grating leads to strong changes in the reflected spectrum: a band-gap is induced in the stop-band structure of the grating, similarly to the effect observed in Phase-Shift Gratings (PSGs) [71, 72]. Differently from them, micro-structured FBGs exhibit a spectral response dependent on the surrounding refractive index able to tune the defect state within the grating stop-band. The principle of operation relies on the optical beating between the spectra of the unperturbed grating regions modulated by the phase shift induced by the thinned region. The stop-band of the new device increases due to the diminution of the length associated to the two lateral grating regions according to the FBG standard rules. Moreover, the destructive interference of the optical signals reflected from the two lateral gratings leads to the formation of allowed state or defect state inside the band-gap according to the Fabry-Perot effect.

The spectral position of the defect state inside the stop-band is related to the phase delay introduced by the perturbed region strongly affected by the perturbation features and the surrounding refractive index n_{out} (SRI). (see Fig. 33) [73]. As n_{out} changes, a consequent modification of the effective refractive index and thus of the phase delay occurs leading to a red wavelength shift of the defect state. The

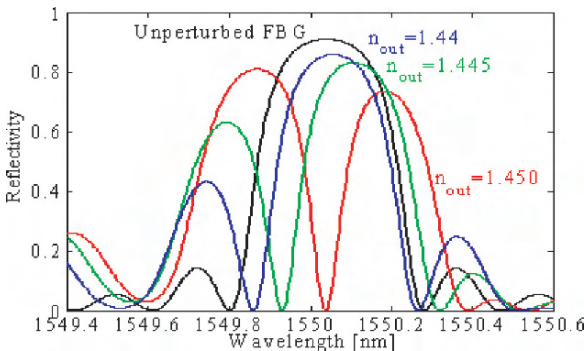


Fig. 33 Change of the defect wavelength due to surrounding refractive index variations

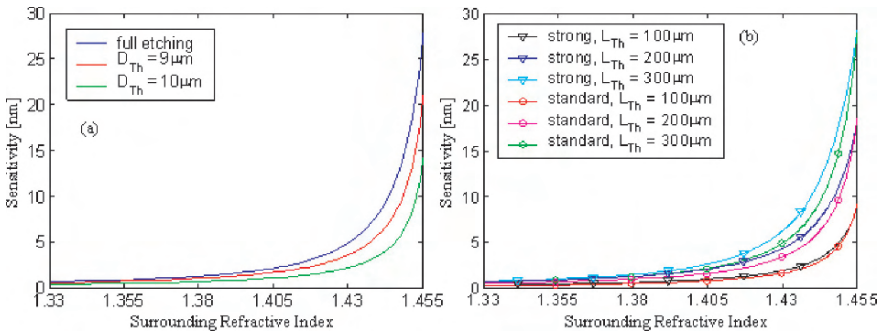


Fig. 34 (a) SRI sensitivity of defect state wavelength in MSFBG for different etching depth, (b) SRI sensitivity of defect state wavelength in MSFBG for different etching length

wavelength shift of the defect state due to n_{out} changes is strongly dependent on the etching length and depth. In particular, as the etching length increases, larger phase shift are induced leading to higher sensitivity of the defect wavelength shift on the surrounding refractive index. On the other side, as the D_{Th} increases, reduced phase shift is induced in the perturbation region, leading to lower sensitivity to n_{out} variations. Figure 34 shows the sensitivity expressed as the derivative of the defect wavelength respect to the SRI for different values of etching depth and length [74].

Since the sensing element is only a limited portion of the overall structure, and thus the main spectral effect relies on the wavelength shift of the defect state without affecting the spectral position of the stop-band [75], intensity measurements based on narrowband interrogation at fixed wavelength seems to be the suitable demodulation strategy to develop low cost and extremely high sensitive in fiber chemical sensors.

The spatial location of the thinned region along the grating length also plays a key role in the reflected spectrum form MSFBGs. Figure 35 shows the comparison between a central thinned region ($L_1 = L_2$) with the case $L_1 = 2 \cdot L_2$, revealing how the asymmetric perturbation induces a weak defect state. This effect is attributable to the unbalance between the reflectivities of the unperturbed grating regions of different length. This means that, in order to work with a well defined defect state inside the original grating stop-band, the device requires a good level of balance between the reflectance of the two lateral grating regions [74].

With regard to the defect bandwidth, it is strongly related to the reflectivity of the original grating. In particular, extremely narrowband features can be achieved using strong gratings (see Fig. 36) [75].

Regarding the optical losses, the thinning of the cladding changes the numerical aperture and losses due to the propagation mode mismatching occur. These coupling losses are responsible for a reduction of the overall reflected power within the stop-band spectrum and, on the other side, for a power unbalancing with effects similar to the case of asymmetric perturbation along the grating. An accurate asymmetric positioning of the thinned region allows compensating for the interferometer unbalancing [74].

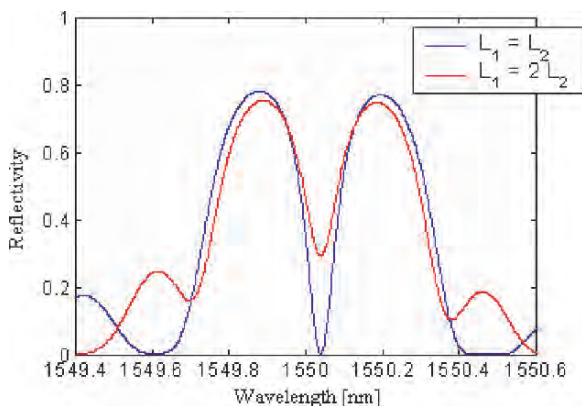


Fig. 35 Spectral Response for asymmetric and symmetric defect for a 125 μm defect MSFBG in case of SRI=1

Particular attention is required for the fabrication of these devices especially with regard to the masking procedure [70]. A first experimental demonstration was carried out by using epoxy resin as masking procedure. However, this first solution would result in a weak control of the etching features leading to low repeatability and reliability. Figure 37 shows the spectral response of a first realization of MSFBGs with a residual diameter of 10.5 μm and central defect 700 μm long for two different values of SRI. Both effects, bandwidth increasing and defect state formation, are evident. Moreover, as the SRI increases, a red shift of the defect state was observed in good agreement with the numerical analysis.

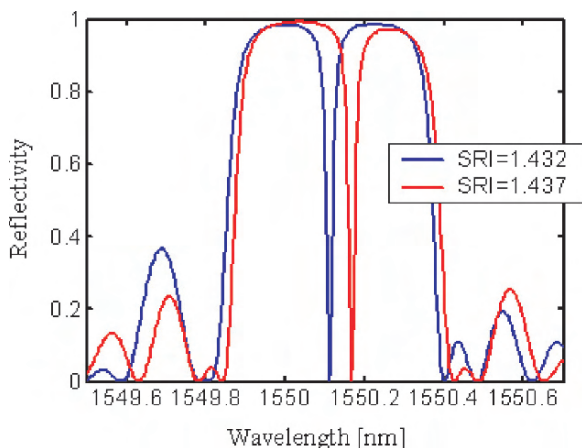


Fig. 36 Reflected Spectra from a MSFBG exploiting strong FBGs for two different values of surrounding refractive index

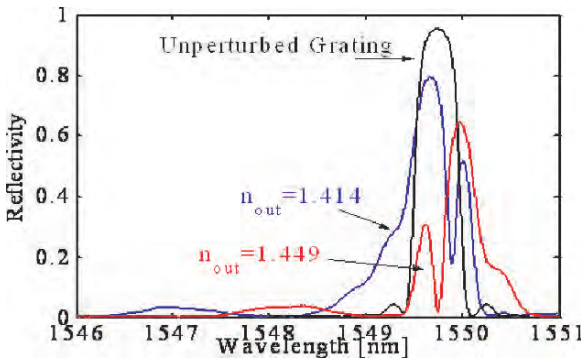


Fig. 37 Reflected Spectra from a MSFBG exploiting strong FBGs for two different values of surrounding refractive index

However, as noticed from experimental results, the wavelength shift is rather limited depending on the original grating bandwidth. This aspect would reduce the measurements performances if detect wavelength monitoring is adopted. Nevertheless, the evidence that SRI changes modify the defect state location leaving unaltered the spectral position suggest to use single and narrowband interrogation [74, 75]. Narrowband interrogation exhibits optimum performances in terms of wavelength resolution compared to typical broadband interrogation techniques. The problem related to this technique is the wavelength stability of the laser and the arrangement used to avoid spurious reflections between the grating itself and the connectors used to join the different fiber optic arms.

Figure 38 shows the relative change in the normalized output signals obtained by the ratio between the reflected signals from the structure and the signal devoted to power monitoring for two operating wavelengths 1549.79 nm and 1549.70 nm, respectively.

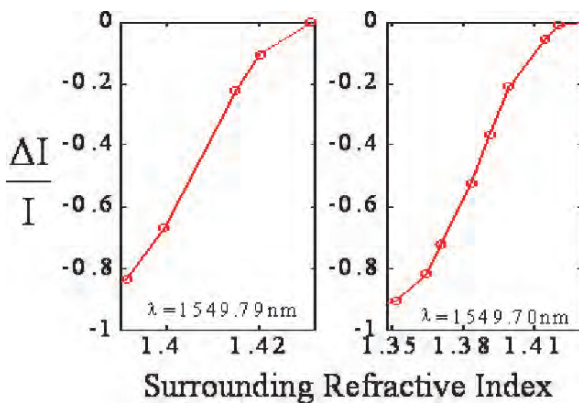


Fig. 38 SRI sensitivity of a 700 μm long defect MSFBG using narrowband interrogation at single optical wavelength

In the investigated refractive index ranges 1.391–1.420 and 1.364–1.40, the normalized signals change of about 75% and 60%, respectively, leading to refractive index resolutions of $4 \cdot 10^{-5}$ and $6 \cdot 10^{-5}$ by using detection units able to resolve 0.1% intensity changes. However, the optimization of the fabrication process combined with integration of microfluidics components is required to provide an advanced technological platform for μ TAS (micro Total Analysis Systems) applications. Successively, other examples of MSFBGs have been demonstrated based on spatially separated FBGs and liquid as filling medium [76], and large beating length MSFBGs [77].

A method to overcome the issues related to the control of the fabrication process was demonstrated using polymeric coatings and UV laser micromachining for accurate definition of the defect features [78]. The base idea of the proposed approach consists on the use of polymeric coatings uniformly deposited along the grating length as protective layers during the successive etching process. To control and define the etched region length, an UV laser micromachining tool was used to selective remove of the coating. Two important requirements in the implementation of the proposed approach are:

- the polymeric coating should absorb light in the UV range and at the same time should avoid the fiber etching in the coated regions of the grating
- the micromachining tool should allow accurate definition of the etching length maintaining the azimuthal symmetry

These requirements have been achieved by using polyamide coatings as protective layers and a special designed UV laser micromachining tool operating at 193 nm able to accurate remove polymeric layers with azimuthal symmetry and a precise spatial control. A schematic of the process is shown in Fig. 39.

The spectral response of a first prototype realized with this solution is shown Fig. 40(a) for different values of SRI, while Fig. 40(b) shows the evolution of the defect wavelength as function of the surrounding refractive index.

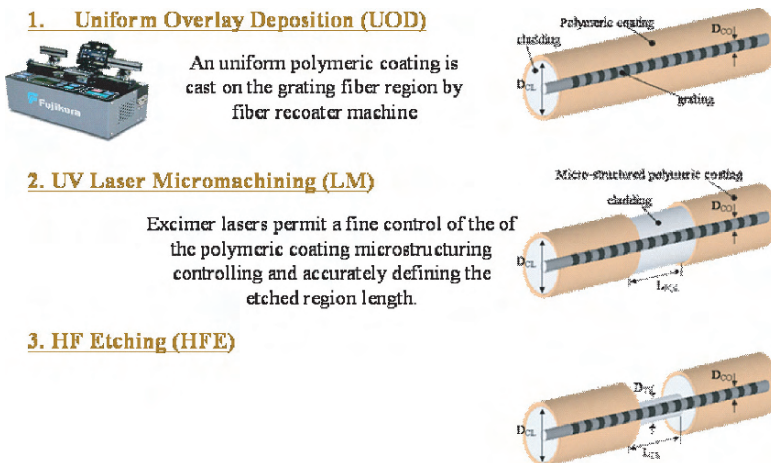


Fig. 39 Schematic of the fabrication process of MSFBGs

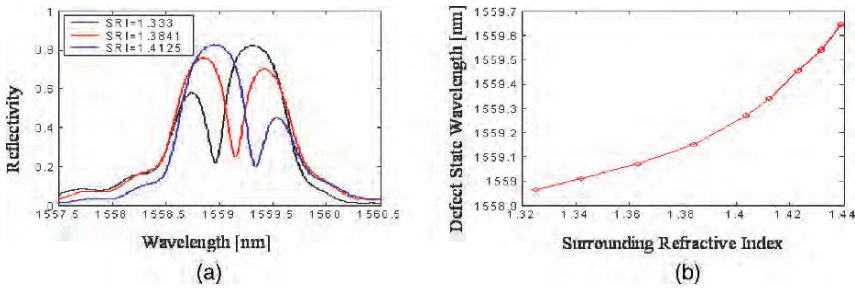


Fig. 40 (a) Reflected spectra of a 190 μm long MSFBG fabricated by using 193 nm laser micro-machining and HF etching, (b) SRI sensitivity of the defect state wavelengths

As observable, good spectral characteristics have been obtained enabling also the fabrication of more complex structures based on multi defect MSFBGs and Chirped MSFBGs [79, 80]. Also, HRI coated MSFBGs could be used to enhance the sensing performances of the basic device as reported in [74, 75, 76, 77, 78, 79].

Recently, techniques using tightly focused femtosecond (fs) laser pulses to produce micro-structures in silica/glass materials have become the focus of research interest. Such processes have been used to induce refractive changes for writing optical waveguides [81] and, with the aid of chemical etching, have been applied in micromachining [82]. It has been reported that regions treated by the fs laser exhibit a remarkably high etching rate compared to pristine material, with contrast ratios up to 100:1 [83]. On this line of argument, Bennion et al. demonstrated a novel micro-slot based FBG refractometer realised by using fs pulses ($\lambda = 800\text{ nm}$) tightly focused on the fiber [84]. The basic structure is shown in Fig. 41. Chemical etching assisted fs laser inscription technique was used to create a $1.2 \times 125 \times 500\text{ }\mu\text{m}$ micro-slot across an FBG. Compared to evanescent wave based refractometers, the working RI range of the device is extended to 1.55 and sensitivity up to 945nm/RIU (or $1.0 \times 10^{-6}/\text{pm}$) is achieved, which is comparable to LPG sensors. The device has great potential in medical, chemical sensing applications.

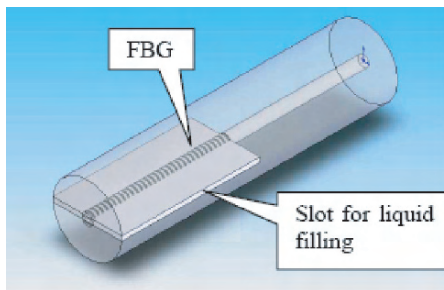


Fig. 41 Schematic of a micro-slot based FBG refractometer
 (Source: Kaiming Zhou, Yicheng Lai, Xianfeng Chen, Kate Sugden, Lin Zhang, Ian Bennion, (2007) ‘A refractometer based on a micro-slot in a fiber Bragg grating formed by chemically assisted femtosecond laser processing’, Opt. Exp. Vol. 15, No. 24 15848)

4.5 Coated FBGs Based on Chemo-Mechanical and Chemo-Thermal Effects

Finally, we report on a different method to chemically sensitize FBGs sensors dealing the deposition of swelling overlays which in turn are able to provide mechanical stress on the grating structure when interact with target analytes. Strain sensitivity is thus used to detect small concentration of chemicals rather than the effective refractive index sensitivity due to evanescent wave interaction. It is important to note that this approach is limited to few examples since a strong energy and thus chemical interactions are required to induce appreciable strain state over the sensing grating.

FBGs covered with palladium have been widely investigated in the past for H₂ sensing [85, 86, 87, 88]. The sensing mechanism is based on the swelling of the Pd-coating, resulting in a stress on the grating. In practice, the Pd-coated sensors suffer from a long response time leading moreover to a hysteresis effect between the responses obtained for increasing and decreasing hydrogen concentrations. A solution has been proposed in [89] to overcome this drawback but it still remains the problem of the environment. Indeed, such sensors have been tested in nitrogen environments and they cannot work properly in air, whereas applications such as the monitoring of storage places and pipes require the use of H₂ sensors in air. An interesting investigation of coated FBGs sensors for hydrogen detection was recently reported by Guemes et al. reporting the results obtained within the framework of the European Space Agency Project: CryFOS (16199/02/NL/ND) [90].

An interesting solution was provided by Caucheteur et al. through the use of FBGs covered by a catalytic sensitive layer made of a ceramic doped with noble metal [91]. In the presence of hydrogen in air, the sensitive layer is the siege of an exothermic reaction and an increase of temperature around the FBG is measured through a shift of its central wavelength. For a 1% concentration of H₂ in dry air, the measured wavelength shift is equal to 2 nm, which is very easy to detect with a low-cost interrogation system. This is equivalent to a local temperature increase of nearly 200°C. The differentiation between ambient temperature variations and temperature elevations due to H₂ can be provided by a bare FBG few centimeters spaced from the covered FBG.

Other examples based chemo-mechanical stress include the use of swelling polymeric coatings for chemicals and relative humidity detection [92, 93].

5 Perspectives and Challenges

Newer applications in chemical, biological and medical sensing demand mass producible, low cost diagnostic micro and nano systems driving the increasing requirements for optical sensor platforms. Arguably, the most important characteristics of

recent systems are the integration of multiple functionalities onto a single platform, the ability to perform multianalyte detection and the production of low cost technological platforms. Multianalyte detection provide an increased amount of information on the system of interest and, combined with techniques such as chemometrics and multivariate analysis, facilitates the development of devices such as high density biochips and novel concepts of ‘noses’ or ‘tongues’ systems. While FBG technology fits well many of the specified requirements and has now reached a well recognised technological maturity for many industrial fields, fiber geometry still remains the main drawback due to the weak compatibility with a broad range of microfabrication/deposition technologies. Also, actual evanescent wave sensors schemes are difficult to be directly considered as reliable technological platforms for robust, disposable commercial product.

However, considerations based on:

- the technological maturity of fiber Bragg grating sensors and related interrogators
- the increasing portfolio of evanescent wave sensors configurations with enhanced and tailorable performances employable in single or multisensor configuration
- the intrinsic capability to perform multi parameters measurements providing unique self referencing features

Are expected to be the driving force to fill the technological gap that separates academic research from field applications providing valuable chemical and biological platforms.

Future efforts will thus rely heavily on cost reduction and development of specialized and application-specific packaging. The prospects of using polymer optical fibers (POF) in sensing applications is expected to open up the door to the development of POF FBGs to be used as inexpensive, simple and low-cost disposable platforms. Similarly, microstructured optical fibers (MOF) are expected to have a major role in the development of new chemical and biological systems based on optofluidics as well as active and passive microfluidics. This scenario prospects a new generation of devices and products that will perform specific agent or parameter sensing functions relying on specially developed and integrated coatings and reagents. To this aim, great effort will be focus toward encapsulating recent exciting developments in the are of nano materials and nanostructures with high functionality degree as well as new integration methodologies taking into account different geometries and constrains whilst in parallel covering the continually expanding world of field trials and application assessments, with special attention to explore new perspectives and outline technological challenges. A novel generation of fiber optic nano-devices for chemical and biological sensing is emerging based on the concurrent addressing of the issues related to the different aspects of a global design concept. A highly integrated approach involving continuous interactions of different backgrounds aimed to optimize each single aspect with a continuous feed-back would enable the definition of valuable technological nano-biophotonics platforms for future applications.

References

1. B. Culshaw and J. Dakin (1988) "Optical Fiber Sensors: Principle and Components", Artech House inc., Norwood.
2. E. Udd (1991) "Fiber Optic Sensors: An Introduction for Engineers and Scientists", John Wiley and Sons, New York.
3. B. Culshaw and J. Dakin (1997) "Optical Fiber Sensors: Applications, Analysis, and Future Trends", Artech House inc., Norwood.
4. B. Culshaw and J. Dakin (1996) "Optical Fiber Sensors: Components and Subsystems", Vol. 3, Artech House inc., Norwood.
5. E. Udd (1995) "Fiber Optic Smart Structures", John Wiley and Sons, New York.
6. R. M. Measures (2001) "Structural Monitoring with Fiber Optic Technology", Academic Press, London.
7. K. O. Hill, Y. Fujii, D. C. Johnson, and B. S. Kawasaky (1978) "Photosensitivity in Optical Fiber Waveguides: Applications to Reflection Filter Fabrication", *Appl. Phys. Lett.* Vol. 32, pp. 647–649.
8. G. Meltz, W. W. Morey, and W. H. Glam (1989) "Formation of Bragg Grating in Optical Fibers by a Transverse Holographic Method", *Opt. Lett.* Vol. 14, pp. 823–825.
9. D. Lam, and B. Garside (1981) "Characterization of Single Mode Optical Fibers", *App. Opt.* Vol. 20, pp. 440–445.
10. K. O. Hill and G. Meltz (1997) "Fiber Bragg Grating Technology Fundamentals and Overview", *J. Lightwave Technol.*, Vol. 15, pp. 1263–1276.
11. A. Yariv (1973) "Coupled-Mode Theory for Guided-Wave Optics", *IEEE J. Quantum Electron.*, Vol. QE-9, pp. 919–933.
12. T. Erdogan (1997) "Fiber Grating Spectra", *J. Lightwave Technol.*, Vol. 15, No. 8, pp. 1277–1294.
13. P. J. Lemaire, R. M. Atkins, V. Mizrahi, K. L. Walker, K. S. Kranz, and W. A. Reed (1993) "High Pressure H₂ Loading as a Technique for Achieving Ultrahigh UV Photosensitivity and Thermal Sensitivity in GeO₂ Doped Optical Fibers", *Electron. Lett.*, Vol. 29, pp. 1191–1193.
14. K. O. Hill et al. (1993) "Bragg Gratings Fabricated in Monomode Photosensitive Optical Fibers by UV Exposure Through a Phase Mask", *Appl. Phys. Lett.*, Vol. 62, pp. 1035–1037.
15. A. Othonos, et al. (1995) "Novel and Improved Methods of Writing Bragg gratings with Phase Mask", *EEE Phot. Techn. Lett.*, Vol. 7, pp. 1183–1185.
16. B. Malo, S. Theriault, D. C. Johnson, F. Bilodeau, J. Albert, and K. O. Hill (1995) "Apodised In-Fiber Bragg Grating Reflectors Photoimprinted using a Phase Mask," *Electron. Lett.*, Vol. 31, pp. 223–225.
17. J. E. Sipe, L. Poladian, and C. M. de Sterke (1994) "Propagation Through Nonuniform Grating Structures", *J. Opt. Soc. Amer. A*, vol. 11, pp. 1307–1320.
18. T. Erdogan and J. E. Sipe (1996) "Tilted Fiber Phase Gratings", *J. Opt. Soc. Amer. A*, Vol. 13, pp. 296–313.
19. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, T. Erdogan, and J. E. Sipe (1996) "Long-Period Fiber Gratings as Band-Rejection Filters", *J. Lightwave Technol.*, Vol. 14, pp. 58–65.
20. W. H. Loh and R. I. Laming (1995) "1.55 μm Phase-Shifted Distributed Feedback Fiber Laser," *Electron. Lett.*, Vol. 31, pp. 1440–1442.
21. B. J. Eggleton, P. A. Krug, L. Poladian, and F. Ouellette (1994) "Long Periodic Superstructure Bragg Gratings in Optical Fibers", *Electron. Lett.*, Vol. 30, pp. 1620–1622.
22. T. Erdogan (1997) "Cladding-Mode Resonances in Short and Long Period Fiber Grating Filters", *J. Opt. Soc. Amer. A*, Vol. 14, No. 8, Aug.
23. S. M. Melle, K. Liu, and R. Measures (1992) "A Passive Wavelength Demodulation System for Guided-Wave Bragg Grating Sensors", *IEEE Phot. Technol. Lett.*, Vol. 4, No.5, pp. 516–518.

24. P. St. J. Russel and J. L. Archambault (1997) "Fiber Gratings" in *Optical Fiber Sensors*, B. Culshaw and J. Dakin Eds., Artech House, pp. 9–67.
25. A. D. Kersey, M. A. Davis, H. J. Patrick, M. LeBlanc, K. P. Koo, C. G. Askins, M. A. Putnam and E. J. Friebele (1997) "Fiber Grating Sensors", *J. Lightwave Technol.*, Vol. 15, pp 1442–1462.
26. A. Othonos and K. Kalli (1999) "Fiber Bragg Gratings Fundamentals and Applications in Telecommunications and Sensing". Artech House, Boston.
27. R. Kashyap (1999) "Fiber Bragg Gratings" Academic Press, San Diego.
28. A. Mendez (2007) "Fiber Bragg grating Sensors: A Market Overview", *Proceedings Vol. 6619 Third European Workshop on Optical Fibre Sensors*, Antonello Cutolo; Brian Culshaw; José Miguel López-Higuera, Editors, 661905.
29. P. N. Prasad (2003) "Introduction to Biophotonics", John Wiley & Sons, Hoboken, New Jersey.
30. F. Baldini, A. N. Chester, J. Homola, and S. Martellucci, Eds. (2004) "Optical Chemical Sensors", NATO Science Series Vol. 224.
31. G. Meltz, S. J. Hewlett, and J. D. Love (1996) "Fiber Grating Evanescent-Wave Sensors", *Proceedings of SPIE Vol. 2836 Chemical, Biochemical, and Environmental Fiber Sensors VIII*, pp. 342–350.
32. S. M. Tseng and Ch. L. Chen (1992) "Side-Polished Fibers", *Appl. Opt.*, Vol. 31, pp. 3438–3447.
33. K. Usbeck, W. Ecke, A. Andreev, V. Hagemann, R. Mueller, and R. Willsch (1998) "Distributed Optochemical Sensor Network Using Evanescent Field Interaction in Fiber Bragg Gratings", *Proceedings of 1st European Workshop on Optical Fibre Sensors, 08–10 July 1998, Peebles, Scotland, SPIE Vol. 3483*, pp. 90–94.
34. K. Schröder, W. Ecke, R. Mueller, R. Willsch, and A. Andreev (2001) "A Fibre Bragg Grating Refractometer", *Meas. Sci. Technol.*, Vol. 12, pp. 757–764.
35. A. Cusano, A. Iadicicco, P. Pilla, L. Contessa, S. Campopiano, and A. Cutolo (2005) "Cladding Mode Reorganization in High-Refractive Index-Coated Long-Period Gratings: Effects on the Refractive-index Sensitivity", *Opt. Lett.*, Vol. 30, No. 19, October 1, pp. 2536–2538.
36. S. W. James, N. Rees, R. P. Tatam, and G. J. Ashwel (2002) "Optical Fiber Long Period Gratings With Langmuir-Blodgett Thin Film Overlays", *Opt. Lett.*, 9, pp. 686–688.
37. I. Del Villar, I. Matías, F. Arregui, and P. Lalanne (2005). "Optimization of Sensitivity in Long Period Fiber Gratings with Overlay Deposition", *Opt. Express*, Vol. 13, pp. 56–69.
38. A. Cusano, A. Iadicicco, P. Pilla, L. Contessa, S. Campopiano, A. Cutolo, and M. Giordano (2006) "Mode Transition in High Refractive Index Coated Long Period Gratings", *Opt. Express*, Vol. 14, pp. 19–34.
39. T. Mizunami, T. V. Djambova, T. Niiho, and S. Gupta (2000) "Bragg gratings in multimode and Few-Mode Optical Fibers", *J. Lightwave Technol.*, Vol. 18, No. 2, pp. 230–235.
40. J. Homola, S. S. Yee, and G. Gauglitz (1999) "Surface Plasmon Resonance Sensors: Review", *Sens. and Actuators B*, Vol. 54, pp. 3–15.
41. J. Čtyroký, F. Abdelmalek, W. Ecke, and K. Usbeck (1999) "Modelling of the Surface Plasmon Resonance Waveguide Sensor with Bragg Grating", *Opt. Quan. Electron.*, Vol. 31, pp. 927–941.
42. J. Čtyroký, W. Ecke, K. Schroeder, and R. Slavík (2000) "Separation of Refractive Index and Temperature Measurements Using Surface Plasmon-Coupled Fiber Grating" *Proceedings of SPIE Vol. 4185 "OFS2000"*, Eds. A.G. Mignani and H.C. Lefevre, pp. 322–325, Venice.
43. R. Willsch, W. Ecke, G. Schwotzer, H. Bartelt (2007) "Nanostructure-Based Optical Fibre Sensor Systems and Examples of their Application" *Proceedings of SPIE Volume: 6585 Francesco Baldini, Jiri Homola, Robert A. Lieberman, Miroslav Miler Eds.*
44. K. Zhou, X. Chen, L. Zhang, and I. Bennion (2004) "High-Sensitivity Optical Chemsonor Based on Etched D-Fibre Bragg Gratings", *Electron. Lett.* Vol. (4), pp. 232–234.

45. M. A. Jensen and R. H. Selfridge (1992) "Analysis of Etching Induced Birefringence Changes in Elliptic Core Fibers", *Appl. Opt.*, Vol. 31, pp. 211–216.
46. K. H. Smith, B. L. Ipson, T. L. Lowder, A. R. Hawkins, R. H. Selfridge, and S. M. Schultz (2006) "Surface-Relief Fiber Bragg Gratings for Sensing Applications" *Appl. Opt.* Vol. 45, pp. 1669.
47. T. L. Lowder, J. D. Gordon, S. M. Schultz, and R. H. Selfridge (2007) "Volatile Organic Compound Sensing Using a Surface Relief D-Shaped Fiber Bragg Grating and a Polydimethylsiloxane Layer" *Opt. Lett.* Vol. 32, No. 17, pp. 2523–2525.
48. A. Asseh, S. Sandgren, H. Ahlfeldt, B. Sahlgren, R. Stubbe, and G. Edwall, (1998) "Fiber Optical Bragg Grating Refractometer", *Fiber and Integrated Optics*, Vol. 17, pp. 51–62.
49. A. Iadicicco, A. Cusano, G.V. Persiano, A. Cutolo, R. Bernini and M. Giordano (2003) "Refractive Index Measurements by Fiber Bragg Grating Sensor" *Proceedings of IEEE Sensors Conference*, Vol. 1, pp. 101–105, Toronto, Canada, October 2003.
50. A. Iadicicco, A. Cusano, A. Cutolo, R. Bernini, M. Giordano (2004) "Thinned Fiber Bragg Gratings as High Sensitivity Refractive Index Sensor" *IEEE Phot. Technol. Lett.*, Vol. 16(No.4), pp. 1149 – 1151.
51. A. Iadicicco, A. Cusano, S. Campopiano, A. Cutolo, and M. Giordano, (2005) "Thinned Fiber Bragg Gratings as Refractive Index Sensors" *IEEE Sens. J.*, Vol. 5, No. 6, pp. 1288–1295.
52. D. Marcuse (1991) "Theory of Dielectric Optical Waveguides". Academic, New York.
53. D. A. Pereira, O. Frazao, and J. L. Santos (2004) "Fiber Bragg Grating Sensing System for Simultaneous Measurement of Salinity and temperature", *Opt. Eng.*, Vol. 43, No. 2, pp. 299–304.
54. A. Iadicicco, S. Campopiano, A. Cutolo, M. Giordano, and A. Cutolo (2005) "Non-Uniform Thinned Fiber Bragg Gratings for Simultaneous Refractive Index and Temperature Measurements" *IEEE Phot. Technol. Lett.*, Vol. 17, No. 7, pp. 1495–1497.
55. A. Iadicicco, S. Campopiano, A. Cutolo, M. Giordano, and A. Cusano (2006) "Self Temperature Referenced Refractive Index Sensor by Non-Uniform Thinned Fiber Bragg Gratings" *Sens. Actuators B: Chem.*, Vol. 120, No. 1, pp. 231–237.
56. A. Iadicicco, S. Campopiano, A. Cutolo, M. Giordano, and A. Cusano (2005) "Simultaneous Measurements of Refractive Index and Temperature by Non-Uniform Thinned Fiber Bragg Gratings" *Proceedings of SPIE*, Vol. 5855, pp. 479–482, Bruges.
57. A. N. Chryssis, S. M. Lee, S. B. Lee, S. S. Saini, and M. Dagenais (2005) "High Sensitivity Evanescent Field Fiber Bragg Grating Sensor" *IEEE Photon. Technol. Lett.*, Vol. 17, No. 6, pp. 1253–1255.
58. A. N. Chryssis, S. S. Saini, S. M. Lee, Y. Hyunmin, W. E. Bentley, and M. Dagenais (2005) "Detecting Hybridization of DNA by Highly Sensitive Evanescent Field Etched core Fiber Bragg Grating Sensors", *IEEE J. Sel. Topics Quantum Electron.*, Vol. 11, No. 4, pp. 864–872.
59. A. N. Chryssis, S. S. Saini, S. M. Lee, and M. Dagenais (2006) "Increased Sensitivity and Parametric Discrimination Using Higher Order Modes of Etched-Core Fiber Bragg Grating Sensors", *IEEE Photon. Technol. Lett.*, Vol. 18, No. 1, pp. 178–180.
60. N. Chen, B. Yun, and Y. Cui (2006) "Cladding Mode Resonances of Etch-Eroded Fiber Bragg Grating for Ambient Refractive Index Sensing" *Appl. Phys. Lett.* Vol. 88, pp. 133902.
61. A. Cusano, A. Iadicicco, P. Pilla, A. Cutolo, M. Giordano, and S. Campopiano (2006) "Sensitivity characteristics in Nanosized Coated Long Period Gratings", *Applied Physics Letters* Vol. 89, pp. 201116.
62. D. Paladino, A. Iadicicco, A. Cutolo, S. Campopiano, M. Giordano, and A. Cusano (2006) "Nano-Scale High Refractive Index Coated Thinned FBGs for Sensing Applications" *Proceedings of the 18th Optical Fiber Sensors Conference SPIE Cancún, México*.
63. G. Laffont and P. Ferdinand, (2001) "Tilted Short-Period Fibre-Bragg-Grating induced Coupling to Cladding Modes for Accurate Refractometry" *Meas. Sci. Technol.* Vol. 12, pp. 765–770.
64. O. Duhem, J.-F. Henninot, M. Warenghem, and M. Douay (1998) "Demonstration of long-period-grating efficient couplings with an external medium of a refractive index higher than that of silica", *Appl. Opt.*, Vol. 37, No. 31, pp. 7223–7228.

65. G. Laffont and P. Ferdinand (2001) "Sensitivity of Slanted Fibre Bragg Gratings to External Refractive Index Higher Than that of Silica" *Electronics Letters* 1st March 2001 Vol. 37, pp. 321–328.
66. X. Chen, K. Zhou, L. Zhang, and I. Bennion (2005) "Optical Chemsensor Based on Etched Tiled Bragg Grating Structures in Multimode Fiber", *IEEE Phot. Technol. Lett.*, 17(4), 864–866.
67. C. Caucheteur and P. Mégret, (2005) "Demodulation Technique for Weakly Tilted Fiber Bragg Grating Refractometer," *IEEE Phot. Technol. Lett.*, Vol. 17, No. 12, pp. 2703–2705.
68. D. Paladino, P. Pilla, A. Cutolo, S. Campopiano, M. Giordano, A. Cusano, C. Caucheteur, P. Mégret (2007) "Effects of Thickness and External Refractive Index in Coated Tilted Fiber Bragg Gratings" *Proceedings of SPIE* Vol. 6619, pp. 68.
69. D. Paladino, A. Cusano, P. Pilla, S. Campopiano, C. Caucheteur, and P. Mégret (2007) "Spectral Behaviour in Nano-Coated Tilted Fiber Bragg Gratings: Effect of Thickness and External Refractive Index" *IEEE Phot. Technol. Lett.*, Vol. 19, No. 24, pp. 2051–2053.
70. A. Iadicicco, A. Cusano, S. Campopiano, A. Cutolo, and M. Giordano (2005) "Microstructured Fiber Bragg Gratings: Analysis and Fabrication", *IEEE Electronics Lett.*, Vol. 41, No. 8, pp. 466–468.
71. L. Wei and J.W.Y. Lit (1997) "Phase Shifted Bragg Grating Filters with Symmetrical Structures", *J. Lightwave Technol.*, Vol. 15, No. 8, pp. 421–426.
72. R. Zengerle and O. Leminger, (1995) "Phase Shifted Bragg-grating Filters with Improved Transmission Characteristics", *J. Lightwave Technol.*, Vol. 13, No. 12, pp. 543–549.
73. A. Cusano, A. Iadicicco, S. Campopiano, M. Giordano, and A. Cutolo (2005) "Thinned and Micro-Structured Fiber Bragg Gratings: Towards New All Fiber High Sensitivity Chemical Sensors", *J. Opt., A: Pure App. Opt.* Vol. 7, pp. 734–741.
74. A. Cusano, A. Iadicicco, D. Paladino, S. Campopiano, A. Cutolo, and M. Giordano, (2007) "Micro-Structured Fiber Bragg Gratings. Part I: Spectral Characteristics", *Opt. Fiber Technol.*, Vol. 13, No. 4, pp. 281–290.
75. A. Iadicicco, A. Cusano, S. Campopiano, A. Cutolo, and M. Giordano, (2005) "Refractive Index Sensor Based on Micro-Structured Fiber Bragg Grating", *IEEE Phot. Technol. Lett.* Vol. 17, No. 5, pp. 1250–1252.
76. P. Domachuk, I. C. M. Littler, M. Cronin-Golomb, and B. J. Eggleton (2006) "Compact Resonant Integrated Microfluidic Refractometers", *Appl. Phys. Lett.* Vol. 88, pp. 093513.
77. W. Liang, Y. Huang, Y. Xu, R. K. Lee, and A. Yariv (2005), "Highly Sensitive Fiber Bragg Grating Refractive Index Sensors", *Appl. Phys. Lett.* Vol. 86, pp. 151122.
78. A. Iadicicco, S. Campopiano, D. Paladino, A. Cutolo, and A. Cusano, (2007) "Micro-Structured Fiber Bragg Gratings: Optimization of the Fabrication Process", *Opt. Expr.*, Vol. 15, No. 23, pp. 15011–15021.
79. A. Cusano, A. Iadicicco, D. Paladino, S. Campopiano, A. Cutolo, and M. Giordano (2007) "Micro-Structured Fiber Bragg Gratings. Part II: Towards Advanced Photonic Devices" *Opt. Fiber Technol.*, Vol. 13, No. 4, pp. 291–301.
80. M. Pisco, A. Iadicicco, S. Campopiano, A. Cutolo, and A. Cusano (2007) "Micro-Structured Chirped Fiber Bragg Gratings: Towards New Spatial Encoded Fiber Optic Sensors", *Proceedings of SPIE*, Vol. 6619, 66192T.
81. K. Davis, K. Miura, N. Sugimoto, and K. Hirao, (1996) "Writing Waveguides in Glass with a Femtosecond Laser", *Opt. Lett.*, Vol. 21, pp. 1729–1731.
82. C. B. Schaffer, A. Brodeur, J. F. García, and E. Mazur (2001) "Micromachining Bulk Glass by Use of Femtosecond Laser Pulses with Nanojoule Energy", *Opt. Lett.* Vol. 26, pp. 93–95.
83. C. Hnatovsky, R. S. Taylor, E. Simova, V. R. Bhardwaj, D. M. Rayner, and P. B. Corkum, (2005) "Polarizationselective Etching in Femtosecond Laser-Assisted Microfluidic Channel Fabrication in Fused Silica", *Opt. Lett.*, Vol. 30, pp. 1867–1869.
84. K. Zhou, Y. Lai, X. Chen, K. Sugden, L. Zhang, I. Bennion (2007) "A Refractometer Based on a Micro-Slot in a Fiber Bragg Grating Formed by Chemically Assisted Femtosecond Laser Processing", *Opt. Exp.* Vol. 15, No. 24, p. 15848.

85. B. Sutapun, M. Tabib-Azar, and A. Kazemi (1999) "Pd-Coated Elastooptic Fiber Optic Bragg Grating Sensors for Multiplexed Hydrogen Sensing", *Sens. Actuators B*, Vol. 60, pp. 27–34.
86. Y. Tang, T. Peng, J. S. Sirkis, B. A. Childers, J. P. Moore, and L. D. Melvin (1999) "Characterization of a Fiber Bragg Grating (FBG)-Based Palladium Tube Hydrogen sensor", in *Proc. SPIE Smart Structures Mater. Conf.*, 1999, vol. 3670, pp. 532–540.
87. Y. T. Peng, Y. Tang, and J. S. Sirkis (1999) "Hydrogen Sensors Based on Palladium Electroplated Fiber Bragg Gratings," in *Proc. SPIE 13th Int. Conf. Opt. Fiber Sensors Workshop Device Syst. Technol. Toward Future Opt. Fiber Commun. Sensing*, Vol. 3746, pp. 171–179.
88. R. R. J. Maier, J. S. Barton, J. D. C. Jones, S. McCulloch, B. J. S. Jones, and G. Burnell, (2006) "Palladium-Based Hydrogen Sensing for Monitoring of Ageing Materials," *Meas. Sci. Technol.*, Vol. 17, pp. 1118–1123.
89. M. Buric, K. P. Chen, M. Bhattarai, P. R. Swinehart, and M. Maklad (2007) "Active Fiber Bragg Grating Hydrogen Sensors for All-Temperature Operation," *IEEE Photon. Technol. Lett.*, Vol. 19, No. 5, pp. 255–257.
90. A. Guemes, J. M. Pintado, M. Frovel, E. Olmo, and A. Obst (2005) "Comparison of Three Types of Fibre Optic Hydrogen sensors Within the Frame of CryoFOS Project", *Proceedings of the 17th International Conference on Optical Fibre Sensors*, SPIE 5855 (SPIE Bellingham, WA, 1000–1003).
91. C. Caucheteur, M. Debliquy, D. Lahem, and P. Mégret (2008) "Catalytic Fiber Bragg Grating Sensor for Hydrogen Leak Detection in Air", *IEEE Phot. Technol. Lett.*, Vol. 20, No. 2, pp. 240–246.
92. G. B. Tait, G. C. Tepper, D. Pestov, and P. M. Boland, (2005) "Fiber Bragg Grating Multi-Functional Chemical Sensor", *Proceedings, 2005 SPIE International Symposium Optics East (Boston, MA)*.
93. T. L. Yeo, T. Sun, K. T. V. Grattan, D. Parry, R. Lade, and B. D. Powell, (2005), "Characterisation of a Polymer-Coated Fiber Bragg Grating Sensor for Relative Humidity Sensing," *Sens. Actuators B*, 110, 148–155.

Optical Fibre Humidity Sensors Using Nano-films

Jesus M. Corres, Ignacio R. Matias and Francisco J. Arregui

Abstract This chapter attempts to approach the fibre optic humidity sensing technology to scientists unfamiliar with the field. A general review of this type of sensors is presented here with emphasis in the techniques based on nanostructured coatings. These devices have been classified according to the sensing mechanism and taking also into account the different methods of fabrication and the sensing materials they are based on.

Keywords Optical fibre sensor · humidity sensor · humidity sensitive nano-films

1 Introduction

The monitoring of humidity is a necessary activity in numerous fields of industry because it may affect both the product, and the health and security of the workers; for instance humidity measurement can be of vital importance in chemical and biomedical industries. Also, humidity sensing is frequently monitored in big structures such as bridges or planes to control possible risk of leakage due to corrosion [1, 2]. Therefore relative humidity measurement has been extensively studied and a great variety of sensors, including capacitive, resistive, thermal conductivity and optical have been developed along the last decades.

So far, electronic humidity sensors cover the main part of the sensors market because their technology of fabrication is very established. However, the field of

Jesus M. Corres
Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra,
Pamplona, Spain, e-mail: jmcorres@unavarra.es

Ignacio R. Matias
Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra,
Pamplona, Spain

Francisco J. Arregui
Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra,
Pamplona, Spain

optical fibre sensors has grown enormously since the 60's and at the present time there exist niches of application where optical fibre humidity sensor technology can advantageously compete with traditional (mainly electronics) technologies.

Fibre optic humidity sensors (FOHS) use optical fibre technology to guide a light signal which is modulated with the ambient humidity and then collected back by a detector, conditioned and processed. Thanks to the low attenuation and large wideband of fibre it is possible to transmit large sensor data quantities over kilometre distances; in addition, the use of several interrogating techniques enable the existence of distributed humidity sensors configurations [3]. There are also a high number of applications where the possible electric hazard produced by the electronic sensor itself or the electromagnetic interferences of the surrounding environment makes difficult the utilization of electronic type humidity sensors. In some fields it is needed a measurement of humidity where the accessibility is limited in space and an electronic sensor could be more difficult to locate. The small dimensions and simple geometries of optical fibre make possible the implementation of light-weight systems that can be easily embedded into construction materials. Finally, the sensitivity, dynamic range and resolution can be potentially much greater than conventional sensors with the use of interferometric techniques.

A number of different approaches have been used in the fabrication of optical fibre humidity sensors. Fibre optic sensors are commonly subdivided into intrinsic (if the transduction between the light and the measurand takes place inside the fibre) and extrinsic (if it takes place outside it). In practice, most of humidity sensors are of extrinsic type because the fibre is immune to humidity, with the exception of special fibres fabricated using porous materials such as sol gel. Depending on the sensing architecture, transmissive and reflexive humidity sensors can be found and more subdivisions are possible depending on the sensing mechanism.

In this chapter, a summary of the most representative architectures for the construction of optical fibre humidity sensors based on nano-films will be presented. Due to the complexity of this field, this text will try to approach the FOHS to engineers and scientists unfamiliar with the field. Classical architectures will be presented here as well as the state of art related to FOHS. Prior to this, an introduction to the sensitive materials and deposition techniques that are usually employed in the fabrication of these sensors will be described next. Among these techniques, nanotechnology has recently emerged as a new and challenging discipline where the scientific community is being greatly involved. One of the more promising techniques included in the commonly called 'bottom-up nanotechnology' is the LbL technique. As this one can summarize by itself the different fibre optic sensing architectures using sensitive nano-films, this chapter will be mainly focused on the LbL technique. Furthermore, since its rediscovery in the nineties, the layer-by-layer (LBL) self-assembly has found a great acceptance between the optical fibre sensor community.

2 Humidity Sensing Materials and Deposition Techniques

Direct detection of water vapour is possible without the use of any sensing material, for instance with the employ of specialised optics to determine the refractive index of air [4] or the air density changes [5]. However these techniques require the use of expensive systems such as radioactive alpha-ray sources.

Otherwise, relative humidity (RH) can be measured through the properties of some bulk material such as the change in either its physical dimensions or refractive index. Because of that, the vast majority of FOHS developed require the presence of a sensing film whose properties change as a function of the ambient humidity. In addition to this, the material should readily absorb and desorb water.

Both organic and inorganic polymers are used for optical sensors fabrication including silicones, PVC, PTFE, PMMA, nafion, nylon, agarose, sol-gels, etc. [6, 7, 8]. These materials usually give good robustness to the sensor design. The humidity sensing can be achieved by the incorporation of a reagent or by the polymer structure itself. The reagent is a dye which reversibly changes its optical properties with the humidity. Several humidity sensors have been reported based on such well-known reagents. Ideally the wavelengths used for detecting the changes are in the visible or near-infrared region which permits the use of low cost instrumentation and standard fibres. This change is marked in complexes of transition metals. CoCl_2 is one of the most commonly used materials for this purpose because it can absorb water and form coordination complexes which change its characteristic colour [9, 10, 11, 12, 13]. When dry, it has strong absorption in the wavelength range 550–700 nm showing a blue colour, while it does not absorb when it gets wet (forming a hexa-hydrated compound) which results in a pink colour. Other type of humidity sensors are those based on fluorescent and phosphorescent materials such as rhodamine [14, 15] or Al-Ferron [16]. The main advantage of these sensors is that lifetime based measurement can be used, which does not depend on the light-source intensity changes caused by the sensing molecule, the transducer, or the optical path [16]. However, these types of sensors have an important cross sensitivity with temperature.

On the other hand, instead of using a reagent, another different strategy consists of using hygroscopic polymers which, as a general rule, swell when water enters the structure. Hydrophilic gels, such as agarose [17], have a high porosity which determines the quantity of water that the gel is capable of absorbing and therefore its performance as working as humidity sensor. When the water content increases, it induces the gel film to swell which generates a reduction in the refractive index. High refractive index materials increase the interaction between the light guided in the fibre and the sensitive film and hence it increases the sensitivity of the sensor; for instance, PMMA has been used for humidity sensing [6, 18] reporting fast response times (5 s) operating in the range 20–80% RH with complete reversibility.

Among the deposition techniques, it could be said that the classical ones, such as physical vapor deposition or spin coating are intended usually for flat semiconductor substrates and cannot deposit easily uniform films on complex geometries, like

in optical fibres. Actually, to our knowledge, there are three techniques that have been successfully used for the deposition of uniform coatings onto optical fibres such as the dip-coating (DC) technique, the Langmuir-Blodgett (LB) technique and the layer-by-layer (LbL) technique. The first one is usually associated to sol-gel or hydrogel coatings, but it is not useful for controlling the thickness of the coatings on the nanometer scale. In contrast to this, using the former techniques it is possible to deposit coatings of specific nanometer thickness.

2.1 Nanostructured Films

A recently open line of research of suitable humidity materials is focused on nanostructured materials. The use of these materials confers several advantages for humidity sensing like shorter response times and enhanced sensitivity [19, 20, 21, 22, 23, 24, 25]. These fibre optic sensors are based on the deposition of a controlled thickness layer onto the surface of the fibre. The light passing through the fibre is modulated by the thin layer whose refractive index changes as a function of the external relative humidity [26]. As it has been previously commented, the LB and LbL techniques can be used for the fabrication of nanostructured films controlling the thickness of the coatings on the nanometer scale. The LbL process has been successfully proved as a useful tool for the fabrication of nanostructured materials that include many diverse species, such as colorimetric dyes, fluorescent indicators, inorganic semiconductors, conducting polymers, ceramics, metals, quantum dots, enzymes, antibodies or even DNA strands.

The use of nanometric scale films enable the fabrication of sensor heads based on nano-films in which the humidity easily interacts with the film and provoke changes in the entire effective detection zone. In particular, using one of these techniques, speeds of response below 1s have been reported [19].

Layer by layer process was reported for the first time by R. Iler [27] and almost thirty years later the technique was rediscovered by G. Decher and co-workers [28], and extended to the layering of polyelectrolytes and many other systems [29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. In the last seven years the number of works on this topic has increased exponentially and some reviews permit to understand the current state of the art [47, 48, 49, 50, 51, 52]. The versatility of LBL method for the synthesis of materials permits the application of this technique to design or fabrication of different structures on the tip or the cladding of the optical fibre.

The deposition procedure for all cases is based on the construction of molecular multilayers by the electrostatic attraction between oppositely charged polyelectrolytes in each monolayer deposited, and involves several steps. The LbL film deposition method is schematically depicted in Fig. 1.

First, a substrate is cleaned and treated to create a charged surface. Then, the substrate is exposed to a solution of a polyion of opposite charge for a short time (minutes) and, by adsorption, a monolayer of polyions is formed on the surface.

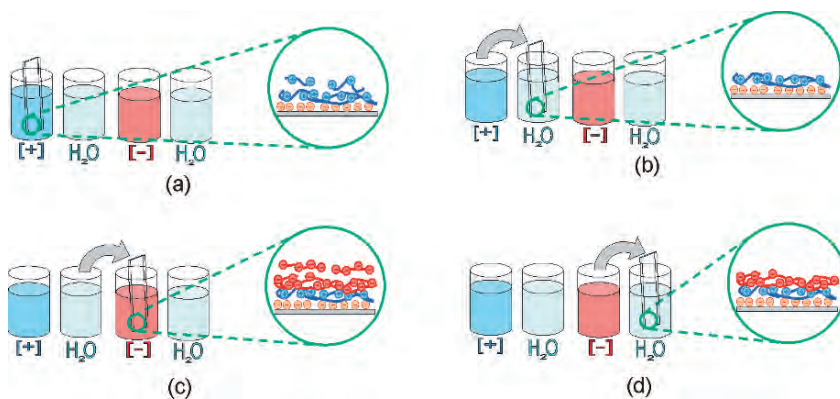


Fig. 1 Schematic representation of the LbL method. (a) Adsorption of a cationic monolayer and inversion of the surface charge. (b) Washing with ultrapure water. (c) Adsorption of anionic monolayer. (d) Washing with ultrapure water

Subsequently, the substrate is alternately dipped into solutions of cationic and anionic polymers (or appropriately charged inorganic clusters) to create a multilayer thin film, a polyanion-polycation multilayer. After each monolayer is formed, the sample is rinsed in pure water to remove the excess of molecules that are not bound and that do not contribute to the monolayer structure.

The molecular species of the cationic and anionic components and the long-range physical order of the layers determine the resulting coating properties. It is important to notice that the polyanions and polycations overlap each other at the molecular level and this produces an optically homogeneous material.

The individual layer composition and thickness can be controlled with different parameters. The most important ones are the temperature, the pH and the concentration of the solutions, the dipping time, the drying time and the substrate where the structure is being deposited. Typical individual layer thickness values range between 0.5 and 15 nm [53]. However, it has been studied with other materials that the range can be widened to 60 nm [54].

An interesting property from the point of view of humidity sensor design is the hydrophobicity and hydrophilicity of the films. If adequately designed, the film can be used for the fabrication of sensors with fast response and low hysteresis.

In addition, other deposition procedures are available; this is the case of the above commented Langmuir-Blodgett technique. This method is based on the deposition of layers with hydrophobic and hydrophilic behavior [55, 56]. Each bilayer to be deposited is spread onto ultrapure water, forming a nanometric surface; when the substrate, in this case the optical fibre, is introduced in the solution, a new layer gets deposited onto the surface [57]. Unfortunately, the LB technique is limited to very specific molecules with combinations of lipophilic and hydrophilic parts.

In any case, most of the sensors based on sensitive nanocoatings, deposited either using sol-gel, LB or LbL techniques, are evanescent wave based sensors. An evanescent wave is a standing harmonic wave located at the core/cladding interface.

This wave penetrates over a small distance (typically below one micron) into the surrounding medium. The evanescent wave and can be described by the exponential decay [58].

$$E = E_o \exp(-z/d_p) \quad (1)$$

where z is the distance normal to the interface and E_o is the wave amplitude at $z = 0$.

The penetration depth is given by

$$d_p = \frac{\lambda}{2\pi n_1} \sqrt{\sin^2 \theta - (n_2/n_1)^2} \quad (2)$$

where λ is the wavelength of the light in the fibre core, n_1 and n_2 are the refractive indices of the core and cladding materials, respectively, and θ is angle of incidence. The thinner is the coating, on the order of the penetration depth, the faster and more efficient its reaction will be.

3 Optical Techniques Used In Fibre Optic Humidity Sensors

Basically, any optical fibre sensor can be classified into a reflexive or transmissive structure. In a reflection-type configuration, the sensing material is located at the end point of the fibre, being especially adequate for applications where the sensor head dimension is important, because it can be easily miniaturized and handled.

In general a higher sensitivity can be obtained with the use of a transmissive configuration. A more important part of the guided power, with respect to the reflective configuration, enters in contact with the sensing film and is modulated and transmitted by the output fibre. There are many possible implementations of evanescent wave sensors. There not exist also other architectures based on wavelength or phase changes, but are no so popular.

3.1 Reflective Sensors Based on Coating Deposition on the Tip of the Fibre

In this scheme, the incident light travels along one optical fibre and illuminates the humidity sensing material. The light reflected or emitted in this case is collected back and transmitted along the same fibre. The sensor head consists of a cleaved or polished end of an optical fibre, onto which the humidity sensitive material is deposited.

In Fig. 2 it is represented the scheme of a fibre tip coated using the LbL method. The thickness of the interferometric cavity increases as the number of adsorbed layers is higher. The refractive index of the film (n_2) can be higher or lower than the fibre (n_1). For example, the nanostructured material [PDDA/PolyR] gives refractive

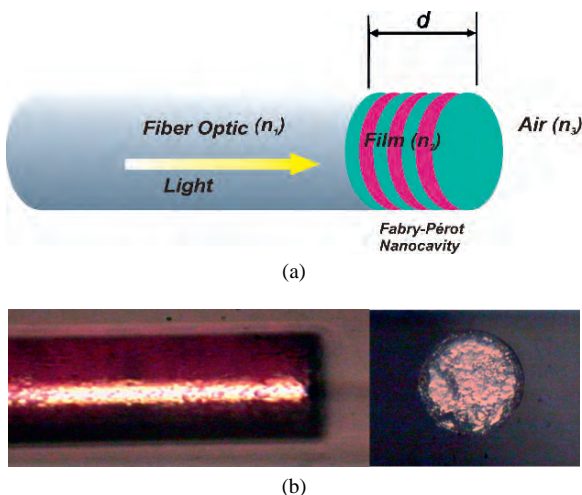


Fig. 2 (a) Coating deposition on the tip of the fibre forming a nanoFabry-Pérot cavity (b) Photograph of a fibre tip humidity sensor

indexes higher than that of the fibre (1.55–1.60), while other porous films based on SiO₂ nanoparticles gives values of n_2 around 1.2–1.25.

The reflected power when n_2 is higher than n_1 is given by [60],

$$R_{FP} = \frac{I_R}{I_0} = \frac{R_1 + R_2 \cdot (1 - A_1)^2 \cdot e^{-\alpha \cdot 4d} + 2 \cdot \sqrt{R_1 \cdot R_2} \cdot (1 - A_1) \cdot e^{-\alpha \cdot 2d} \cdot \cos \phi}{1 + R_1 \cdot R_2 \cdot e^{-\alpha \cdot 4d} + 2 \cdot \sqrt{R_1 \cdot R_2} \cdot e^{-\alpha \cdot 2d} \cdot \cos \phi} \quad (3)$$

where ϕ is the round-trip phase shift of the optical beam in the cavity formed by the coating with thickness d , and R_1 is the reflection coefficient at the first interface (optical fibre–coating) and R_2 is the reflection coefficient at the second interface (coating–air), α is the absorption coefficient of the coating and A_i is a factor associated to the scattering losses.

The following expression permits to estimate the thickness of each nanocavity bilayer [59, 60]:

$$\phi = \frac{4n_2d\pi}{\lambda} \quad (4)$$

where n_2 is the real refractive index of the coating, and λ is the wavelength of the LED. The destructive interference occurs when ϕ is an odd multiple of π , which indicates that it appears for a lower length d .

As can be seen in Fig. 3, the light is coupled from the source and is guided until the sensor head, where the light interacts with the humidity sensing material. The reflected signal from the sensitive film is transmitted to the optical detector. To achieve this, an optical coupler is necessary to drive the response signal to the detection system. Since the cavity created at the end of fibre is much shorter than the coherence length of a LED source, this permits to avoid the necessity of an expensive laser device to monitor interferometric phenomenon caused by changes in the refractive index of the material deposited.

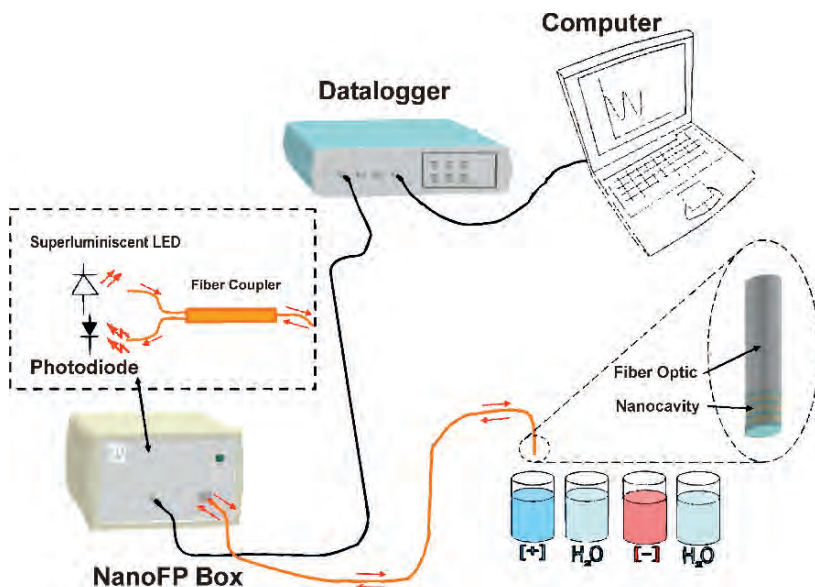


Fig. 3 Experimental set-up used for monitoring the reflected optical power during the LBL process deposition at the end of the fibre. The same scheme is used for humidity monitoring

In addition to the measurement of the target parameter by means of the variation of the output optical power, it is also important to note that with the same scheme it is possible to monitor the power fluctuations during the building of the nanocavity. As most of the materials deposited in the nanocavity are highly lossy, the reflected power changes with a sinusoidal shape during the construction process, which is caused by the light interference in two beams nanoFabry-Pérot (or Fizeau) cavity. Using an optimization algorithm it is possible to estimate the film growing ratio each layer is deposited.

In [60] this scheme was used to measure RH using a solution of poly (diallyl-dimethyl ammonium chloride) (PDDA) as the cation solution and the molecular dye PSS for the anionic solution. The combination of PDDA⁺ and PSS⁻ was deposited onto the sensor head using the LbL technique, creating a nanoFabry-Pérot cavity which attracts water molecules. Consequently, the thickness of the device is a function of the humidity. As the refractive index and the thickness of the nanocavity get modified by the environmental humidity, there is a change in the reflected optical power detected. The response time is so fast that these sensors can be used even for human breathing monitoring [60], as can be seen in Fig. 4. In addition, the sensors showed a repetitive response.

In [61] the same structure was used, but changing the coating by a super hydrophilic nano-film based on SiO₂ nanoparticles, and thus highly sensitive to humidity.

The response time of a fibre optic humidity sensor based on SiO₂ nanoparticles [61] is shown in Fig. 5. In this case, the rise and fall times are only 150 ms and

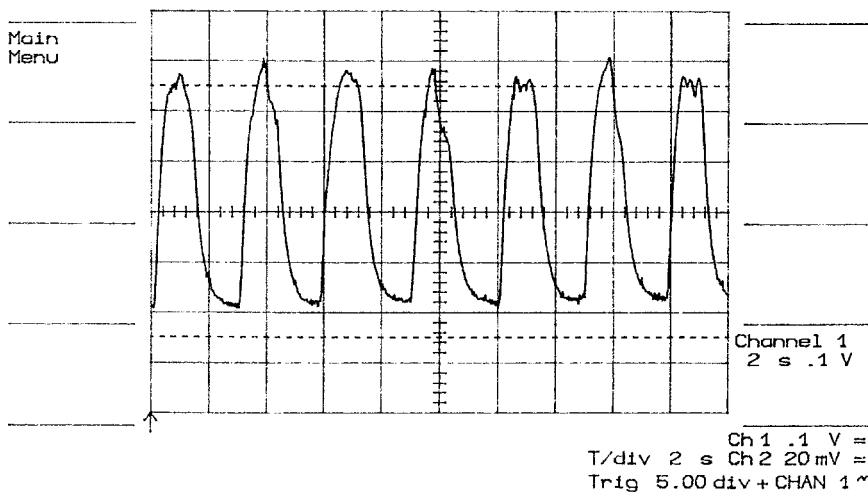


Fig. 4 Response to human breathing of an optical fibre humidity sensor based on a LbL nanoFabry-Perot formed. As indicated in the lower part of the figure, every division along the horizontal axis is 2 s, and every division along the vertical axis is 0.1 V. The materials used were $[\text{Au:PDDA/PSS}]_n$. It has been taken from [60]

100 ms, respectively. The main improvement of the sensor proposed here is that the recovery time after an increase of the relative humidity is much shorter than other sensors based on polymeric films.

In Fig. 6 an AFM image gives an idea of the morphology of the film deposited using SiO_2 nanoparticles. In this image the nanoparticle nature of the film is clearly seen and also the roughness of the sensitive coating.

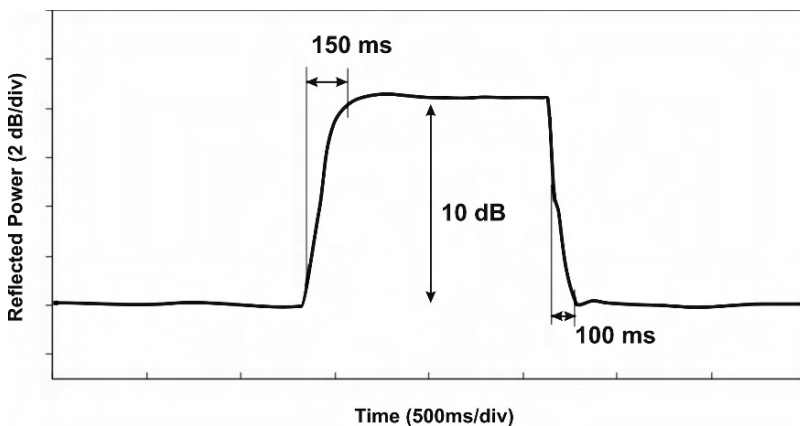


Fig. 5 Dynamic performance of a SiO_2 -based optical humidity sensor taken from [61]

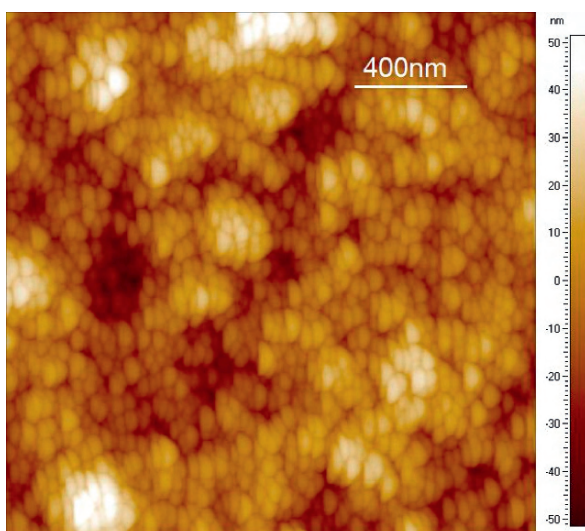


Fig. 6 AFM height image of a film composed by 20 bilayers of 20 nm nanoparticles and 3 bilayers of 7 nm nanoparticles of SiO₂. It has been taken from [61]

3.2 Transmissive Evanescent Wave Sensors

Intrinsic fibre optic humidity sensor can be fabricated using a porous segment of fibre. This can be accomplished using sol gel which when dried has a refractive index similar to that of the fibre. The special segment of fibre can be doped with a humidity sensing material when it is synthesized, for example using CoCl₂ [12]. The main drawback of these sensors is the effect of sol gel aging.

On the other hand, a widely employed method to modulate the transmitted power consists of modulating the evanescent wave. The passive cladding of the optical fibre is replaced along a small section by a humidity sensitive material; so any change in the optical or structural characteristic of the chemical dye due to the presence of the humidity, provokes a change in the effective index of the optical fibre, changing its transmission properties [62, 63]. There are several manners to do this; for instance, it is possible to polish the fibre [64] or to use etching with hydrofluoric acid [65]. An easy to implement alternative consists of using plastic cladding fibres (PCS) [66]. In these fibres, the cladding can be removed easily by mechanical methods allowing highly reproducible sensors [67].

The transmittivity is characterized by the refractive index of the dye, the length and the thickness of the new cladding among other parameters. If the new cladding has a lower refractive index than the core the sensor response is governed by the intensity modulation caused by light absorption of the evanescent wave which is guided through the cladding [68]. Otherwise, for higher refractive index, part of the light is refracted into the cladding and part is reflected back to the core [63].

A bent fibre is also an interesting choice to increase the power coupled in the evanescent field and hence the penetration depth in the humidity sensing area [18, 63]. An important part of the light comes out at the curved portion of the fibre, so a bent probe can lead to higher sensitivities in less exposed regions [15].

In the following sections, three structures that use evanescent waves to modulate the transmitted power and whose performance has been studied for nanostructured material are described: tapered optical fibres, hollow core fibres and long period grating fibres.

3.2.1 Tapered Optical Fibres

In this case, the transducer element is a single mode tapered optical fibre coated with a humidity sensitive material. By tapering of the fibre, it is possible to obtain a more fragile but much more sensitive sensor [17].

In the light guided by an optical fibre it can be distinguished two components, the component inside the core, and the evanescent component propagating through the cladding. However, as the cladding is much thicker than the core, the interaction of the exponentially decaying component with the external medium is insignificant. When a single mode fibre is tapered the core/cladding interface is redefined in such a way that the single mode fibre in the central region of the taper (Zone II in Fig. 4) can be seen as a new waveguide formed by the tapered fibre (which acts as the core of the new structure) and the surrounding medium (which acts as the cladding of this waveguide). In consequence, the evanescent field interacts with the outer medium changing the optical power transmitted.

The shape that the fibre acquires after the tapering process, which depends on the method employed in its fabrication, has a high impact on the light transmission properties. The tapering of the fibres can be achieved by heat pulling or chemical etching [69]. Usually chemically etched tapers are characterized by the removal of part of the cladding, while heat pulling tapers maintain the geometrical ratio between cladding and core. Heat pulling tapers can be fabricated using a flame, a laser or an electrical arc. In Fig. 7 a profile of a typical taper is shown. The transition zones have normally different lengths because of the technique used for tapering. As shown in Fig. 7, a tapered fibre can be divided into three regions: a contracting tapered region L_1 , a central region (waist) L_c , and an expanding tapered region L_2 .

From now on it is supposed a conical distribution taper with step refractive-index profile. Nevertheless, assuming the refractive-index of the optical fibre core as a parabolic distribution, a more realistic theoretical study could be done [70]. Another supposition is that the ratio $S = \rho_{\text{cladding}} / \rho_{\text{core}} = 12.7$ remains constant under tapering and the V-parameter is defined as

$$V_{\text{core}} = \frac{2\pi}{\lambda} \cdot \rho_{\text{core}} \cdot \left(n_{\text{core}}^2 - n_{\text{cladding}}^2 \right)^{\frac{1}{2}}, \quad (5)$$

where λ is the wavelength, ρ_{core} the radius of the core, n_{core} and n_{cladding} are the refractive indices of the core and the cladding respectively. The V-parameter is the

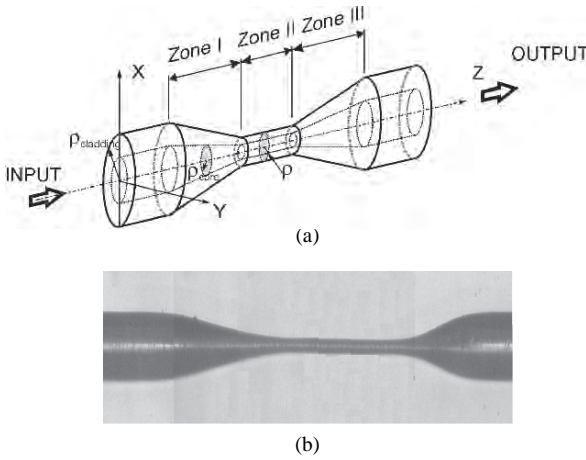


Fig. 7 (a) Optical fibre taper profile. (b) Photograph of a tapered fibre humidity sensor

normalized frequency of the fibre, a dimensionless parameter, and it gives information about the number of modes propagating in a fibre [71]. When $V = 2.405$ only the lowest order mode is propagating.

Stretching the fibre diameter leads to decrease the value of V , consequently the intensity distribution of the fundamental core mode LP_{01} changes from a narrow profile to a broader one. If V continues decreasing it can reach its cut-off value (V_{cc}) and ρ_{cc} will be the radius of the core when $V = V_{cc}$. This V_{cc} is the minimum value of V to have light guided through the core, for smaller values of V the light cannot be confined in the core. The numerical value of V_{cc} has been determinate as [72]

$$V_{cc} = \sqrt{2/\ln(S)}. \quad (6)$$

From this equation and for typical single-mode fibres with $S \geq 10$ it is predicted $V_{cc} \approx 0.93$. Thus, therefore a good approximation for the following experimental results is to suppose a value of unity for V_{cc} [73].

In other words, for $V = V_{cc}$ the modal field expands and reaches completely the cladding producing a set of cladding modes. Thus, the effect of tapering (for $\rho = \rho_{cc}$) is to create a region where cladding modes exist with a V value much greater than 2.405 [72]. For this reason in a taper we can find a singlemode-multimode transition and a multimode-singlemode transition. This gives us the classical modal domain interferometric structure (singlemode-multimode-singlemode). The coupling mechanism is determined between modes of same angular eigennumbers (LP_{01} to LP_{0x}), mainly between LP_{01} and LP_{02} . It should be remarked that the total optical power (core + cladding) in the taper remains constant [74]. Once the light passes through the waist, the remaining light in the LP_{01} cladding mode is transferred to the LP_{01} core mode which propagates through the conventional optical fibre. The remaining light in other modes gets lost. Theoretical aspects of this phenomenon have been exhaustively studied in [72, 73, 74, 75, 76].

The cladding modes that are the cause of the power loss can be modulated once the tapering process has finished by the deposition of an overlay surrounding material. Small changes in the index of refraction or the thickness of this overlay greatly influence the transmission properties in the multimode central region. The sensitive overlay is used both for creating a humidity sensing surface and for adjusting the working point of the sensor at the point of optimal sensitivity.

As can be seen in the experimental set-up of Fig. 8, the light is coupled from the laser source and is guided until the sensor head, where the light interacts with the humidity sensing material. The signal is transmitted to the optical detector.

In [19, 20, 21, 22] LBL was used to fabricate an optical fibre humidity sensor based on [PDDA/ Poly R-478] nanostructured overlay.

In this work it was shown that the thickness can be controlled in order to optimize the sensor sensitivity, by stopping the deposition process at the maximum slope of the transmitted optical power, achieving 26.8 times better sensitivity with half the thickness.

In Fig. 9(d) the transmittivity of a 20 μ m waist diameter taper, coated with the humidity sensitive polymer [PDDA+/Poly-R-] is shown. When the working point is located at a zone of high derivative in the transmission optical power curve, the final sensor was much more sensitive. In Fig. 9(a-c) the experimental response to relative humidity of the three sensors corresponding to the three working points is shown. It is deduced that a higher slope of the transmission characteristic results in sensors with a higher sensitivity. By using the experimental output optical transmission characteristics curve it is possible to find out the maximum slope zone of the

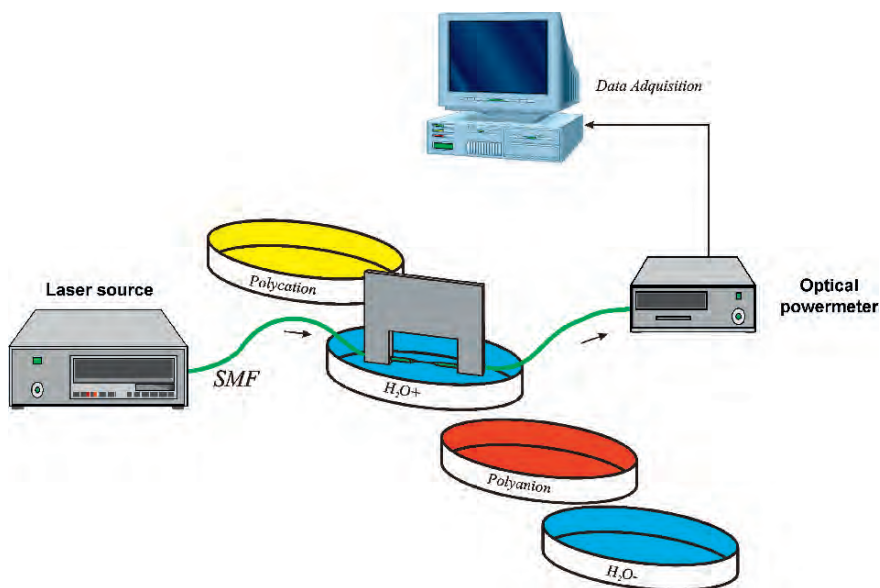


Fig. 8 Transmission set up with a tapered optical fibre based evanescent wave sensor

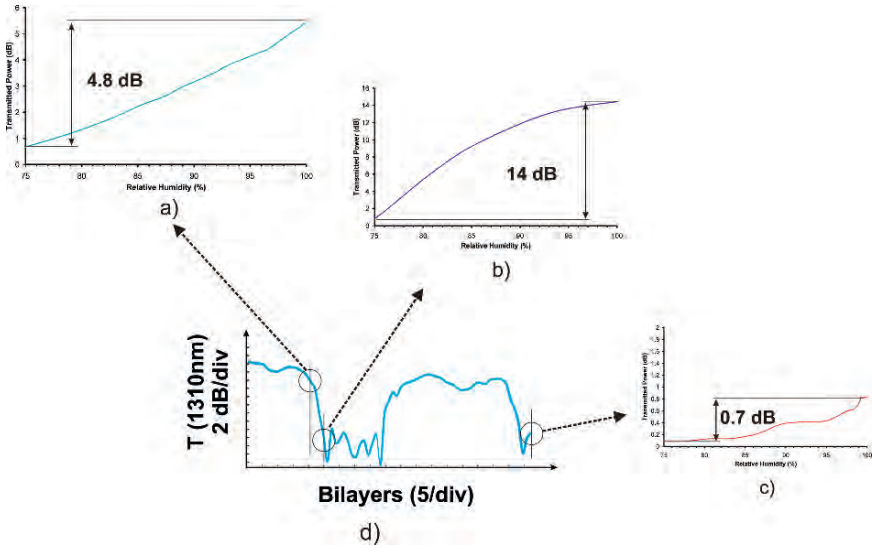


Fig. 9 Experimental response of 20 μm waist diameter TOF based humidity sensors to RH corresponding to three working points of coating thicknesses: (a) 23 bilayers, (b) 26 bilayers and (c) 62 bilayers. (d) Transmittivity as a function of the sensitive deposited nano-film thickness. Taper length: 2.2 mm. Nano-film: [PDDA+/Poly-R-]. $\lambda = 1310\text{nm}$ [19]

potential sensor in order to fabricate an optimized sensitivity sensor with a coating thickness tuned around this working point.

3.2.2 Hollow Core Fibres

Another possibility for increasing the amount of optical power coupled in the evanescent field is by using new types of fibres that have appeared a few years ago. Among them, one can mention hollow core fibres. The simpler type of hollow core fibres, consist basically of tubular fibres where the optical signal is guided mainly by the cladding, and hence, it is easy to reach the evanescent fields. If a short portion of a hollow core fibre is spliced to a multimode fibre (MMF) using the appropriate electric arc conditions, the HCF collapses, forming a tapered solid fibre in the interface between both fibres (see Fig. 10). In these devices, the light that is guided in the core of a lead-in MMF can be coupled to the cladding of the HCF due to the tapered region instead of being confined in the air core. When the light reaches the lead-out MMF, it is coupled into the silica core again. Because the light is guided by the silica cladding in the HCF region, these devices, called MHM, can be used as evanescent field sensors that are sensitive to any coating deposited onto this region and have been also used to build optical fibre humidity sensors in [22, 23, 24].

In order to understand the influence of some important parameters on the behavior of the MHM, in this and case in contrast to the previous architecture, a ray model

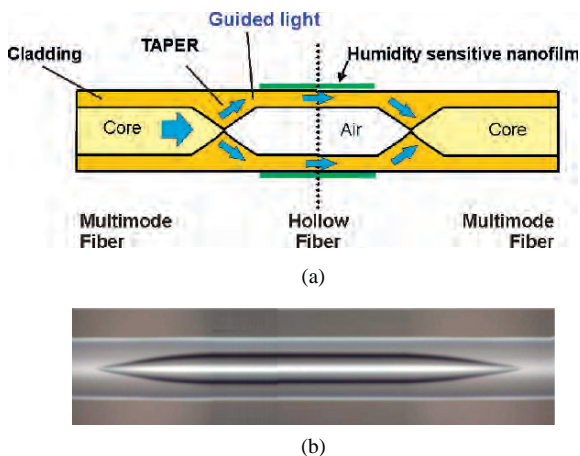


Fig. 10 (a) Hollow core fibre with microstructured cladding, connected between two standard multimode optical fibre sections. A sensitive material can be also fixed onto the cladding. (b) Photograph of a MHM sensor

for a one-dimensional waveguide is developed (see Fig. 11). This model calculates the transmitted power in the HCF after the multiple reflections of light at the interface between the HCF and the exterior. The light source is assumed to be a lambertian one oriented in the direction of propagation of the waveguide. The transmitted power when no deposition has been performed on the HCF will be taken as a reference value (0 dB). This one-dimensional simple model is enough to qualitatively

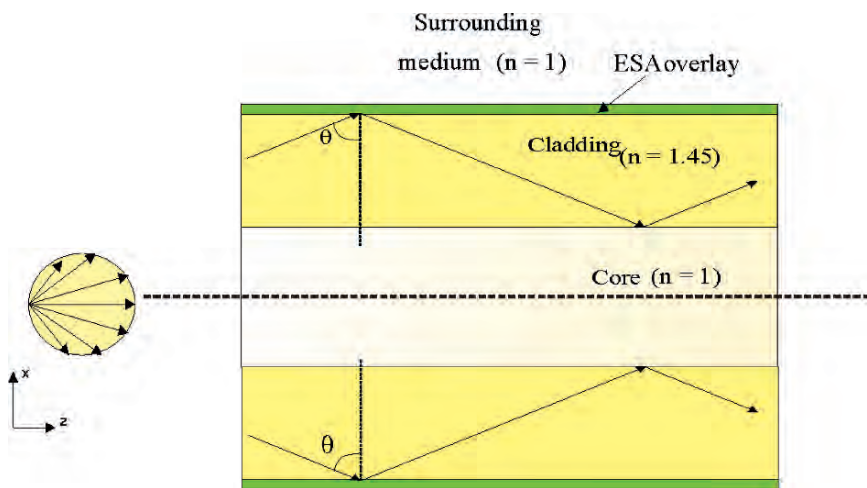


Fig. 11 Theoretical model: a lambertian light source impinges on a one-dimensional slab waveguide with an LBL overlay on one side

demonstrate the behavior of the structure. If more accuracy is required, a more complex two-dimensional model is necessary.

Working with evanescent field sensors one important characteristic of the material deposited is the refractive index. In this work it has been only deposited [PDDA/Poly R], which refractive index is higher than that of the fibre. In order to know what response the devices will have with different refractive index coatings some simulations were done. The simulations were done with three different refractive indexes, the one obtained in the experiments, another lower and the other one higher; exactly $1.54 + 0.004i$, $1.37 + 0.074i$ and $1.67 + 0.004i$, respectively. The length of the HCF section used for the simulations was 20 mm and the wavelength of the light source was 1310 nm. The results are shown in Fig. 12.

The response of the device is very different when the refractive index changes. The transmitted optical power of the devices when $n = 1.65$ and $n = 1.54$ falls oscillatory as the thickness of the material deposited is increased unlike the power of the device when $n = 1.37$ which has no oscillations. As can be seen, the depth and period of the theoretical curve are bigger when $n = 1.54$ than when $n = 1.65$. On the other hand, although the fall of the theoretical curve for $n = 1.37$ is not the biggest one, since its transmission slope in the first few nanometers of thickness is very high it could be used to develop very sensitive sensors, depositing materials with the appropriate refractive index and with a thickness of few decades of nanometers.

Some experimental results are presented in Fig. 13; the HCF segment used in the MHM has a length of 20 mm, and diameters of 50/150 μm .

The power transmitted by MHM structures follows an oscillatory way as the number of nano-bilayers deposited gets increased. As it can be seen, the period of this oscillation depends, among other parameters, on the wavelength of the light source as expected due to the modal interferometer behaviour it exhibits. Furthermore, it is observed a gradual decrease of transmission due to the non-zero

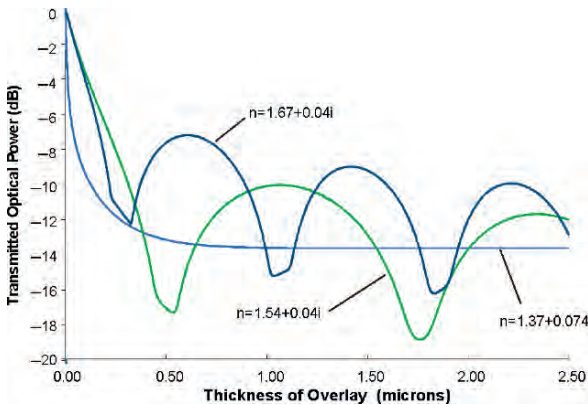


Fig. 12 Theoretical optical power transmitted by three MHM devices 20 mm long when the refractive index of the material deposited are: $1.54 + 0.004i$, $1.37 + 0.074i$ and $1.67 + 0.004i$ at 1310 nm. Waveguide thickness: 50 microns [23]

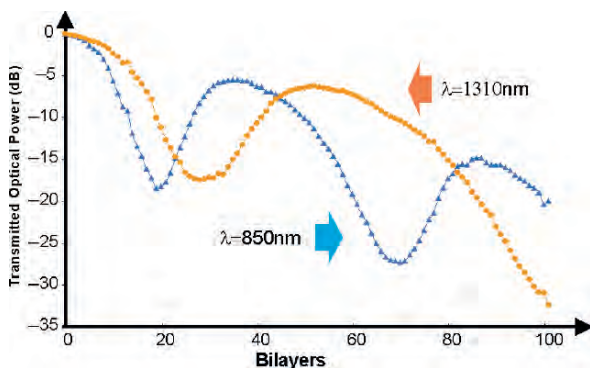


Fig. 13 Experimental data of the optical power transmitted by one MHM device 20 mm long with HCF 50/150 μm inner and outer diameters at two different wavelength: 850 nm and 1310 nm [23]

imaginary part of the refractive index of the nano-film deposited. The amplitude of the oscillatory behaviour of the transmitted power trends to reduce when the inner diameter decreases because the evanescent field ratio respect to the total transmitted optical power is higher. There exists others parameters that also affect to the optical power transmission of this device as the length of the HCF section, the splice machine parameters used to tapered the multimode fibres (that is, the slope of the taper) the index of refraction of the overlay, etc.

Also, to evaluate the response of these proposed sensors, we exposed the sensor head to rapid changes of the RH. Human breathing contains more water vapour than the normal room environment. Accordingly, the sensor was set 3 cm from a subject’s mouth. The results obtained are shown in Fig. 14. The observed rise response times were around 300 ms. The fall times were less than a couple of seconds.

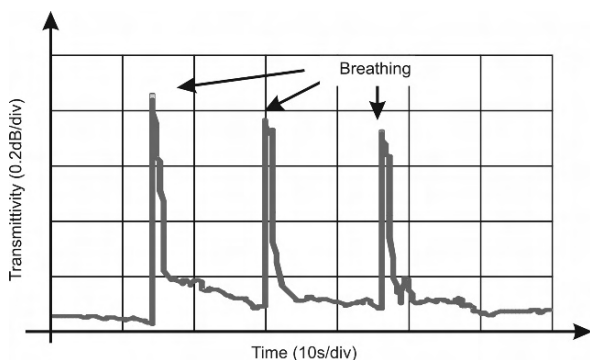


Fig. 14 Experimental response to human breathing using a 20 mm-long HCF. Taken from [77]

3.2.3 Long Period Fibre Bragg Grating

Long Period Fibre Gratings (LPFGs) are based on a periodic index modulation of the refractive index of the core of a single mode fibre (SMF), with a period between 100 microns and 1 mm. LPFGs induce attenuation bands in the transmission spectrum based on the coupling between the core mode and the copropagating cladding modes. Because of that, the influence of the surrounding medium on the LPFGs transmission is more important than in fibre Bragg gratings (FBGs), where there is a contrapropagative coupling only between core modes. It has been experimentally proved that the deposition of thin layers onto the surface of the fibre using Langmuir Blodgett (LB) technique, LBL or dip-coating, can induce important changes in the resonance wavelengths [78, 79, 80, 81, 82, 83, 84, 85]. In this way, long-period fibre gratings (LPFGs) can be used for the construction of evanescent wave sensor; the changes of the deposited overlay due to the humidity induce changes in position of the attenuation bands that are collected in the OSA.

In Fig. 15 it is represented the typical setup used for monitorization of the transmission spectra in a long-period fibre grating (LPFG). The flat spectrum from a broadband source is launched into the optical fibre and at the output, the transmission spectrum with the attenuation bands generated due to the coupling of light from the core mode to the cladding modes, is collected in an optical spectrum analyzer (OSA).

There are two key points that define each attenuation band. The first one is its depth. This can be approximated with this expression [86]:

$$T_i = \cos^2(k_i L) \quad (7)$$

where i is the cladding mode order, k_i is the coupling coefficient and L the length of the grating.

The second key point is the resonance wavelength. Though the simple phase matching condition is sufficient to give an approximation of the central wavelength of the attenuation band, the modified phase matching condition is often preferred [87]:

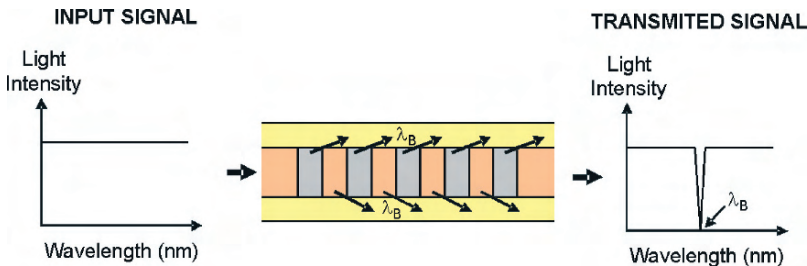


Fig. 15 Schematic working principles of LPFG

$$\beta_{01}(\lambda) + s_0 \zeta_{01,01}(\lambda) - (\beta_{0j}(\lambda) + s_0 \zeta_{0j,0j}(\lambda)) = \frac{2\pi N}{\Lambda} \quad (8)$$

where β_{01} and β_{0j} are the propagation constants of the core and the j cladding modes respectively, $\zeta_{01,01}$ and $\zeta_{0j,0j}$ are the self-coupling coefficients of the core and the j cladding modes, s_0 is the coefficient of the zero-frequency Fourier component of the grating, Λ is the period of the grating, and N is the diffraction order.

The development of numerical methods has helped to understand the phenomena involved in these experiments. The wavelength shift of the attenuation bands was explained in [88] with a scalar analysis of modes (LP mode approximation) and the application of coupled mode theory [87]. If an overlay of higher refractive index than the cladding is deposited on this LPFG, as the overlay thickness increases, cladding modes shift their effective index to higher values. When the overlay is thick enough, one of the cladding modes is guided by the overlay. This causes a reorganization of the effective index of the rest of modes.

In Fig. 16 the effective index of the first ten cladding modes are represented as a function of the coating thickness. The notation used is $LP_{0,2}$ for the first cladding mode, $LP_{0,3}$ for the second cladding mode and so on. Cladding modes with lower effective index than the one that is guided by the overlay will shift their effective index value towards the effective index of the immediate higher effective index mode.

As more material is deposited, the effective index distribution before deposition is recovered. The effective index of the eighth cladding mode will be now that of the seventh one, the effective index of the seventh cladding mode will be that of the sixth mode, and so on. During the first redistribution there is a sudden effective index increase for the $LP_{0,2}$ mode. The same is true for the $LP_{0,3}$ mode during the second redistribution, for the $LP_{0,4}$ in the third one, and for the $LP_{0,5}$ in the fourth one. A more detailed explanation of these statements can be found in [88].

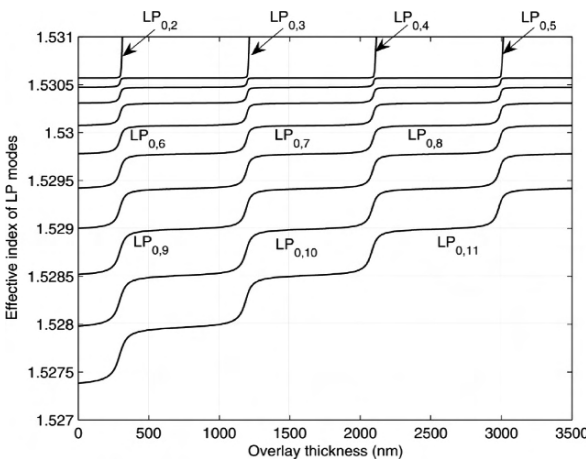


Fig. 16 Effective index of the first ten cladding modes as a function of the overlay thickness of the nano-film deposited

The same phenomenon observed for the effective index of the modes is true for the resonance wavelength values because there is a close relation between the resonance wavelength of each attenuation band and the effective index of its corresponding cladding mode.

Several studies have been developed in the last years related to the experimental behaviour of the LPFG to thin layers [25, 78, 79, 80, 81, 82, 83, 84, 85, 89]. When an intermediate external medium (a coating) is added between the surrounding medium and the cladding of the LPFG two additional parameters are added: the overlay refractive index and the overlay thickness which can be controlled using the appropriate deposition technique. By adequate parameterization of the overlay thickness and overlay refractive index the device can be optimized for a maximum wavelength shift, thus a maximum sensitivity, as a function of specific parameters [90, 91]. An usually employed method for improving the sensitivity is based on the deposition of a thin overlay on the cladding of an LPFG. In this way, it is possible to design a device with optimal sensitivity. The phenomenon has also been experimentally analyzed with an improvement in the sensitivity of the device by more than ten times [85].

In [92] a humidity sensor using a long-period fibre grating (LPG) and a SiO_2 -nanospheres coating was developed. Figure 17 shows a humidity cycle. The dashed line represents the relative humidity variation, and the straight line represents the peak of absorption of the LPG. At different humidity values, the polymeric overlay changes its optical properties yielding to a shift in the resonance wavelength of the LPG. Wavelength shifts up to 12 nm in a range from 20 to 80% of humidity level were obtained. When the relative humidity increases, the resonance wavelength peak shifts to lower wavelengths. This effect is due to changes accomplished in the refractive index of the SiO_2 -nanospheres film.

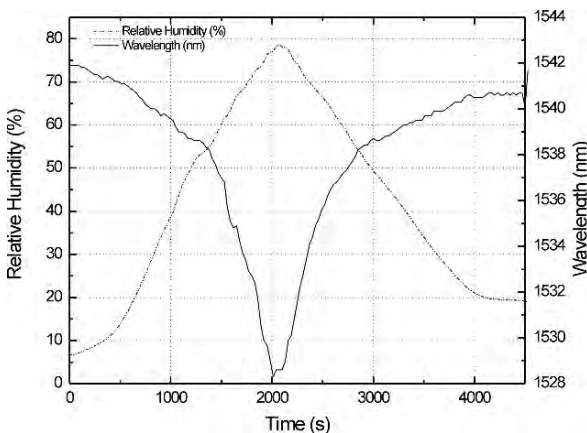


Fig. 17 Resonance wavelength shift for different relative humidity levels. It has been taken from [94]

4 Conclusions

The field of optical fibre humidity sensors has been reviewed, with special attention to those coated with nanostructured films. The basics of the LBL method applied to the fabrication of humidity sensitive coatings have been described. The main sensing mechanisms and architectures have been studied paying attention to the different methods of fabrication; both reflection and transmission configurations have been taken into account. The first one is based on the deposition on the tip of optical fibre creating a nanocavity, while the second group is based mainly on evanescent field based modulation.

Acknowledgments This work was funded in part by the Spanish Ministry of Education and Science – FEDER TEC2006-12170/MIC Research Grant and Government of Navarra-Feder Euroinnova Research Grants.

References

1. Fuhr, P.L.; Huston, D.R., Corrosion detection in reinforced concrete roadways and bridges via embedded fibre optic sensors. In: *Smart Mater. Struct.*, 1998, 7, 217–228.
2. Cooper, K.R.; Elster, J.; Jones, M.; Kelly, R.G., Optical fibre-based corrosion sensor systems for health monitoring of aging aircraft, In: *Autotestcon Proceedings, IEEE Systems Readiness Technology Conference*, Aug. 2001, 847–856, 20–23.
3. Bownass, D.C.; Barton, J.S.; Jones, J.D.C., Detection of high humidity by optical fibre sensing at telecommunications wavelengths, *Optics Communications*, 15 January 1998, 146 (1), 90–94(5).
4. Mc Murtry, S.; Wright, J.D.; Jackson, D.A., Sensing applications of a low-coherence fibre-optic interferometer measuring the refractive index of air, *Sensors and Actuators B: Chemical*, 5 January 2001, 72 (1), 69–74.
5. Matsumoto, S., New air density and absolute humidity sensors using optical fibre cable and alpha-rays, *Meas. Sci. Technol.* 2001, 12, 865–870.
6. McMurtry, S.; Wright, J.D.; Jackson, D. A., A multiplexed low coherence interferometric system for humidity sensing, *Sensors and Actuators B*, 2000, 67, 52–56.
7. Brook, T.E.; Narayanaswamy, R., Polymeric films in optical gas sensors, *Sensors and Actuators, B: Chemical*, 1998, 51(1–3), 77–83.
8. Raimundo, I.M., Jr. Narayanaswamy, R., Evaluation of Nafion-Crystal Violet films for the construction of an optical relative humidity sensor, *Analyst*, 1999, 124(11), 1623–1627.
9. Corera, F. P.; Gaston A.; Sevilla, J.; Relative humidity sensor based on side-polished fibre optic, *Proceedings of the 17th IEEE Instrumentation and Measurement Technology Conference*, 2000. IMTC 2000.
10. Ballantine, D.S.; Wohltjen, H., Optical waveguide humidity detector, *Analytical Chemistry*, 1986, 58(13), 2883–2885.
11. Russel, A.P.; Fletcher, K.S., Optical sensor for the determination of moisture, *Anal. Chim. Acta* 1985, 170, 209–216.
12. Zhou, Q.; Shahriari, M.R.; Kritz, D.; Sigel G.H. Jr., Porous fibre optic sensor for high sensitivity humidity measurements, *Anal. Chem.* 1988, 60, 2317–2320.
13. Boltinghouse, F.; Abel, K. Development of an optical relative humidity sensor. Cobalt chloride optical absorbency sensor study, *Analytical Chemistry*, 1989, 61(17), 1863–1866.

14. Choi, M.M.F.; Ling, T.O., Humidity-sensitive optode membrane based on a fluorescent dye immobilized in gelatin film, *Analytica Chimica Acta*, 1999, 378(1–3), 127–134.
15. Otsuki, S.; Adachi, K.; Taguchi T., A novel fibre-optic gas-sensing configuration using extremely curved optical fibres and an attempt for optical humidity detection, *Sensors and Actuators, B: Chemical*, 1998, 53(1–2), 91–96.
16. Campo, J.C.; Perez, M.A.; Gonzalez, M.; Ferrero, F.J., Measurement of air moisture by the phosphorescence lifetime of a sol-gel based sensor, *Proceedings of the 17th IEEE Instrumentation and Measurement Technology Conference*, 2000, IMTC 2000, 1, 273–276.
17. Bariain, C.; Matias, I.R.; Arregui, F.J.; Lopez-Amo, M., Optical fibre humidity sensor based on a tapered fibre coated with agarose gel, *Sensors and Actuators B*, 2000, 69, 127–131.
18. Gupta, B.D.; Ratnanjali, A novel probe for a fibre optic humidity sensor, *Sensors and Actuators B: Chemical*, 20 November 2001, 80 (2), 132–135(4).
19. Corres, J.M.; Arregui, F.J.; Matias, I.R., Sensitivity optimization of tapered optical fibre humidity sensors by means of tuning the thickness of nanostructured sensitive coatings, *Sensors and Actuators, B: Chemical*, 2007, 122(2), 442–449.
20. Corres, J.M.; Bravo, J.; Matias, I.R.; Arregui, F.J., Nonadiabatic tapered single-mode fibre coated with humidity sensitive nano-films, *IEEE Photonics Technology Letters*, 2006, 18(8), 935–937.
21. Corres, J.M.; Arregui, F.J.; Matias, I.R., Design of humidity sensors based on tapered optical fibres, *Journal of Lightwave Technology*, 2006, 24(11), 4329–4336.
22. Matias, I.R.; Arregui, F.J.; Corres, J.M.; Bravo, J., Evanescent field fibre-optic sensors for humidity monitoring based on nanocoatings, *IEEE Sensors Journal*, 7(1), 89–95.
23. Bravo, J.; Matias, I.R.; Del Villar, I.; Corres, J.M.; Arregui, F.J., Nano-films on hollow core fibre-based structures: An optical study, *Journal of Lightwave Technology*, 2006, 24(5), 2100–2107.
24. Matias, I.R.; Bravo, J.; Arregui, F.J.; Corres, J.M., Nano-films on a hollow core fibre, *Optical Engineering*, 2006, 45(5), Art. No. 050503.
25. Del Villar, I.; Corres, J.M.; Achaerandio, M.; Arregui, F.J.; Matias, I.R., Spectral evolution with incremental nanocoating of long period fibre gratings, *Optics Express*, 2006, 14(25), 11972–11981.
26. Jindal, R.; Tao, S.; Singh, J.P.; Gaikwad, P.S., High dynamic range fibre optic relative humidity sensor, *Opt. Eng.*, May 2002, 41(5), 1093–1096.
27. Iler, R.J.J. Multilayers of colloidal particles, *Journal of Colloid and Interface Science*, June 1966, 21 (6), 569–594.
28. Decher, G.; *Fuzzy Nanoassemblies: Toward Layered Polymeric Multicomposites*, *Science*, 1997, 277: 1232–1237.
29. Lvov, Y., Ariga, K., Ichinose, I., Kunitake, T., Assembly of multicomponent protein films by means of electrostatic layer-by-layer adsorption, *J. Am. Chem. Soc.*, 1995, 117(22), 6117–6123.
30. Ariga, K.; Lvov, Y.; Kunitake, T., Assembling alternate dye-polyion molecular films by electrostatic layer-by-layer adsorption, *J. Am. Chem. Soc.*, 1997, 119(9), 2224–2231.
31. Liu, Y.; Wang, A.; Claus, R., Molecular self-assembly of TiO₂/polymer nanocomposite films, *J. Phys. Chem. B.*, (Article), 1997, 101(8), 1385–1388.
32. Liu, Y.J.; Wang, A.B.; Claus, R.O., Layer-by-layer electrostatic self-assembly of nanoscale Fe₃O₄ particles and polyimide precursor on silicon and silica surfaces. *Appl. Phys. Lett.*, 1997, 71, 2265–2267.
33. Liu, Y.J.; Wang, A.B.; Claus, R.O., Layer-by-layer ionic self-assembly of Au colloids into multilayer thin-films with bulk metal conductivity, *Chem. Phys. Lett.*, 1998, 298, 315–319.
34. Lenahan, K. M.; Wang, A. B.; Liu, Y. J.; Claus, R. O., Novel polymer dyes for nonlinear optical applications using ionic self-assembled monolayer technology. *Adv. Mater.* 1998, 10, 853–855.
35. Donath, E.; Sukhorukov, G. B.; Caruso, F.; Davis, S.A.; Mohwald, H., Novel hollow polymer shells by colloid-templated assembly of polyelectrolytes. *Angewandte Chemie-International Edition*, 1998, 37, 2202–2205.

36. Caruso, F.; Caruso, R. A.; Mohwald, H., Nanoengineering of inorganic and hybrid hollow spheres by colloidal templating, *Science*, 1998, 282, 1111–1114.
37. Caruso, F.; Lichtenfeld, H.; Donath, E.; Mohwald, H., Investigation of electrostatic interactions in polyelectrolyte multilayer films: binding of anionic fluorescent probes to layers assembled onto colloids, *Macromolecules*, 1999, 32, 2317–2328.
38. Chen, L. H.; McBranch, D. W.; Wang, H. L.; Helgeson, R.; Wudl, F.; Whitten, D. G., Highly sensitive biological and chemical sensors based on reversible fluorescence quenching in a conjugated polymer, *Proceedings of the National Academy of Sciences of the United States of America*, 1999, 96, 12287.
39. Dubas, S. T.; Schlenoff, J. B., Factors controlling the growth of polyelectrolyte multilayers, *Macromolecules*, 1999, 32, 8153–8160.
40. Mamedov, A. A.; Kotov, N. A.; Prato, M.; Guldi, D. M.; Wicksted, J. P.; Hirsch, A., Molecular design of strong single-wall carbon nanotube/polyelectrolyte multilayer composites, *Nature Materials* 2002, 1, 190.
41. Nakagawa, M.; Oh, S. K.; Ichimura, K., Photopatterning and visualization of adsorbed monolayers of bis(1-benzyl-4-pyridinio)ethylene moieties, *Adv. Mater.*, 2000, 12, 403–407.
42. Ho, P.K.H.; Kim, J.S.; Burroughes, J.H.; Becker, H.; Li, S.F.Y.; Brown, T.M.; Cacialli, F.; Friend, R.H., Molecular-scale interface engineering for polymer light-emitting diodes, *Nature*, 2000, 404 (6777), 481–484.
43. Shiratori, S. S.; Rubner, M. F., pH-dependent thickness behavior of sequentially adsorbed layers of weak polyelectrolytes, *Macromolecules*, 2000, 33, 4213–4219.
44. Bertrand, P.; Jonas, A.; Laschewsky, A.; Legras R., Ultrathin polymer coatings by complexation of polyelectrolytes at interfaces: Suitable materials, structure and properties, *Macromolecular Rapid Communications*, 2000, 21, 319–348.
45. Mendelsohn, J.D.; Barret, C.J.; Chan, V.V.; Pal, A.J.; Mayes, A.M.; Rubner, M.F., Fabrication of microporous thin films from polyelectrolyte multilayers, *Langmuir*, 2000, 16, 5017–5023.
46. Mattoussi, H.; Mauro, J.M.; Goldman, E.R.; Anderson, G.P.; Sundar, V.C.; Mikulec, F.V.; Bawendi, M.G., Self-assembly of CdSe-ZnS quantum dot bioconjugates using an engineered recombinant protein, *J. Am. Chem. Soc.* 2000, 122, 12142–12150.
47. Schreiber, F., Structure and growth of self-assembling monolayers, *Progress in Surface Science*, 2000, 62, 151–256.
48. Caruso, F., Nanoengineering of particle surfaces, *Adv. Mater.* 2001, 12, 11–22.
49. Adams, D.M.; Brus, L.; Chidsey C.E.D.; Creager, S.; Creutz, C.; Kagan, C.R.; Kamat, P.V.; Lieberman, M.; Lindsay, S.; Marcus, R.A.; Metzger, R.M.; Michel-Beyerle, M.E.; Miller, J.R.; Newton, M.D.; Rolison, D.R.; Sankey, O.; Schanze, K.S.; Yardley, J.; Zhu, X.Y., Charge transfer on the nanoscale: Current status, *J. Phys. Chem. B*, 2003, 107, 6668–6697.
50. Schönhoff, M., Self-assembled polyelectrolyte multilayers, *Current Opinion in Coll. Interf. Sci.* 2003, 8, 86.
51. Thünnemann, A.F.; Müller, M.; Dautzenberg, H.; Joanny, J.-F.; Löwen, H., Polyelectrolyte Complexes, *Advances in Polymer Science* 2004, 166, 113–171.
52. Hammond, P.T., Form and function in multilayer assembly: New applications at the nanoscale, *Adv. Mater.* 2004, 16, 1271–1293.
53. Choi, J.; Rubner, M.F., Influence of the degree of ionization on weak polyelectrolyte multilayer assembly, *Macromolecules*, 2005, 38, 116–124.
54. Del Villar, I.; Matias, I.R.; Arregui, F.J., LBL-based in-fibre nanocavity for hydrogen-peroxide detection, *IEEE Trans. on Nanotech.* 2005, 4, 187–193.
55. Hu, W.; Liu, Y.; Xu, Y.; Liu, S.; Zhou, S.; Zeng, P.; Zhu, D.B., The gas sensitivity of Langmuir-Blodgett films of a new asymmetrically substituted phthalocyanine, *Sensor. Actuat. B-Chem.*, 1999, 56, 228–233.
56. Bariain, C.; Matias, I.R.; Fernandez-Valdivielso, C.; Arregui, F.J.; Rodríguez-Méndez, M.L.; DLbLja, J.A., Optical fibre sensor based on lutetium bisphthalocyanine for the detection of gases using standard telecommunication wavelengths, *Sensor. Actuat. B-Chem.*, 2003, 93, 153–158.

57. Gutierrez, N.; Rodríguez-Méndez, M.L.; De Saja, J.A., Array of sensors based on lanthanide bisphthalocyanine Langmuir-Blodgett films for the detection of olive oil aroma. *Sensor. Actuat. B-Chem.* 2001, 77, 437–442.
58. Dakin, J.; Culshaw, B., *Optical fibre sensors. Principles and components*, Norwood, MA: Artech House. 1988, pp. 63–64.
59. Arregui, F.J.; Matias, I.R.; Liu, Y.J.; Lenahan, K.M.; Claus, R.O., Optical fibre nanometer-scale Fabry-Perot interferometer formed by the ionic self-assembly monolayer process, *Opt Lett.*, 1999, 24, 596–598.
60. Arregui, F.J.; Liu, Y.; Matias, I.R.; Claus, R.O.; Optical fibre humidity sensor using a nano Fabry-Perot cavity formed by the ionic self-assembly method, *Sensors and Actuators B* 59 1999.54–59
61. Corres, J. M.; Matias, I. R.; Hernaez, M.; Bravo, J.; Arregui, F. J., Optical fibre humidity sensors using nanostructured coatings of SiO₂ nanoparticles, *IEEE Sensors Journal*, Vol. 8, Issue 3, March 2008, pp. 281–285.
62. Khalil, S.; Bansal, L.; El-Sherif, M., Intrinsic fibre optic chemical sensor for the detection of dimethyl methylphosphonate. *Opt. Eng.*, 2004, 43, 2683–2688.
63. Otsuki, S.; Adachi, K.; Taguchi, T.; A novel fibre-optic gas-sensing configuration using extremely curved optical fibres and an attempt for optical humidity detection, *Sensors and Actuators B*, 1998, 53, 91–96.
64. Senosiain, J.; Díaz, I.; Gastón, A.; Sevilla, J., High sensitivity temperature sensor based on side-polished optical fibre. *IEEE Trans. Instrum. Meas.* 2001, 50, 1656–1660.
65. Sumdia, S.; Okazaki, S.; Asakura, S.; Nakagawa, H.; Murayama, H.; Hasegawa, T., Distributed hydrogen determination with fibre-optic sensor. *Sensor. Actuat. B-Chem.* 2005, 108, 508–514.
66. Cherif, K.; Mrazek, J.; Hleli, S.; Matejec, V.; Abdelghani, A.; Chomat, M.; Jaffrezic-Renault, N.; Kasik, I., Detection of aromatic hydrocarbons in air and water by using xerogel layers coated on PCS fibres excited by an inclined collimated beam. *Sensor. Actuat. B-Chem.*, 2003, 95, 97–106.
67. Suzuki, O.; Miura, M.; Morisawa, M.; Muto, S., POF-type optic humidity sensor and its application (as breathing-condition monitor), in: *Proceedings of the 15th Optical Fibre Sensors Conference (OFS 2002)*, Technical Digest, Orlando, OR, 2002, pp. 447–450.
68. Yuan, J.; El-Sherif, A. Fibre-optic chemical sensor using polyaniline as modified cladding material. *IEEE Sensor. J.* 2003, 3, 5–12.
69. Haddock, H.S.; Shankar, P. M.; Mutharasan, R., Fabrication of biconical tapered optical fibres using hydrofluoric acid. *Materials science and engineering B*, 2003, 97, 87–93.
70. Yuan, L.; Qui, A., Analysis of a single-mode fibre with taper lens end, *J. Opt. Soc. Am. A*, 1992, 9, 950–952.
71. Senior, J.M., *Optical fibre communications. Principles and practice*, Prentice Hall, Hertfordshire, 2nd edn., 1992, pp. 40–58.
72. Black, R.J.; Bourbonnais, R., Core-mode cutoff for finite-cladding lightguides, *IEE Proceedings-J.*, 133 1986, (6), 277–384.
73. Love, J.D.; Henry, W.M.; Stewart, W.J.; Black, R.J.; Lacroix, S.; Gonthier, F., Tapered single-mode fibres and devices. Part I: Adiabatic criteria, *IEE Proceedings-J*, 1991, 138(5), 343–353.
74. Bobb, L.C.; Shankar, P.M.; Krumboltz, H.D., Bending effects in biconically tapered single-mode fibres, *J. Light. and Tech.* 1990, 8, 1084–1090
75. Birks, T. A.; Russell, P.; St. and Pannel, C. N.; Low power acousto-optic device based on a tapered single mode fibre, *IEEE Phot. Tech. Let.* 1994, 6 725–727.
76. Shankar, P.M.; Bobb, L. C.; Krumboltz, H.D.; Coupling of modes in bent biconically tapered single-mode fibres, *J. of Light. Techn.*, 1991, 9(7), 832–837.
77. Matías, I.R.; Corres, J. M.; Arregui, F. J.; Bravo, J., Humidity sensors using nano-films deposited on hollow core fibres, *SPIE Newsroom*, International Society for Optical Engineering.
78. Rees, N. D.; James, S. W.; Tatam, R. P.; Ashwell, G. J., Optical fibre long-period gratings with Langmuir-Blodgett thin-film overlays, *Opt. Lett.*, 2002, 27, 686–688.

79. Del Villar, I.; Achaerandio, M.; Matias, I. R.; Arregui, F. J., Deposition of overlays by electrostatic self-assembly in long-period fibre gratings, *Opt. Lett.*, 2005, 30, 720–722.
80. Wang, Z. Y.; Heflin, J. R.; Stolen, R. H.; Ramachandran, S., Analysis of optical response of long period fibre gratings to nm-thick thin-film coatings, *Opt. Exp.*, 2005, 13, 2808–2813.
81. Kim, D.W.; Zhang, Y.; Cooper, K.L.; Wang, A., In-fibre reflection mode interferometer based on a long-period grating for external refractive-index measurement, *App. Opt.*, 2006, 44, 5368.
82. Chen, Q.; Lee, J.; Lin, M.R.; Wang, Y.; Yin, S.S.; Zhang, Q.M.; Reichard, K.A., Investigation of tuning characteristics of electrically tunable long-period gratings with a precise four-layer model, *J. Lightwave Technol.*, 2006, 24, 2954–2962.
83. Cusano, A.; Iadicicco, A.; Pilla, P.; Contessa, L.; Campopiano, S.; Cutolo, A.; Giordano, M., Cladding mode reorganization in high-refractive-index-coated long-period gratings: Effects on the refractive-index sensitivity, *Opt. Lett.*, 2005, 30, 2536–2538.
84. Pilla, P.; Iadicicco, A.; Contessa, L.; Campopiano, S.; Cutolo, A.; Giordano, M.; Cusano, A., Optical chemo-sensor based on long period gratings coated with δ form syndiotactic polystyrene, *IEEE Photon. Technol. Lett.*, 2005, 17, 1713–1715.
85. Cusano, A.; Iadicicco, A.; Pilla, P.; Contessa, L.; Campopiano, S.; Cutolo, A.; Giordano, M., Mode transition in high refractive index coated long period gratings, *Opt. Express*, 2006, 14, 19–34.
86. Erdogan, T.; Fibre grating spectra, *J. Lightwave Technol.*, 1997, 15, 1277–1294.
87. Anemogiannis, E.; Glytsis E. N.; Gaylord, T. K., Transmission characteristics of long-period fibre gratings having arbitrary azimuthal/radial refractive index variations, *J. Lightwave Technol.*, 2003, 21, 218–227.
88. Del Villar, I.; Matias, I. R.; Arregui, F. J.; Lalanne, P., Optimization of sensitivity in long period fibre gratings with overlay deposition, *Opt. Express*, 2005, 13, 56–69.
89. Del Villar, I.; Matias, I.R.; Arregui, F.J.; Achaerandio, M., Nanodeposition of materials with complex refractive index in long-period fibre gratings, *Journal of Lightwave Technology*, 2005, 23(12), 4192–4199.
90. Chung, K.W.; Yin, S., Analysis of a widely tunable long-period grating by use of an ultrathin cladding layer and higher-order cladding mode coupling, *Opt. Lett.*, 2004, 29, 812–814.
91. Lyons, E.R.; Lee, H.P., Demonstration of an etched cladding fibre Bragg grating filter with reduced tuning force requirement, *IEEE Photon. Technol. Lett.*, 1999, 11, 1626–1628.
92. Viegas, D.; Goicoechea, J.; Corres, J.M.; Matias, I.R.; Araújo, F.M.; Santos, J.L., Humidity sensing based on SiO_2 -nanospheres onto a Long-Period Fibre Grating, *OFS-2008 Optical Fibre Sensors International Conference*.

Overview of the OPTO-EMI-SENSE Project: Optical Fibre Sensor Network for Automotive Emission Monitoring

E. Lewis, G. Dooly, E. Have, C. Fitzpatrick, P. Chambers, J. Clifford, W.Z. Zhao, T. Sun, K.T.V. Grattan, J. Lucas, M. Degner, H. Ewald, S. Lochmann, G. Bramann, F. Gili, and E. Merlone-Borla

E. Lewis

Department of Electronic & Computer Engineering, University of Limerick, Ireland

G. Dooly

Department of Electronic & Computer Engineering, University of Limerick, Ireland

E. Have

Department of Electronic & Computer Engineering, University of Limerick, Ireland

C. Fitzpatrick

Department of Electronic & Computer Engineering, University of Limerick, Ireland

P. Chambers

Department of Electronic & Computer Engineering, University of Limerick, Ireland

J. Clifford

Department of Electronic & Computer Engineering, University of Limerick, Ireland

W.Z. Zhao

School of Engineering & Mathematical Sciences, City University, London EC1 0HB, UK

T. Sun

School of Engineering & Mathematical Sciences, City University, London EC1 0HB, UK

K.T.V. Grattan

School of Engineering & Mathematical Sciences, City University, London EC1 0HB, UK

J. Lucas

Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 3GJ, UK

M. Degner

Department of Electrical Engineering & Information Technology, University of Rostock, Germany

H. Ewald

Department of Electrical Engineering & Information Technology, University of Rostock, Germany

S. Lochmann

Department of Electrical Engineering & Computer Science, Hochschule Wismar, Germany

G. Bramann

Department of Electrical Engineering & Computer Science, Hochschule Wismar, Germany

F. Gili

Centro Ricerche Fiat, Strada Torino 50, 10043 Orbassano (TO), Italy

E. Merlone-Borla

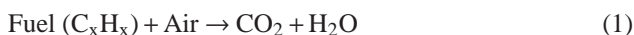
Centro Ricerche Fiat, Strada Torino 50, 10043 Orbassano (TO), Italy

Abstract An optical fibre based system has been developed which is capable of monitoring the presence of exhaust gas emissions and measuring their temperature on line in the exhaust system of a modern vehicle. There exists at present no commercial sensor, which is capable of providing online measurements of these exhaust gases as required by European legislation. The design of this sensor using low cost and compact optical components, which make it suitable for operation on board a vehicle, is discussed. The sensor is capable of detecting NO, NO₂, SO₂ to a minimum detection threshold of 5ppm, CO and HCs to a minimum threshold of 200 ppm, CO₂ in the range 300 ppm to 20% and temperature from 0°C to 900°C. Results measured in the exhaust of a modern engine are presented for each of these parameters.

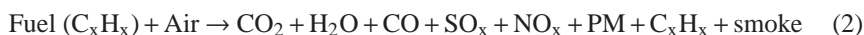
Keywords Mid-infrared gas detection · UV gas detection · in-fibre Bragg Grating temperature sensor · optical fibre sensor · vehicle emission detection

1 Introduction

Automotive emissions typically consist of water vapour, carbon dioxide (CO₂), carbon monoxide (CO), oxides of nitrogen (NO_x), oxides of sulphur (SO_x), smoke particles (diameters of 0.05 μm to 1 μm) and also particulate matter (diameters greater than 1 μm). Under perfect combustion conditions the following relationship would hold:



As carbon dioxide (CO₂) and water vapour (H₂O) are both present as trace gases in the atmosphere, no pollution would result from this process. However in reality perfect combustion does not occur and the following relationship holds as fuel is burnt in an engine:



Research has shown that each of these species is a threat to either human health or the environment [1]. Carbon monoxide (CO) is known to be poisonous to humans at concentrations above 400 parts-per-million. While CO₂ is not strictly considered a pollutant, as it exists naturally as a trace gas in the atmosphere, it is believed that the relatively high levels of CO₂ produced by combustion are a prime contributor to global warming [2]. As both CO and CO₂ have high absorption in the mid-infrared wavelength range [3], as shown in Fig. 1.

Similarly, detection of the other gases namely SO₂, NO and NO₂ is possible in the UV/Visible part of the spectrum. The theoretical spectra for these gasses are shown in Fig. 2.

Optical fibre sensors are particularly well suited to monitoring vehicle exhaust emissions, as they can be made small, lightweight, and as they are made purely

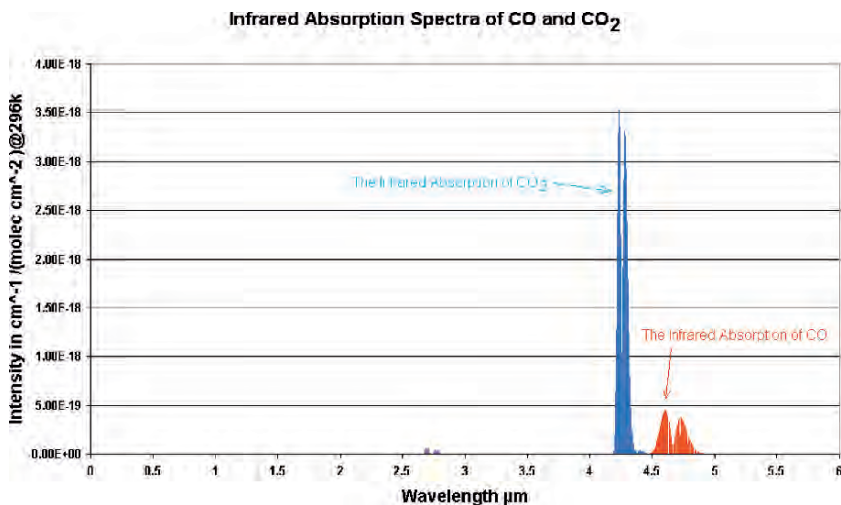


Fig. 1 Theoretical absorption Spectra for CO and CO₂ gases [3]

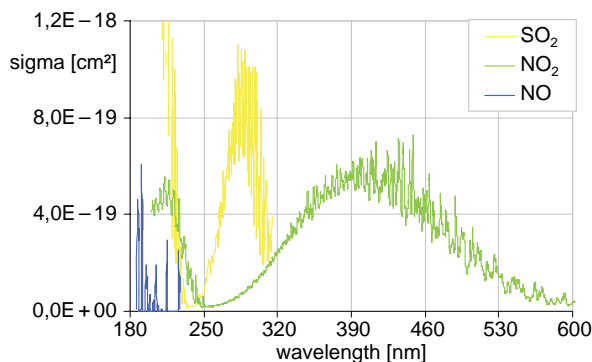


Fig. 2 Theoretical absorption Spectra for NO, NO₂ and SO₂ gases [3]

from silica glass (doped for high temperature measurement), quartz lenses (for UV) or chalcogenide with Calcium Flouride Lenses for Mid IR they can withstand the high temperature of the gases present in the exhaust system [4, 5].

2 Theoretical Background

The Beer-Lambert Law is used to calculate how much incident radiation is absorbed by a sample. The sample may be an aqueous solution or a gaseous quantity. If radiation of intensity I_0 is directed at a sample of path length l , radiation of intensity I_t leaves the sample. The absorbance A can be defined as:

$$A = \log_{10} \frac{I_o}{I_t} = \epsilon c l \quad (3)$$

where ϵ is called the molar absorption coefficient of the species in question, and c is the concentration of the sample. The ratio I_t/I_o is defined as the transmittance T , by substituting T into Eq. 3, manipulating the equation gives [6]:

$$T = 10^{-\epsilon c l} = \frac{I_t}{I_o} \quad (4)$$

This relationship was used to determine the concentration of the gas based upon experimental results of absorption observed in the mid infra red and UV part of the spectrum in the case of the gas sensor.

The Reference Forward Model (RFM) was developed at Oxford University to simulate the absorption spectra of gases in the HITRAN Database [7] such as CO at different concentrations, pressures, and temperatures. It was possible to use the RFM to vary the path length of the sample, to simulate the experimental results in the wavelength range of interest (i.e. within the pass band of the optical filter fitted to the pyroelectric detector).

Having performed the simulations at the various concentrations (1000 ppm, 800 ppm, etc) using RFM, the absorption spectra at these concentrations was then interpolated using MATLAB against the filter wavelength data so that it was over the same wavelength scale as the transmission spectrum of the band pass filter fitted to the pyroelectric detector. Having manipulated the data so that it was over the same wavelength range for both data sets (the absorption spectrum of the gas at ambient temperature and pressure over a path length of 360 mm, and the filter transmission spectrum), the absorption spectrum was converted to a transmission spectrum and this was multiplied by the filter transmission spectrum. This resulting spectrum corresponds to the transmission by CO at a particular concentration over a path length of 360 mm as measured by the pyroelectric detector. It is shown in Fig. 3 for concentrations of 0 ppm and 1000 ppm of CO over a path length of 360 mm, at 23°C and 1 bar of pressure (i.e. ambient temperature and atmospheric pressure). By calculating the area under the curve at a particular concentration, the theoretical values for I_o and I_t in Eq. 4 can be calculated. The area under the curve at 0 ppm corresponds to I_o , while the area under the curve at a particular concentration corresponds to I_t .

Figure 4 shows the analysis of the 200 ppm step test. The transmittance was calculated as the concentration of CO in the cell was increased from 0 ppm (when the cell was filled with N₂) to 1000 ppm and then decreased in steps of 200 ppm.

It is clear that the theoretical values (calculated by RFM, using the Beer-Lambert Law model) are in close agreement with the measured values. The largest deviation (1%) is at a concentration of 1000 ppm. The difference between the measured and theoretical results can be attributed to experimental uncertainty e.g. electrical noise on the outputs of the pyroelectric detectors. This could be reduced in future by improved the coupling of the emitter and detector to fibre which would increase the amount of radiant flux arriving at the detector which would increase the signal to noise ratio.

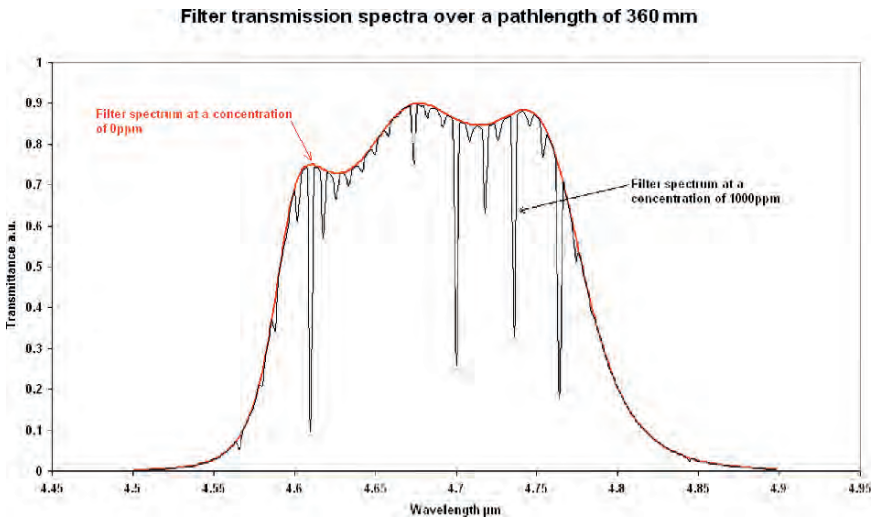


Fig. 3 A comparison of the transmission spectra for CO at 0 ppm and 1000 ppm calculated using RFM

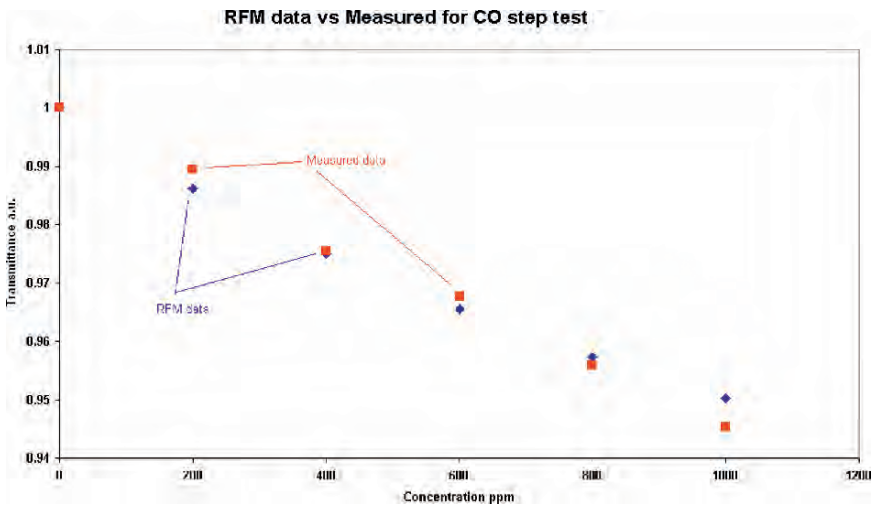


Fig. 4 A comparison of the transmittance values generated by the RFM simulation and those produced experimentally for CO

3 Experimental Results

3.1 Gas Measurement in the Mid Infra Red Range

The experimental rig for measuring CO, HC and CO₂ in the mid infra red region in the exhaust is shown in Fig. 5. The response of the sensor to 200 ppm step changes

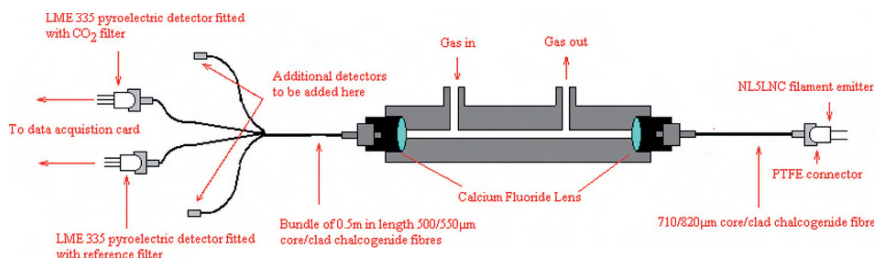


Fig. 5 The gas sensor for measuring in the mid infra red

of CO was carried out using the gas mixing facility at Centro Ricerche Fiat. The NL5LNC filament emitter was pulsed at 2 Hz with a 50% duty cycle. The radiant flux from the infrared emitter was guided to the 380 mm long test cell (effective path length of 360 mm) using a 710/820 μm core/clad chalcogenide fibre. The infrared radiation from the fibre was guided across the cell using a 25.4 mm CaF_2 collimating lens. A second CaF_2 lens was used to guide the collimated beam into a 500/550 μm fibre bundle, which guided the infrared beam to a pyroelectric detector fitted with a narrow band CO filter (centred at 4.66 μm with a 180 nm bandwidth) and a reference detector pyroelectric detector (centred at 3.95 μm with a 90 nm bandwidth). A Dell Latitude D610 notebook with a National Instruments PCIMCIA 6024E data acquisition card was used to acquire the output voltages of the pyroelectric detector and a Lab View Virtual Instrument was used to store these voltages to a file. Figure 1 shows the experimental set-up.

The gas cell was initially purged with nitrogen (N_2) supplied from a cylinder using a mass flow controller (MFC). The gas cell was then filled with 1000 ppm of CO for several minutes before the cell was purged with N_2 . Following this the cell was filled with 800 ppm of CO before again being purged with N_2 . This process was repeated in steps of 200 ppm of CO. The concentration of CO in the test cell was measured using a conventional ABB gas analyser. Figure 6 shows the results of the experiment. The ratio of the voltage on the CO pyroelectric detector to the reference pyroelectric detector was calculated and is given as the response of the sensor (to eliminate any drift in the output of the filament emitter). The concentration of CO measured by the ABB (Advance Optima) gas analyser during the experiment is also shown.

The above cell was modified in order to measure CO_2 emissions. The requirements for CO_2 emissions measurement in the modern car are quite different from CO and HCs as CO_2 is generated in the range 0 to 15% depending on the driving conditions. In this case the path length can be made much shorter (according to the Beer Lambert Law) and consequently the sensor can be correspondingly smaller. With this scenario two different sensors were designed and fabricated, a short transmissive one and a reflective version which also has a short path length (the path length in the two cases were 20 mm for the transmissive and 37 mm for the reflective).

The tests in these cases were conducted using a mass flow controller system to accurately control the levels of CO_2 entering the sensors. The CO_2 concentration

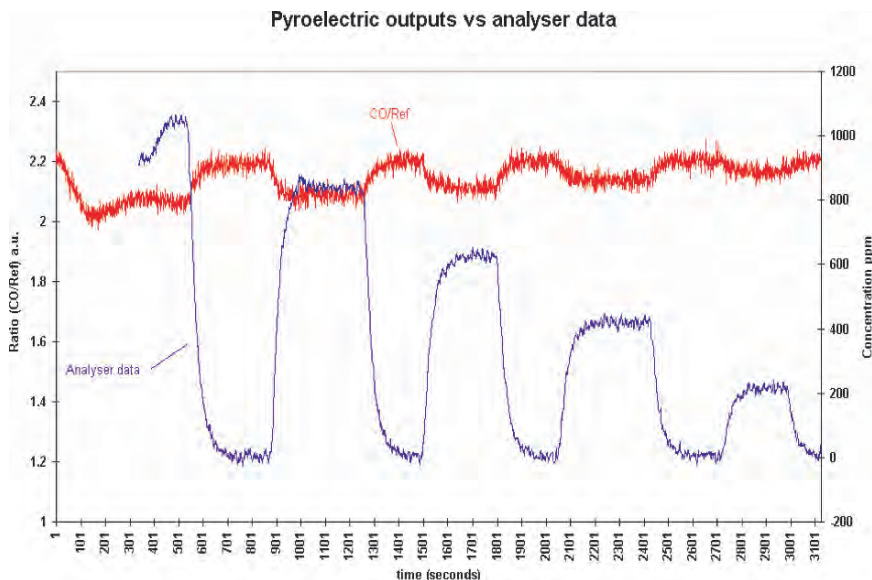


Fig. 6 Results of CO recorded in 200 ppm steps compared with a reference commercial Instruments

was accurately measured using the ABB (Advance Optima) gas analyzer. The test cycle used was to start with a level of 12 percent CO₂ in nitrogen reduce this to zero after 2 minutes and increase it to 10% following a further 2 minutes. This cycle was repeated until a level of only 1 percent was input to the system. The results of changing gas concentration in this manner in the case of the transmissive element is shown in Fig. 7 and the reflective sensor in Fig. 8.

The results of Fig. 7 and 8 show that the short path length CO₂ sensors is able to reproduce the fluctuation in CO₂. The transmissive sensor (Fig. 7) shows excellent rise time response being at least as good as the commercial gas analyzer. The reflective sensor exhibits rapid rise and fall times (Fig. 8).The two versions of the Mid infra red CO₂ optical fibre sensors were also mounted on the exhaust of the Fiat Croma. These are shown photographically in Fig. 9 (a and b).

The transmissive optical fibre sensor was connected in the exhaust system of the vehicle as shown in Fig. 9 and the output of the optical fibre sensor and reference instrument were recorded simultaneously whilst the car was driven according to the standard driving pattern i.e. the Extra Urban Driving Cycle (EUDC) whilst on a roller test bench (instrumented rolling road). These are shown in Fig. 10.

It is clear from Fig. 10 that the Optical Fibre Sensor is capable of faithfully reproducing the variation of CO₂ concentration over the whole NEDC cycle. A similar series of tests were conducted using the reflective mode mid infra red optical fibre sensor. The results of these tests are shown in Fig. 11.

The results of Fig. 11 again show that the reflective mode optical fibre sensor is capable of tracking the CO₂ concentrations during the full duration of the NEDC

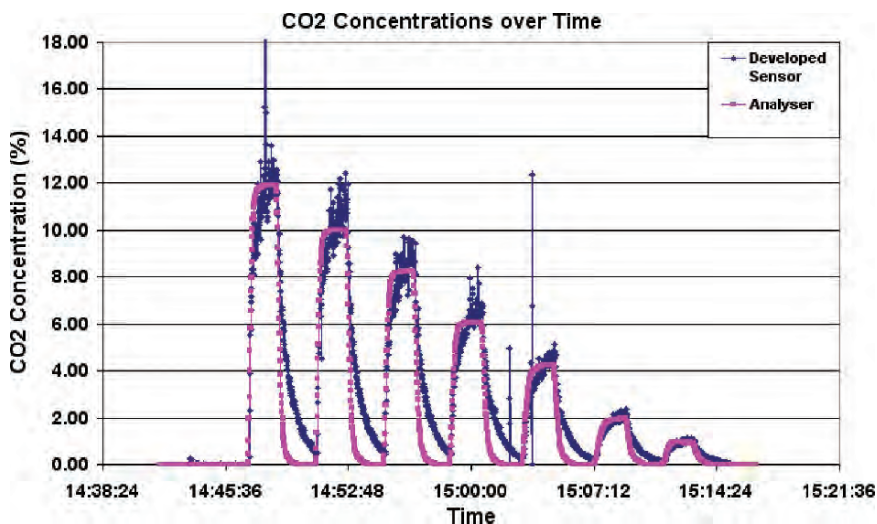


Fig. 7 Measurements of CO₂ concentration versus time for the 20 mm transmissive mode Mid-IR optical fibre sensor compared with reference instrumentation values acquired simultaneously

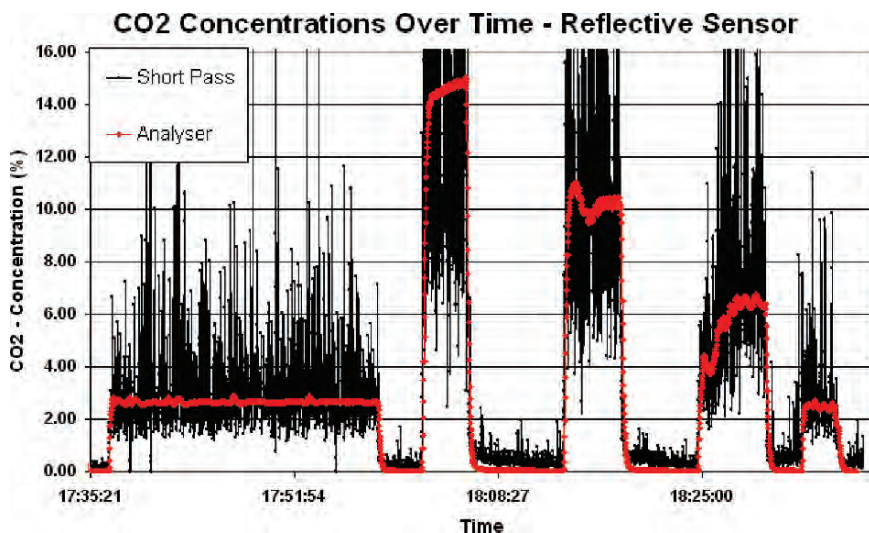


Fig. 8 Measurements of CO₂ concentration versus time for the 20 mm reflective mode mid-IR optical fibre sensor compared with reference instrumentation values acquired simultaneously

cycle. The optical fibre sensor output in this case is more noisy than the transmissive sensor, but this is due to the lower received signal power at the detector which is due to the reflective geometry of the sensor. However, the reflective sensor is a compact plug-like sensor and could be screwed in directly into the exhaust system. Further optimisation of this geometry will yield a superior signal to noise ratio.

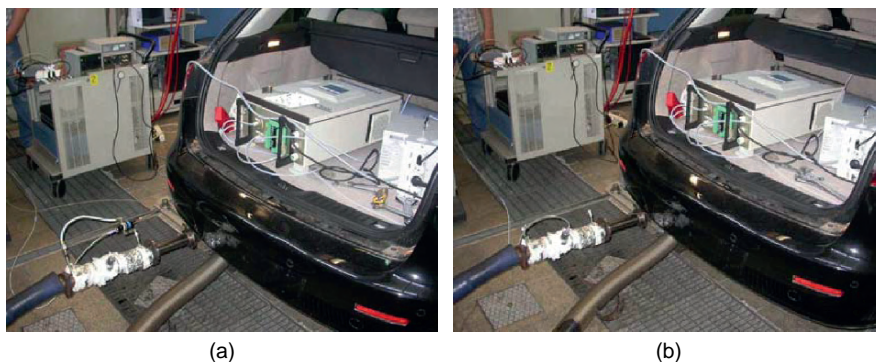


Fig. 9 Installation of the short path length mid IR sensor on the exhaust line of fiat crom a (a) transmissive (b) reflective

3.2 Gas Measurement in the Ultra Violet Range

The system for measuring the gases NO, NO₂ and SO₂ in the UV range is shown schematically in Fig. 12. The Principal of operation for the UV cell for inclusion in the exhaust of the vehicle is shown in Fig. 12 and Figs. 13 and 14).

Figure 13 shows the output spectrum from a number of LEDs which are simultaneously connected to the input side of the optical gas cell via individual optical fibres. An alternative to this is to use a broadband optical source (Deuterium Source) through a single optical fibre (This is shown schematically in Fig. 12. The detector responses shown in Fig. 14 are for two separate filtered detectors e.g. photodiodes. Alternatively, a low cost spectrometer can be used to give spectral resolution.

In a further development from this principle a reflective version of the UV sensor has been developed which operates on the same optical principal as the above (absorption spectroscopy) but uses reflection from a single surface to interrogate the absorption signal i.e. a different geometry. The packaged system is shown photographically in Fig. 15 below.

Figure 16 shows the pipe section with the optical fibre cell mounted within it prior to installation under the demonstrator vehicle. A number of test results were obtained using this arrangement using accurately controlled gas concentrations both individually and in mixture in the test facility at CRF.

The reflective version of the same sensor is shown photographically in Fig. 17. This has the advantage that it can be screwed into an aperture in the exhaust system (in this case the removal pipe section).

The optical fibre UV gas concentration sensor was mounted on the detachable exhaust pipe section as shown in Fig. 16. The following results of gas concentration versus time in the UV/VIS spectral range were obtained simultaneously with reference instrumentation in the laboratory, MIR 9000 for SO₂, H₂O (IR principle), TOPAZE 3000 for NO, NO₂, NO_x (Chemiluminescent principle).

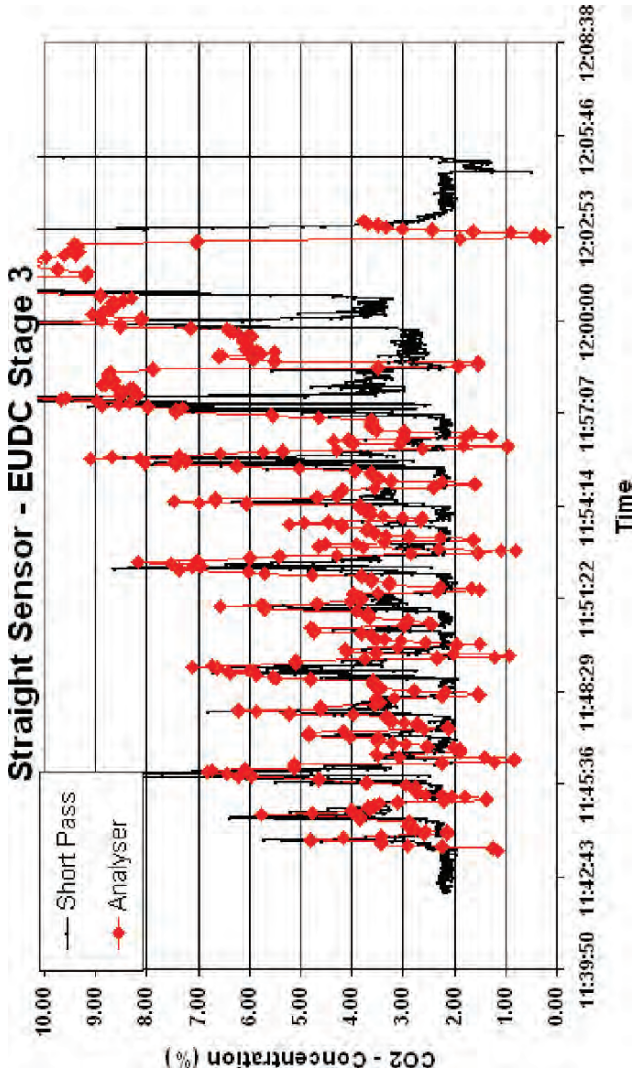


Fig. 10 CO₂ concentration recorded during the NEDC test cycle for the fiat cruma on the rolling road. Simultaneous measurements shown for the transmissive optical fibre and reference sensors

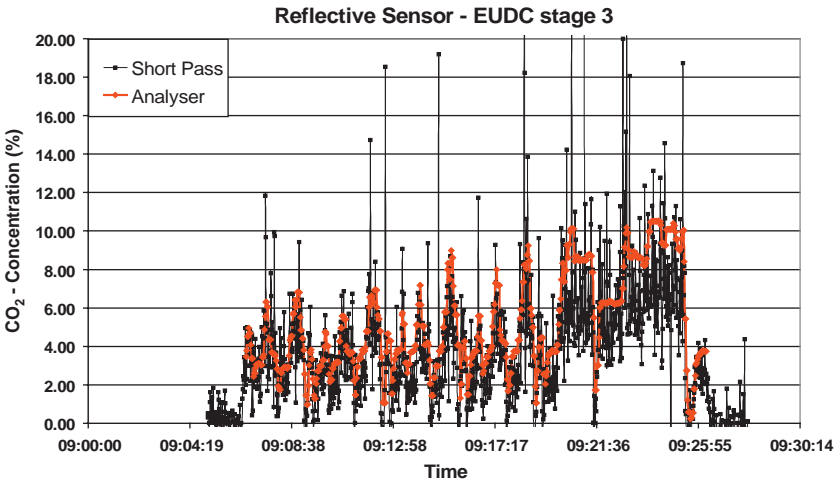


Fig. 11 CO₂ concentration recorded during the NEDC test cycle for the fiat roma on the rolling road. Simultaneous measurements shown for the reflective optical fibre and reference sensors

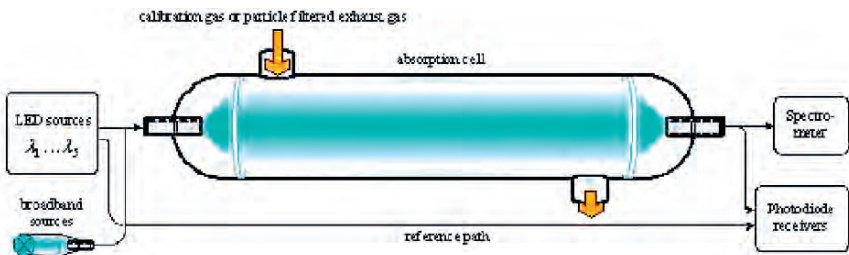


Fig. 12 The gas sensor for measuring in ultra violet range

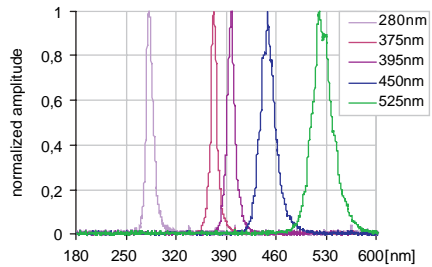


Fig. 13 UV source LEDs spectral output

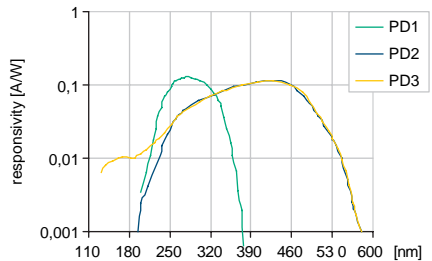


Fig. 14 Detector responses for the UV gas sensor

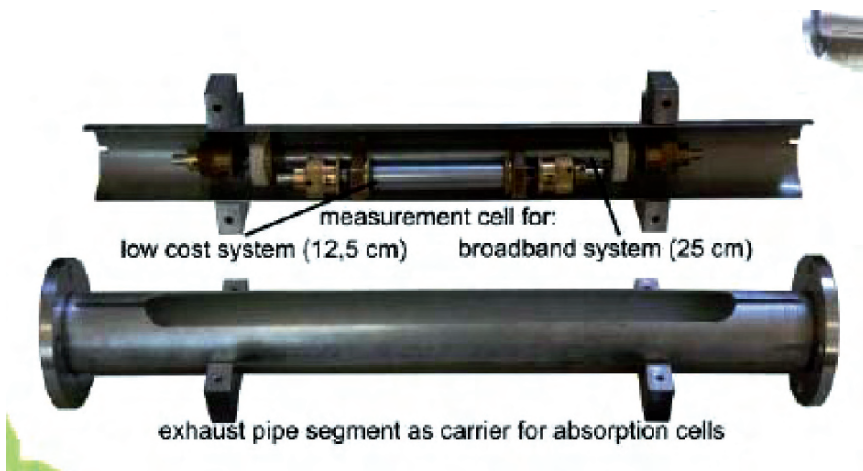


Fig. 15 UV sensor physical construction

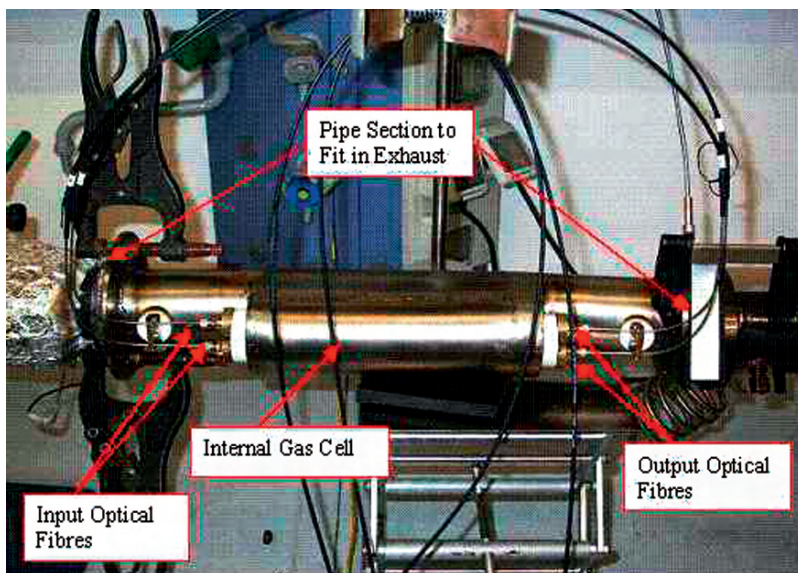


Fig. 16 Photograph of the modified pipe section including the UV optical fibre sensor for gas concentration measurement for vehicle mounting

Figure 18 shows results of optically measured NO_2 concentration while the concentration has been varied in a 'stepwise' manner in the controlled conditions available at the CRF test lab.

The results of Fig. 18 clearly show that the values of NO_2 concentration as measured by the optical fibre sensor using a 20 cm path length transmissive cell are in close agreement with those obtained using the lab based reference instrumentation.

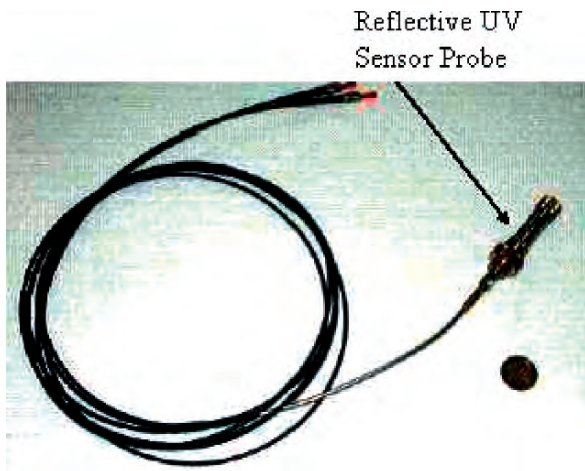


Fig. 17 The reflective UV optical fibre sensor for mounting under the vehicle

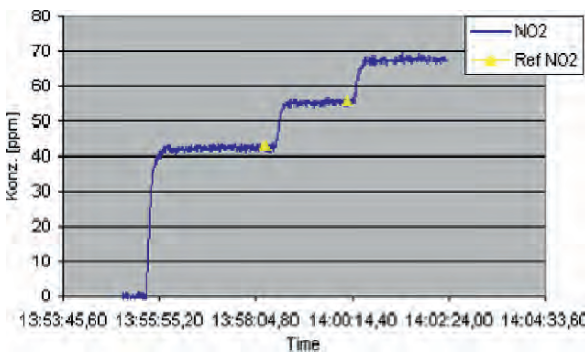


Fig. 18 Measurements of NO₂ concentration versus time for the LED based UV optical fibre sensor compared with reference instrumentation values acquired simultaneously

In this case the excellent agreement is such that the two sets of data (single point value for the reference instrumentation are indistinguishable. Figure 19 shows a similar sequence of tests but in the reverse direction.

The results of Fig. 19 show a very similar trend to that in Fig. 18 and therefore shows that the sensor is capable of faithfully reproducing NO₂ concentrations in increasing as well as decreasing quantities.

During performing these tests it was clear that the speed of response of the optical fibre sensor was superior to that of the reference instrumentation and was of the order of 100 msec. The lower detection limit of the optical fibre sensor appears to be at a level of about 3 ppm (Clear in Fig. 19.)

Figure 20 shows results of optically measured SO₂ concentration while the concentration has been varied in a 'stepwise' manner in the controlled conditions available at the CRF test lab.

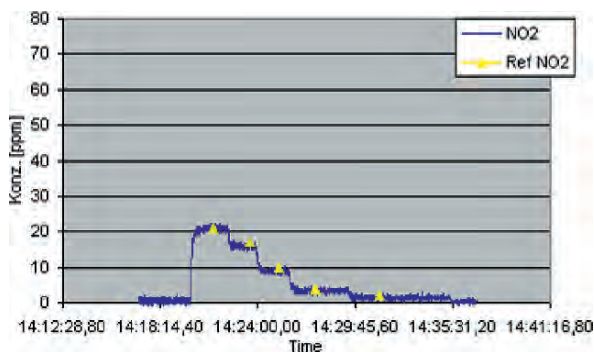


Fig. 19 Measurements of NO_2 concentration versus time for the LED based UV optical fibre sensor compared with reference instrumentation values acquired simultaneously

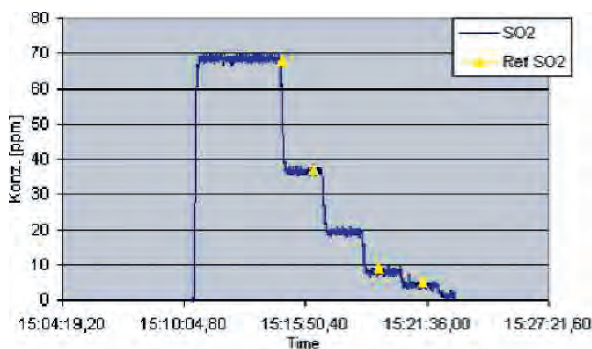


Fig. 20 Measurements of NO_2 concentration versus time for the LED based UV optical fibre sensor compared with reference instrumentation values acquired simultaneously

The results of Fig. 20 again show excellent agreement between the values measured by the optical fibre UV absorption cell and the reference instrument in the laboratory of CRF. The resolution of the optical fibre sensor at low concentration is such that the lower detection limit appears to be around 2 ppm. Again it was clear that the speed of response of the optical fibre sensor was superior to that of the reference instrumentation and was of the order of 100 msec.

The cell shown in Fig. 16 was inserted in line in the exhaust system under the car. This is shown photographically in Fig. 21.

The cell was used to record the levels of NO , NO_2 and SO_2 for a full cycle of the standard acceleration/ deceleration test with the car mounted on a rolling road at the test facility of CRF in Turin. The results of these tests corresponding to NO_2 are shown in Fig. 22.

It is clear from Fig. 22 that the value of NO_2 recorded on the optical sensor faithfully reproduces the values measured on the reference gas analysis instrumentation. The optical fibre sensor has therefore been proved to be capable of measurement within the exhaust of the vehicle.



Fig. 21 The UV gas sensor mounted underneath the car

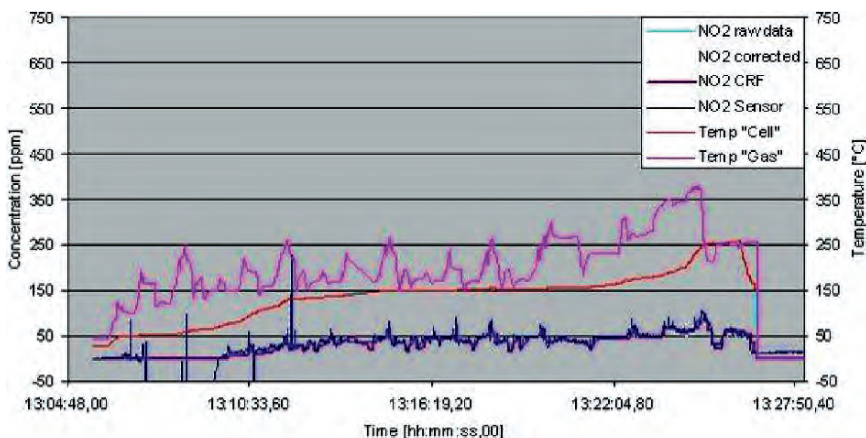


Fig. 22 NO₂ test results under the car with simultaneous reference instruments recording

3.3 Optical Fibre Temperature Measurement

As well as measuring the gas concentrations using optical fibre technology, the OPTO-EMI-SENSE project has been concerned with the measurement of temperature of the exhaust gases using in-fibre Bragg Gratings. The system for the temperature measurement is shown schematically in Fig. 23.

The FBG-based temperature sensor system utilizes a broad band (Superluminescent LED) light source (centred on a wavelength of 1550 nm) and a Fabry-Perot tunable filter for FBG wavelength interrogation. The temperature sensor located on the test exhaust system at the laboratory of CRF is shown photographically in Fig. 24 and the results of a typical temperature cycle corresponding to values encountered in a standard test cycle shown in Fig. 25.

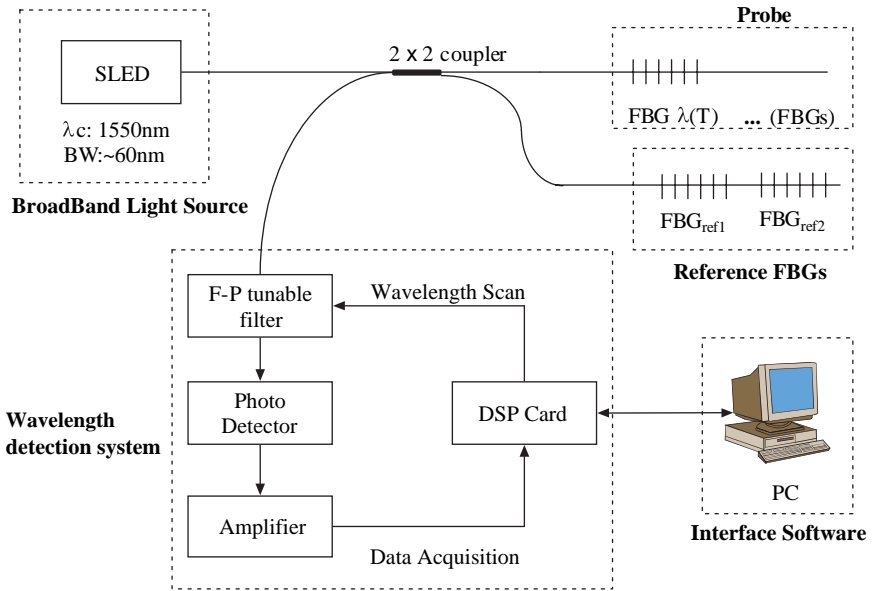


Fig. 23 The optical fibre sensor for measuring temperature

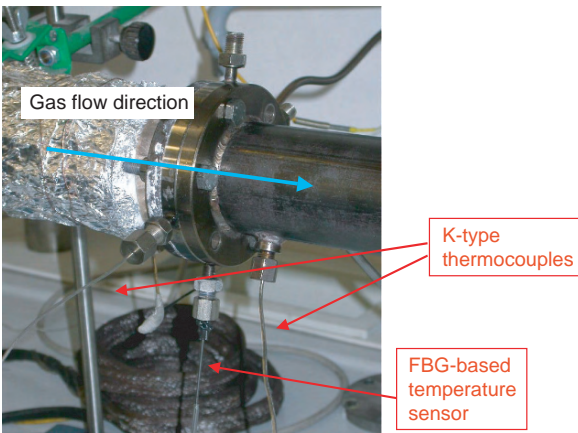


Fig. 24 The FBG optical fibre sensor mounted on the experimental exhaust system at the laboratory of CRF, turin

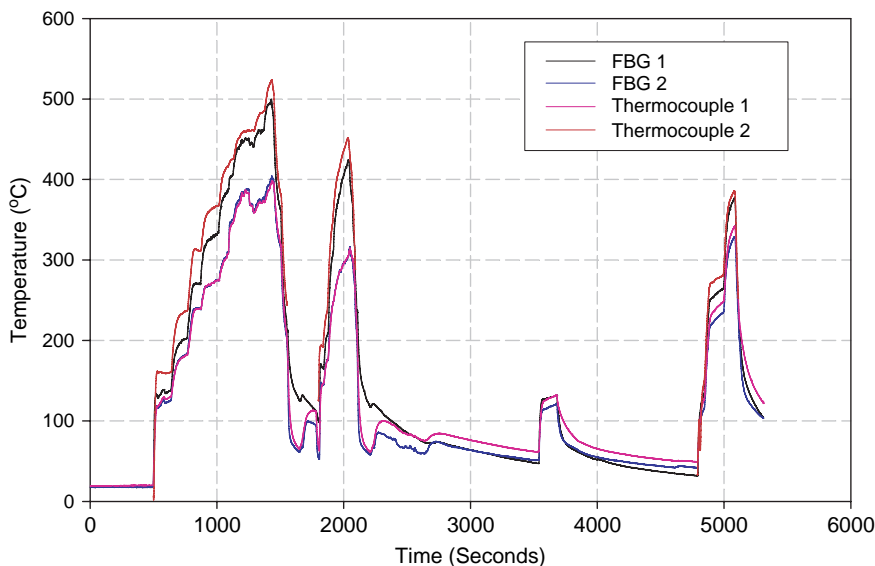


Fig. 25 Temperature measurements on the experimental exhaust test facility with simultaneous reference instruments recording

4 Conclusions

Optical fibre sensor suitable for the detection of exhaust gas emissions and temperature has been described in this paper. The development of the sensors are novel as they use a low cost and compact components coupled to optical fibre, to provide a practical solution for the measurement in the harsh environment of the car exhaust system.

This sensors have proved to be capable of detecting gas concentrations as low as single ppm values for NO, NO₂ and SO₂, 200 ppm of CO (and CO₂) and Hydrocarbons (Non Methane HCs). An analysis of these results using the Reference Forward Model (RFM) and MATLAB indicated that the measured and theoretical values are in close agreement. Optical Fibre Temperature Measurements have been performed in the exhaust of a Diesel engine and these have demonstrated the sensor’s capability of accurately (to within one degree) measuring the exhaust gas temperature over a range of many 100 degrees centigrade (at least 800°C).

Acknowledgments The authors would like to thank the staff of the Emissions Group at Centro Ricerche Fiat for their assistance during testing.

The authors would also like to acknowledge the support of the EU FP6 project Opto-Emi-Sense (Contract number: TST3-CT2003-506592) for funding this work.

References

1. Hillier, V. A. W. (1991) "Fundamentals of Motor Vehicle Technology", 4th Ed., pp. 131–136.
2. Energy Information Administration (Official Energy Statistics from the U.S. Government) <http://www.eia.doe.gov/oiaf/1605/ggcebro/chapter1.html>
3. Rothman, L.S. et al. (2004) "The Hitran Molecular Spectroscopic Database and HAWKS (HI-TRAN Atmospheric Workstation)" *Journal of Quantitative Spectroscopy & Radiative Transfer* Vol. 96, pp. 139–204.
4. "Handbook of Optical Fibre Sensing Technology" (2002) edited by José Miguel López-Higuera pp. 287–286.
5. Stewart, G., Jin W., Culshaw B. (1997) "Prospects for Fibre Optic Evanescent Field Gas Sensors Using Absorption in the Near-infrared," *Sensors & Actuators B38*, pp. 42–47.
6. McCash E. B. C. (1994). "Fundamentals of Molecular Spectroscopy", McGraw, Hill p. 19.
7. Reference Forward Model <http://www.atm.ox.ac.uk/RFM/>

Part III
Wireless Sensors

Wireless Sensor Networks and Applications

Yueh-Min Huang, Meng-Yen Hsieh and Frode Eika Sandnes

Abstract Wireless sensor networks are collections of highly distributed, small, and lightweight wireless sensor nodes that monitor the environment or systems through physical measurement. Once the Zigbee and the IEEE 802.11 standards are approved and embraced, wireless sensor nodes will be capable of supporting interoperability among a wide range of mobile and fixed devices from different manufacturers. This paper reviews some of the fundamental mechanisms of wireless sensor networks including their architecture, topology, data integration, routing techniques, and applications. Sensor network applications include both military and civilian monitoring in both rural and urban environments. Wireless sensor networks hold great potential for improving control, conservation, convenience, efficiency, reliability flexibility, and safety in network environments.

Keywords Wireless sensor networks · topology · routing · data integration · architecture · heterogeneity · WSN applications

1 Introduction

Recent advances in micro-electro-mechanical systems (MEMS) technology, wireless communication and digital electronics have allowed the realisation of wireless sensor networks. A wireless sensor network is a highly distributed or organized network of many small and lightweight sensor nodes that are used for monitoring, detecting, and tracking operations. Low cost, low power and multifunctional

Yueh-Min Huang

Department of Engineering Science, National Cheng-Kung University, Taiwan, ROC,
e-mail: huang@mail.ncku.edu.tw

Meng-Yen Hsieh

Department of Computer Science and Information Engineering, Providence University, Taiwan,
ROC, e-mail: mengyen0501@gmail.com

Frode Eika Sandnes

Faculty of Engineering, Oslo University College, Norway, e-mail: frodes@hio.no

sensor nodes can detect physical characteristics of their environment and perform primitive computation. Measurements sensed throughout the network are collected, aggregated and forwarded to a central point for processing.

Wireless sensor networks (WSN) can provide low-cost and flexible wireless communication for system automation applications, especially in urban areas. Different sensor nodes can capture data such as seismic, magnetic, thermal, visual, infrared, acoustic, and radar. Possible measurements include temperature, pressure, humidity; vibration, radiation, noise level, vehicular movement, soil erosion, the presence or absence of objects, mechanical stress levels on objects, and the speed, direction, and size of an object. WSN can also be applied to continuous sensing, event detection, location sensing, and local control. Some WSN applications must operate in terms of a long time-perspective. Therefore, only a small fraction of the sensors are active during network operation, while the others are in sleeping mode to conserve energy. WSN applications also need to simplify network operation and data management to support faster time-to-deployment, and easier application integration. Various research projects for wireless sensor networks include SensorNet, WINS, SPINS, SINA, mAMPS, LEACH, SmartDust, SCADDS, PicoRadio, PACMAN, Dynamic Sensor Networks, Aware Home, COUGAR and Device Database Project DataSpace [1, 2].

This overview focuses on some wireless sensor applications. In the next section, low-rate wireless communication technology is described. Then, fundamental techniques and network architectures are presented. Finally, sensing objectives and wireless sensor applications are introduced.

2 Low-Rate Wireless Communication Technology

Wireless sensor networks nodes rely on low data rates, very long battery life (several months or even years) and very low computational complexity associated with the processing and communication of the collected information across the WSN. The first edition of the 802.15.4 standard [3], released May 2003, specifies the physical layer and medium access control for low-rate wireless personal area networks (LR-WPAN's) [4] intended for sensor networks, home automation, and remote controls. The 802.15.4 standard addresses low power consumption, low-cost and low-speed ubiquitous communication between devices. Low communication rates require less power than high speed communication. Hence, nodes designed according to the standard operate on one of three unlicensed frequency bands, namely 868–868.8 MHz in Europe with 1–3 channels, 902–928 MHz in North America with 10–30 channels, and 2400–2483.5 MHz worldwide with up to 16 channels. The 2003 version of the standard specifies the use of direct sequence spread spectrum (DSSS) at the physical layers. The 868/915 MHz bands can therefore deliver transfer rates of 20 and 40 kbit/s, respectively, and the 2450 MHz band delivers a rate of 250 kbit/s. The 2006 revision includes improvements that allow maximum data

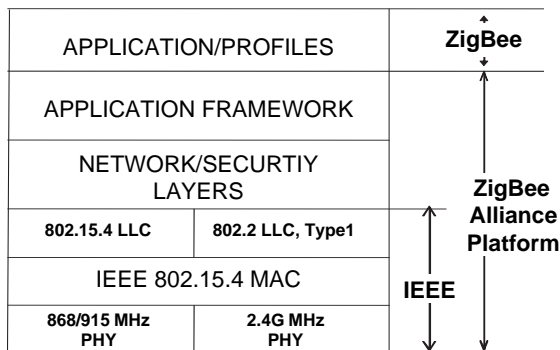


Fig. 1 Zigbee/802.15.4 architecture

rates for the 868/915 MHz bands with up to 100 and 250 kbit/s, respectively. The standard specifies a communication range of 10 to 20 meters. CSMA/CA is used for collision avoidance with four basic types of frame transport (data acknowledgment, beacon, and MAC command frames).

ZigBee [5] is designed to perform high level communication using small, low-power digital radios based on the IEEE 802.15.4 standard. ZigBee addresses the unique needs of low-cost, low-power, wireless sensor networks in remote monitoring, home control, and building automation network applications. Zigbee Alliance [6] is an industry working group that has developed standardized application software on top of the IEEE 802.15.4 wireless standard in collaboration with the IEEE. Manufacturers have supported the development of fully compliant IEEE 802.15.4 and Zigbee single chip solutions. Figure 1 depicts the Zigbee architecture.

The Zigbee specification supports inexpensive and robust networking in environments with a very large number of nodes. It should be noted that, although similar, Zigbee, Bluetooth, and WLAN(Wi-Fi) are designed for different purposes and different applications: The Wi-Fi standard results in high node costs and complex power-hungry RF ICs. Wi-Fi is therefore not suitable for low-power, low-cost and low data rate applications. Typical Wi-Fi applications include wireless local area network connectivity and broadband Internet access. Bluetooth™ was designed with low power consumption, short range and low-cost transceiver technology in mind. However, they are also designed to support a variety of duty cycles, medium data rates and high quality of service (QoS) networks with few active nodes. For example, wireless connectivity between devices such as phones, PDAs, laptops, audio headsets is a common Bluetooth application. On the contrary, Zigbee is capable of very low duty cycle for both static and dynamic networks. For example, the transition from sleep mode to data transition is much faster with ZigBee than Bluetooth. Table 1 summarizes Zigbee, Bluetooth and Wi-Fi characteristics.

Table 1 Zigbee, Bluetooth, and Wi-Fi characteristics

	Zigbee (802.15.4)	Bluetooth (802.15.1)	Wi-Fi (802.11)
Data rate	20–250 kbps	1 Mbps	11 & 54 Mbps
Range	10–100 meters	10 meters	50–100 meters
Frequency band	868 MHz (Europe) 900–928 MHz (NA), 2.4 GHz (worldwide)	2.4 GHz	2.4 and 5 GHz
Transmit power	0.5, 1, or 3 mW	1, 2.5 or 100 mW	100 mW
Nodes per network	256+	8	unlimited (Depending on applications)
Topology	Ad hoc, peer-to-peer, star, mesh	Ad hoc, infrastructure	Ad hoc, very small networks
Complexity (Device and application impact)	Low	High	High
Power Consumption	(Very) Low	High	Medium

3 WSN Fundamentals

Fundamental WSN issues including network topology, routing protocols, and data integration modes are introduced in this section.

3.1 Network Topology

Two well-know WSN network structures are the mesh and star topologies, shown in Fig. 2. A star topology network is a single-hop system where any of wireless sensor nodes can directly connect to a gateway/base station. The gateway/base station can

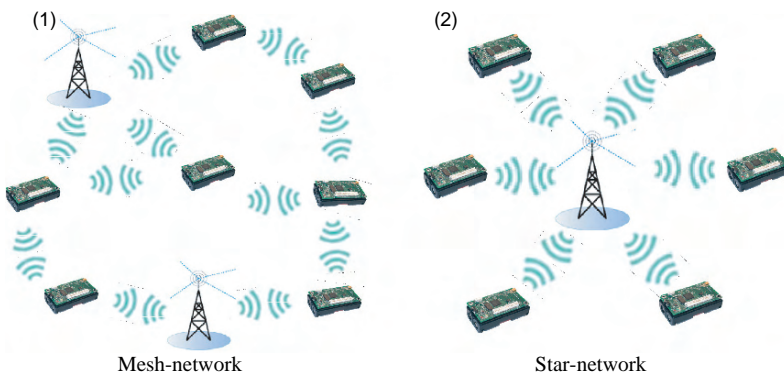


Fig. 2 Two well-known WSN topologies

be a computer, PDA, dedicated control device, or some other embedded server. The wireless sensor nodes in such networks can employ short-range communication links that support distances from ten to a few hundreds meters, which helps reduce the network power consumption. However, the connection between a sensor node and the gateway/base station can become obstructed, especially if many sensor nodes are close to the gateway/base station or the communication channel is being jammed. Alternatively, WSN with a mesh topology may need to send data via multiple hops to communicate for data to reach its destination. However, two nodes may exchange messages through several different paths, and mesh topologies can therefore be used to balance traffic loads via the available paths. Mesh networks are also more fault tolerant as alternative routes can be chosen if nodes become unavailable.

Star and mesh topologies can be combined to exploit both the low power and simplicity of the star topology, and the extended range and fault tolerance of the mesh topology.

3.2 Routing Protocols

Several WSN applications provide self-organizing and self-healing wireless sensor nodes through the routing protocols of the network layers. A self-organizing sensor node can discover alternative routes and deliver data to the gateway/base station or other nodes interested in the data, even when radio links along their established route are broken. Most WSN routing protocols are inspired by traditional wired and wireless routing protocols. In general, there are two classes of routing techniques, namely proactive and reactive. Proactive routing protocols employ static route information, typically in the form of sensor node routing tables. Consequently, such networks require fixed topologies since the proactive route update algorithms is resource intensive in terms of power consumption and network bandwidth. Reactive routing protocols require each node in the WSN to establish and maintain routes on demand. Protocols that are effective in wireless mobile environments, such as dynamic source routing (DSR) [7] and ad hoc on demand distance vector (AODV) protocols [8], may not necessarily be suitable for all WSN applications due to the resource-constrained micro-sensors (e.g. MICA motes) [9]. Most WSN applications need to operate more efficiently with more efficient routing protocols that consume less power. With limited resources, routing protocols may be implemented by optimizing the duty cycle and minimizing overheads.

There is a comprehensive set of routing techniques for WSN applications [10]. WSN routing protocols can be divided into three categories according to their network structure, namely flat-based routing, hierarchical-based routing, and position-based routing.

Furthermore, depending on the characteristics and operations of the protocols, they can also be classified into negotiation-based routing, multi-path routing, query-based routing, QoS-based routing, and coherent-based routing. There is a trade-off between energy consumption and communication overhead savings, but the common objective is to extend the battery lifetime of the wireless sensor nodes.

3.3 Data Integration Modes

Data integration models address the flow of data in terms of how those data are used. Wireless sensor networks allow three modes of data integration, namely data collection, data broadcast, and bidirectional dialog.

3.3.1 Data Collection Mode

Wireless sensor network applications may require the gateway/base station to collect data from the wireless sensor nodes. Three strategies for data collection include event driven data collection, periodic sampling, and store & forward. An active fire alarm system that detects fire, or the effects of fire, is an example of an even-driven system. Wireless sensor nodes are used to collect fire information such that occupants can be notified, the fire brigade can be summoned and fire alarm components in a building can be controlled. Periodic sampling is typically performed in environmental monitoring applications where measurements are taken from sensor nodes surrounding a location or a target of interest at regular intervals. Store and forward is typically used in cold chain systems [11] to perform a series of storage and distribution operations at a given temperature range for a network of data collecting sensor nodes deployed in a temperature-controlled supply chain.

3.3.2 Broadcast Data Mode

A gateway/base station may need to broadcast data to sensor nodes. Typically, wireless sensor nodes simply execute the commands given by the control node during a broadcast. Smart lighting is an example of an application that employs data burst-broadcast across a network of wireless sensor nodes. A smart lighting application is triggered by positioning signals and is therefore sensitive and responsive to the presence of people. Such lighting systems provide different levels of light in context of time, location and the presence of people. This is known as ambient intelligence.

3.3.3 Bidirectional Dialog Mode

WSN applications require two-way communication between the sensor/actuator nodes and the gateway/base station. Bidirectional dialogue can be performed through polling or on-demand. Controller applications in building automation use polling. The controller polls each device associated with a device ID in the network, typically by sending a serial query message and then waiting for a response. For example, the controller of an energy management application polls thermostats, variable air volume sensors, and climate control nodes. The on-demand data model allows a mobile gateway to bind to a network, actively gather data from assigned sensor nodes, and then leave the network. The on-demand model allows one mobile sensing device to

bind to multiple networks and multiple mobile gateways to bind to a given network. For example, patients in hospitals can wear sensors to monitor various medical conditions. Then, doctors can access those data via a PDA. The PDA constitutes the mobile gateway that binds to the local sensor network on-demand.

4 WSN Architectures

The organization of a wireless sensor network is affected by factors such as scalability, fault tolerance, power consumption and environmental conditions. Self-organizing wireless sensor network mechanism has received significant attention in the sensor network research community. Clustered or connected dominating sets, star, tree, grid and mesh-based topologies have all been applied to wireless self-organization networks [12, 13, 14, 15, 16, 17]. Self-organizing networks rely on the autonomous and correct operations of their individual sensor nodes. Most sensor nodes are therefore equipped with an intelligent self-organizing mechanism for wireless computing. This section reviews the prevalent self-organizing sensor node architectures including tier-based and cluster-based architectures.

4.1 Tier-based Architectures

Multi-tier wireless sensor network architectures (often referred to as n-tier architectures) have been proposed in the literature [16, 17]. A multi-tier wireless sensor network can be viewed as a wireless transport mechanism for connecting hierarchical heterogeneous nodes. Benefits of heterogeneous multi-tier sensor networks include:

- Heterogeneous wireless nodes can be equipped with different transport medium with different range of coverage and different specifications including CPU, memory, and peripherals to meet specific needs.
- Multi-tiered architectures are better suited for static and dynamic topologies for immobile and mobile wireless sensor nodes than planar network architectures.
- Heterogeneous sensor nodes support wireless networks with long-distance remote access, robust connectivity, and scalability.
- Wireless sensor network types such as mesh, location-based, real-time, query-based and event-driven networks are adaptable to multi-tier network architectures.

4.1.1 Planar Wireless Sensor Network

Homogeneous sensor nodes scattered over a sensing region form a planar wireless sensor network (see Fig. 3). The sensor nodes are responsible for collecting information from its sensors and deliver the result to a fixed device, or a high-performance computing system. The sensing data from the sensors, called source

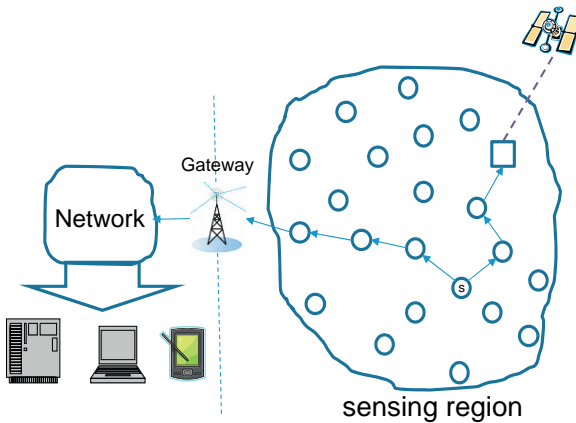


Fig. 3 A planar wireless sensor network

nodes, is forwarded to a remote sink, or gateway, by the means of a multi-hop routing protocol. Homogeneous planar wireless sensor networks are problematic. For example, packet loss along the multi hop delivery path will occur if an intermediate node suffers a link failure. Multi-hop planar networks are not scalable as the hop count is related to the total number of nodes. It takes longer to traverse a longer path than necessary and the network nodes are unnecessarily drained for valuable power. Moreover, nodes neighboring a fixed sink or a gateway will need to handle heavy traffic loads and hence consume more energy.

4.1.2 Two-Tiered Sensor Network Architectures

A two-tiered wireless sensor network has been proposed as a solution to the problems of planar wireless sensor networks. Two-tiered heterogeneous wireless sensor networks consist of a low tier and an upper tier (see Fig. 4). The low-tier network comprises many energy constrained sensor nodes with low data rates,

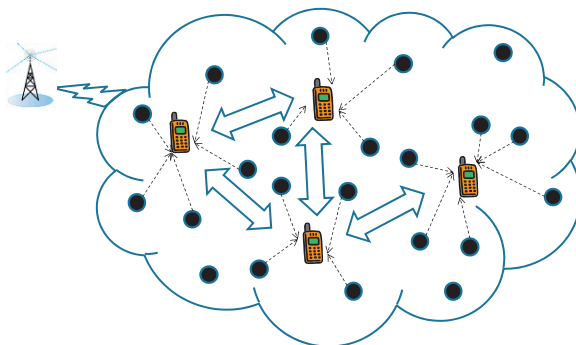


Fig. 4 A two-tiered wireless sensor network

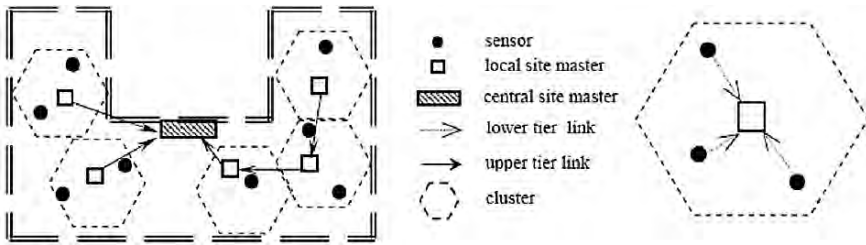


Fig. 5 Two-tier wireless sensor network for structural health monitoring (Figure taken from [17])

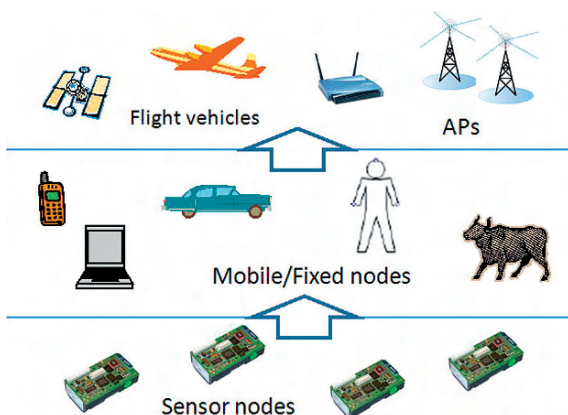
short transmission ranges and limited computing capabilities. The upper tier network comprises several coordinator devices with high data rates, large transmission ranges, and powerful computing capabilities. Example coordinator devices include mobile phones, laptops and PDAs. Nodes in the upper tier may be mobile and easily moved to any part of the low-tier network to gather information packets from local sensor nodes. Sensing data therefore flows from the fixed sensor nodes to mobile coordinator nodes instead of sensor nodes in the vicinity of the gateway. Consequently, energy is conserved throughout the network. In addition, wireless access technologies such as Wi-Fi (802.11) and Bluetooth can be utilized by the coordinator nodes.

Figure 5 illustrates a two-tier network architecture for structural health monitoring. The low tier comprises battery powered sensor units communicating with low data rates in the order of tens of Kbps in the 915 MHz ISM band. The upper tier network consists of several local site masters powered from the mains outlets with a battery backup. Data flows at rates in the order of Mbps in the 2.4 GHz ISM band. A local site master is responsible for coordinating, and collecting measurements from a cluster of assigned fixed peripheral sensor units. This network can also be regarded as a cluster-based wireless sensor network. Cluster-based architectures are discussed in the next section.

4.1.3 Three-Tier Sensor Network Architectures

A three-tier heterogeneous wireless sensor network can be established by layering one tier upon a two-tier network by using access points, flight vehicles, or other high capability machines (see Fig. 6). The three tiers in the network are termed the bottom, middle, top tier, respectively. Wireless sensor nodes distributed randomly in the lowest layer network communicate with the middle layer network comprising fixed or mobile coordinator nodes in close proximity. The packet forwarding distance between the bottom tier and the middle tier should be no longer than a few hops. The top tier network usually consists of a number of high capacity access points that connects to other wireless and wired networks. Top tier nodes share the coordinating responsibility with nodes in the middle tier. For example, the coordinator role can be dynamic whereupon the coordinator nodes in the middle tier just have to forward packets from the bottom tier to the top tier without aggregation. The two upper tiers must help reduce the power consumption to extend the lifetime of

Fig. 6 A three-tier wireless sensor network architecture



the whole network by minimizing the work of the wireless sensor nodes. The middle tier must simultaneously employ multiple transmission frequencies to communicate with both the bottom tier and the top tier networks. The two upper tiers in a three-tier architecture network should be designed with different movement in mind. The middle tier network should allow mobility while the top tier network remain static. Alternatively, the coordinator nodes can be fixed with a dynamic top tier network, such as flight or mobile vehicles.

The traffic monitoring system MULE [18] is an example of a three-tier sensor network architecture that provides wide-area connectivity for a sparse sensor network by exploiting mobile agents such as people, animals, or vehicles moving in the environment. The top tier comprises WAN connected devices, the middle tier comprises mobile transport agents and the bottom tier comprises fixed wireless sensor nodes. Key traits of a MULE include large storage capacities (relative to sensors), renewable power, and the ability to communicate with the sensors and networked access points. To ensure data reliability acknowledgments are sent between sensor nodes and their upper devices. Throughput can be improved, without the aid of upper tiers, though always-on connections provided by cellular or satellite phones. Consequently, a number of tiers in a multi-tier architecture could be collapsed onto one node according to different applications, situations and needs. SensEye [19] is a multi-tier network of heterogeneous wireless sensor camera nodes organized hierarchically across multiple tiers, unlike two-tier surveillance networks with low power cameras at the bottom tier that trigger higher resolution cameras at the upper tier. The multi-tier network achieves an order of magnitude reduction in energy consumption compared to single-tier networks, without sacrificing reliability.

4.2 Cluster-based Architectures

A cluster-based network architecture groups wireless sensor nodes into a number of clusters controlled by part of nodes playing a particular role denoted as cluster

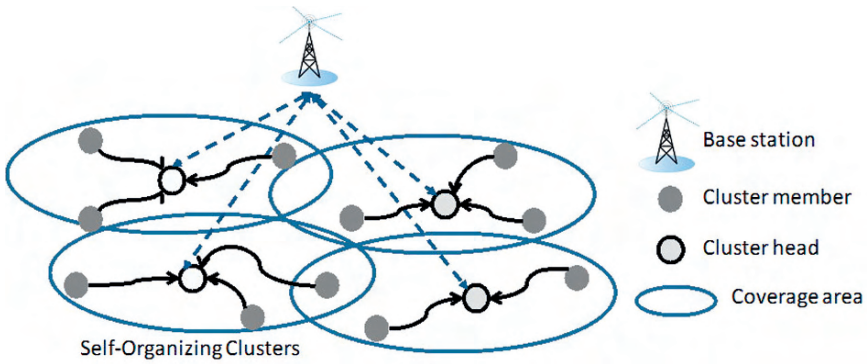


Fig. 7 The architecture of a cluster-based wireless sensor network

heads (CH) (see Fig. 7). CH neighbors its member nodes. Member nodes are associated with a cluster via a one-hop link and located inside the edge of the CHs' coverage area and these member nodes performs sensing and forwarding. After gathering or aggregating localized sensing information from its cluster members nodes a CH sends packets to the base station (BS), which usually is an access point connected to a wired network. Cluster heads may act as aggregation nodes by combining local sensing packets and forward the processed information. An important distinction between the cluster-based and tier-based wireless sensor network architectures is that the clustering-operations are performed spontaneously through self-organization. The election of CHs and the formation of clusters should be autonomous. Cluster management involves handling members joining and leaving the network. Moreover, the members of each cluster may operate according to an agreed schedule such that radio components of each non-cluster-head node can be tuned off at all times except during transmission. Figure 8 shows how sensor nodes in a cluster with active/sleep modes are scheduled into a number of transmission subgroups.

A well-known cluster-based network, the Low-Energy Adaptive Clustering Hierarchy (LEACH) [20], rotates CHs during different cluster rounds. LEACH uses localized coordination to achieve scalability and robustness for dynamic networks.

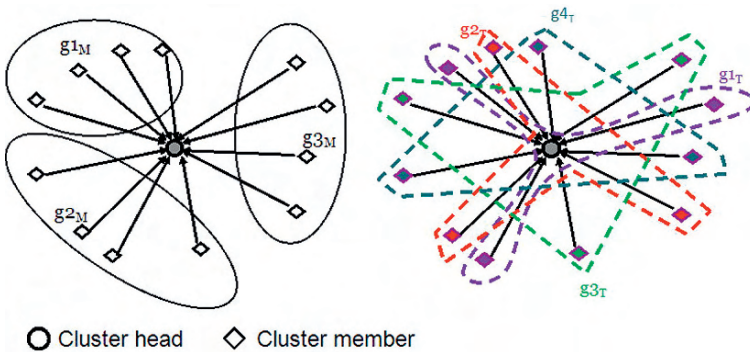


Fig. 8 Cluster transmission subgroups

Data fusion is incorporated into the routing protocol to reduce the amount of information transmitted to the sink. The rotating of CHs has the effect of averaging the energy consumption across the sensor nodes, thus reducing the probability of network disruption due to dead nodes. Local CHs are elected from sensor nodes at any given time with a certain probability. These CHs advertise their status to the network. Each of other sensor nodes then joins the CH that requires the least communication energy. Although, CHs consume more energy during the relay of compressed or aggregated data to the BS, this only affects a small subset of nodes. The details of LEACH are as follows: Node n selects a random number between 0 and 1. The time threshold, $T(n)$, for becoming a CH for the current round is given by:

$$T(n) = \begin{cases} \frac{x}{1 - x \times (r \bmod 1/x)} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases}$$

where x is the portion of nodes that should be CHs and r is the current round, and G is the set of nodes that have not been CHs in the last $1/x$ rounds.

5 WSN Applications

A wide range of WSN applications and systems have been developed in recent years. WSN applications can be classified according to their design dimension [21] or areas of deployment [22].

WSN design dimensions include deployment, mobility, resources, cost, energy, heterogeneity, modality, infrastructure, topology, coverage, connectivity, size, lifetime and QoS. Example applications include Great Duck (bird observation on Great Duck island), ZebraNet, Glacier Monitoring, Cattle Herding, Bathymetry, Ocean Water Monitoring, Grape monitoring, Cold Chain Management, Rescue of Avalanche Victims, Vital Sign Monitoring, Power monitoring, Parts Assembly, Tracking Military Vehicles, and Self-healing Mine Field and Sniper Localization [11, 21, 23].

According to areas of deployment WSN applications can be industrial, military, location oriented, public safety oriented, automotive, airport oriented, agricultural, emergency handling, medical and oceanic. Some WSN applications employ motes and smart dust sensors in the networks, such as general indoor/outdoor environmental monitoring, industrial and vibration monitoring, test and measurement, advanced wireless and available sensor boards systems, etc. Other WSN applications have been made with MEMS-based sensor solution for applications in the electric sector.

Early developments in wireless sensor networks were motivated by military application [24]. However, in recent years there has been a diversification towards civilian applications, including industrial, classroom/home, buildings, environmental monitoring, habitat monitoring, structural monitoring, health monitoring [11, 23, 25, 26, 27, 28, 29, 30, 31] to mention a few.

5.1 WSN Applications Supporting Static Routing

Several point-to-point and point-to-multipoint WSN applications are intended for the commercial market. In those applications, the gateway/base station utilizes static communication paths to connect the sensor nodes. Static routing WSN applications includes home automation, building automation, industrial automation, medical control, residential control, transportation, and various remote control and monitoring systems. These applications use wireless sensing techniques (IEEE 802.15.4) for real time control and connectivity for all types of sensing devices. This is different to that of other wireless applications such as metropolitan transport (IEEE802.15.3/WiMax), enterprise wireless LANs (IEEE802.11 series/WiFi), and wireless personal networks (IEEE802.15.1/Bluetooth). The interoperability and RF performance characteristics of the IEEE 802.15.4 conformant Zigbee technology provide the foundation for wide scale deployment of static routing WSN applications, such as those mentioned in [32].

5.1.1 Home Control/Automation

Home control/automation wireless heterogeneous sensor networks applications provide humans with control, conservation, convenience, efficiency, and safety. Wireless sensor networks with anything from a few nodes to hundreds of nodes can be deployed in home networks. Figure 9 shows an example of a distributed home control/automation network with real-time control mechanisms.

Sensor nodes may wirelessly control the lights, switches, blinds and thermostats. A single press of a remote control button could trigger a chain-reaction among the controlled devices. For example, a personal video player may trigger other events such as light dimming, lowering of blinds, and turning on the TV. Another remote control button could activate a change in the room atmosphere. For example, there

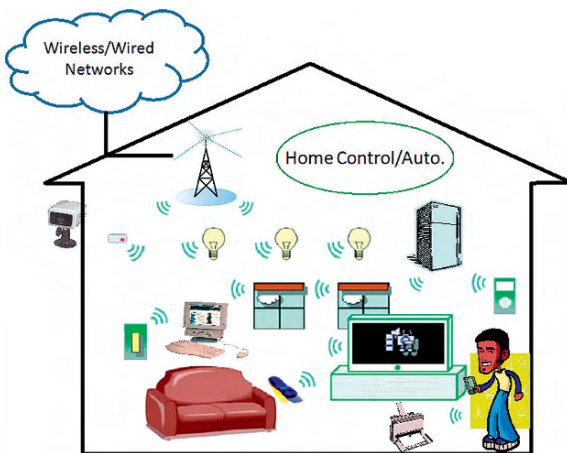


Fig. 9 Zigbee home automation network

could be a work mode button which would lower the air conditioner or heating in the room except the atrium and turn off all lights in the home. Appliances, personal computers and peripherals, entertainment systems, and other devices in the home control/automation network are easily configured by using simple remote controls or the internet access.

Zigbee home automation systems have enabled the control of lighting, window shades, vehicle ventilation, heating and air conditioning, and security systems. Specific functions include flexible management of lighting, heating, cooling, and shade systems from anywhere in the home, optimal consumption of natural resources in the home, real-time and precise data capture of electricity usage, water usage, and gas usage, notification of detection of unusual events, and software installation and updates through wireless connections;

5.1.2 Building Automation

Building automation applications based on wireless heterogeneous sensor networks provide offices and buildings with control, conservation, flexibility, and safety (see Fig. 10). Traditional building automation systems (BAS) connect sensor nodes and actuator to controllers using wires through a bus network topology. Wireless BAS can be realized using a distributed wireless bus controller, also called a wireless field bus. BAS that utilize wireless sensor technologies can achieve cost-savings as labor costs traditionally associated with wiring are eliminated. Wireless sensor networks easily adapt to changing floor plans with IEEE 802.11 enabled devices such as PDAs, cameras and laptops. For examples, a wireless BAS application can control and improve the indoor climate in a building. Other wireless BAS applications such as surveillance and fire detection must support point-to-point quality of service, including real-time operation, the minimum bandwidth guarantees, and delay limits.

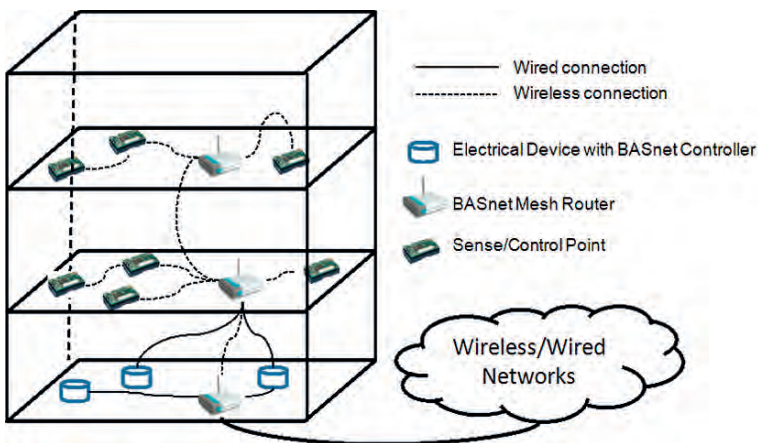


Fig. 10 A wireless mesh-based BAS network

Furthermore, temporary BAS installations are possible without cabling. However, a number of fundamental technological issues hinder the deployment of wireless heterogeneous BAS sensor networks. For example, BAS wireless sensor network nodes have insufficient battery capacity to sustain a video surveillance application which may be needed in a building. As building automation installation and maintenance expenses continue to rise there has to be shift towards affordable wireless sensor nodes. Several examples of wireless sensor networks integrated into building automation bus system exist including EIB and its successor KNX, LON, BACNET, the IBBT WBA project, and Arts Centre Vooruit [30].

Zigbee nodes provides building automation networks with several functions that help achieve control, conservation, flexibility, and safety, namely (1) management of integrating and centralizing the lighting, heating, cooling, and security mechanisms, (2) automatic control systems for improving conservation, flexibility and security, (3) energy reduction with optimized heating, ventilation and air conditioning (HVAC) management, (4) equitable allocation of utility costs based on actual consumption, (5) adaptive reconfigurable workspaces, (6) simple extension and upgrading of building infrastructure, (7) data integration and networking from multiple access points, and (8) wireless monitoring for enhanced perimeter protection.

5.1.3 Industrial Automation

Wireless heterogeneous sensor network industrial automation applications provide industries with control, conservation, efficiency, and safety. Wiring and maintaining sensor networks can be costly, dangerous, and difficult due to incompatible sensor nodes and control system protocols. The introduction of wireless sensor networks, such as wireless FieldBus and wireless industrial Ethernet, in industrial environments, can help reduce the problems associated with traditional cabling. New opportunities emerge for OEMs if they introduce wireless sensor network based remote monitoring, predictive maintenance, or precision instruments. Industrial automation wireless sensor networks scale to hundreds of nodes, can operate for many years on batteries to support remote deployment, bring new possibilities, new markets, new value to plant and warehouse managers, and allow faster retrofitting that reduce labor costs and delays. Because of their reliability, power-efficiency, adaptability, and scalability, wireless sensor networks have been applied to many industrial automation applications such as pressure/flow/temperature monitoring, machine condition monitoring, precision instrumentation, plant-wide telemetry, compliance and quality measurements, overlay monitoring, supervisor control and data acquisition (SCADA) systems, machine health diagnostics, waste water and tank monitoring, utility power-line monitoring and automotive performance monitoring.

Zigbee based industrial automation networks provide several functions that help achieve control, conservation, efficiency, and safety, namely reliability extension in existing manufacturing and process control systems, asset management improvement with continuously monitoring of critical equipment, energy cost reduction by optimizing manufacturing processes, identification of inefficient operation or poorly

performing equipment, automation of data acquisition from remote sensors for reducing user intervention, improvement of preventive maintenance programs, improved employee and public safety, and collection of streamlining data for improved compliance reporting.

5.2 WSN Applications Supporting Dynamic Routing

Several WSN applications utilize multiple hop radio connectivity between wireless sensor nodes in a mesh network, such as military applications, habitat monitoring applications, and various self-organization or reconfigurable wireless sensing applications. These applications may be fixed or may evolve. Applications are designed to detect events or monitor the environment through a large number of unattended wireless sensor nodes. Usually, the wireless sensor nodes are randomly scattered and self-organize in unreachable regions. Some WSN applications are equipped with camera-imaging or complementary metal-oxide semiconductor (COMS-based) sensors to track or observe objects. Design issues associated with WSN applications, in context of randomly distributed and uncontrolled wireless sensor nodes, include scalability, connectivity, reliability, and security. Two examples [35, 36] of such networks can be found in military and environmental monitoring applications (see Figs. 11 and 12).

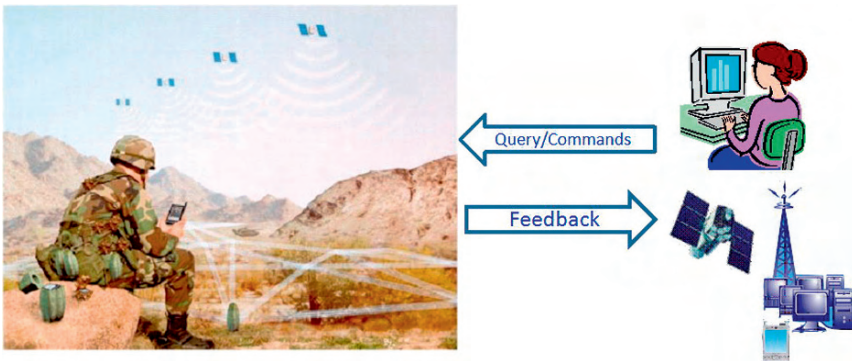


Fig. 11 WSN military applications (Reproduced with permission from USC information sciences institute [35])

5.2.1 Military Applications

Military sensing networks are designed to detect and gain as much information as possible about enemy movements, explosions, and other phenomena. Therefore, wireless sensor nodes are integrated with military command, control, communications, computing, intelligence, surveillance, reconnaissance and targeting systems.

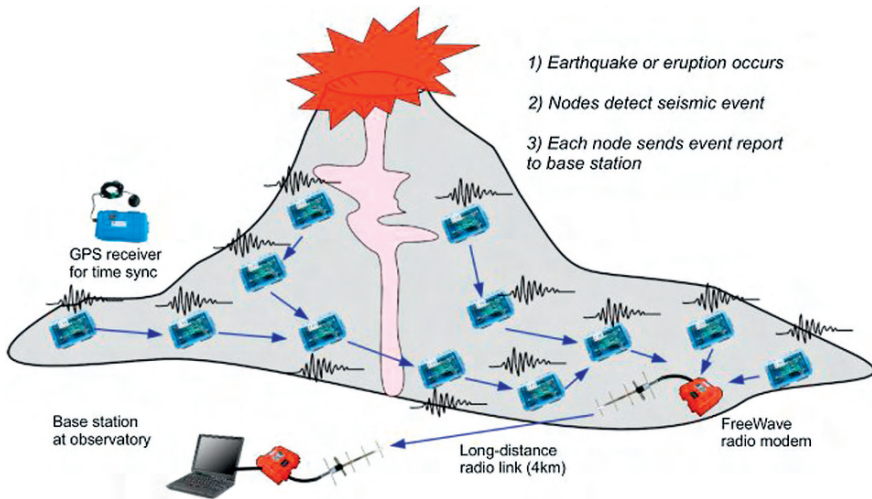


Fig. 12 A WSN environmental monitoring application (Figure taken from [36])

Examples of military wireless sensor network applications [2, 24, 25, 35] are battlefield surveillance, guidance systems for intelligent missiles, detection of attacks by weapons of mass destruction such as nuclear, biological, or chemical, and other monitoring applications (e.g. board monitoring, friendly forces, equipment and ammunition monitoring, etc).

For security and battlefield surveillance applications, wireless sensor networks are applied to area and theater monitoring, instead of single high-cost sensing assets. The fault tolerance and self-organizing characteristics of wireless sensor nodes means that they can be deployed easily and rapidly by untrained troops in any situation to accomplish various missions, self protection, and the prevention of danger. For example, critical terrains, approach routes, paths and straits can be covered and the activities of the opposing forces can be closely watched.

Wireless sensor networks can also be incorporated into guidance systems of intelligent ammunition to achieve improved targeting. Sensing techniques has greatly improved the accuracy of precision-guided weapons systems (cruise missile, surface-to-air missile, air-to-surface missiles, air-to-air missiles, surface-to-surface missiles, ship-borne missiles and antitank missiles): The precision of terminal weapons systems have also improved greatly.

To detect attacks by weapons of mass destruction, it is necessary to develop a nuclear, biological, and chemical (NBC) communications network with various wireless sensor nodes. In chemical and biological warfare environments, it is important to be close to ground zero for real-time and accurate detection of any hazardous material. The WSN NBC systems allow wireless sensor nodes to connect to a gateway to warn military personnel within critical reaction time if a hazardous material id detected. Casualties drastically minimized as a result. Wireless sensor nodes can be used as an invaluable nuclear reconnaissance tool that prevents exposing a rescue team to nuclear radiation after an NBC attack.

For condition-based monitoring, Rockwell Scientific has developed a wireless sensor network specifically tailored for monitoring complex machinery and processes on board U.S. Navy ships. Monitoring of machinery systems, also known as condition-based maintenance (CBM), has become increasingly important. Adequate machinery monitoring can help optimize machinery usage and minimize manufacturing costs.

5.2.2 Environment Applications

Examples of environmental WSN applications include tracking the movements of animals, and insects, forest fire detection, habitat monitoring, flood detection, precision agriculture and civil and environmental engineering monitoring.

Forest fires are uncontrolled fires that occur in wild areas which cause significant damage to natural and human resources. A forest fire wireless sensor network detection system can relay the exact origin of the fire and minimize the scale of the disaster. A forest fire detection system can also report vital information about the fire to a processing center which can coordinate the alerting of local residents and the dispatching of fire fighters. Wireless sensor networks can help provide real time fire detection with high accuracy. Wireless sensor nodes are deployed uniformly and randomly in the wild forest by, for example, throwing them from an aircraft.

Wireless sensor network habitat monitoring holds enormous potential for monitoring plants and animals. Data needs to be collected for a long time. For example, a habitat monitoring system used for controlling the climate comprises a collection of for monitoring and controlling humidity and temperature. INSIGHT (INternet-Sensor InteGration for HabitaT monitoring) [33] is an example of a habitat monitoring system. INSIGHT is energy-efficient, as the sensor node batteries only need replacing about every 6 months (Lithium batteries last a year). Furthermore, researchers can remotely query and reconfigure the INSIGHT wireless sensor network via the Internet.

A flood detection system forecasts the rainfall runoff patterns as heavy long term rainfall often results in flooding in urban regions. Wireless sensor networks can detect water level using ultrasonic water level sensors, monitor the water situation with video and still image cameras, and connect to a back-end server. Several systems for detecting flood have been proposed, such as the COUGAR Device Database Project at Cornell University, the DataSpace project at Rutgers, and the ALERT system [2].

WSN precision agriculture applications focus on observing, assessing and controlling agricultural practices such as the pesticides level in the water, the level of soil erosion, and the level of air pollution. For example, wireless sensor networks can monitor micro-climates in a crop field.

Civil and environment engineering monitoring involves monitoring building, bridges, and other large structures. Several wireless sensor networks for structural health monitoring are currently being developed. Such monitoring systems need data acquisition, data compression and global clock synchronization. Wireless sensor nodes are capable of self-diagnosis for potential problems and self-repairs. For

example, a wireless sensor network for structural health monitoring is deployed and tested on the 4200ft long main span and the south tower of the golden gate bridge [34]. The Wisden WSN application is another example that incorporates reliable data transport using a hybrid of end-to-end and hop-by-hop recovery in a multi-hop topology, and low-overhead data time-stamps that does not require global clock synchronization.

6 Conclusions

Wireless sensor networks are well suited to remote sensing applications due to their flexibility, scalability, fault tolerance, high sensing fidelity, low-cost and rapid deployment and reconfiguration. Wireless sensor networks technology was surveyed with particular emphasis on their low-rate wireless communication, routing protocols, network architecture and applications. The Zigbee and 802.15.4 standards have triggered a wide range of application that are becoming increasingly commonplace and this development is expected to continue as the sensing technology and wireless communication technologies evolve. Wireless sensor network design issues related to various application types has been discussed including synchronization, connectivity, security, and real-time communication.

References

1. Gungor V.C., Lambert F.C. (2006) A survey on communication networks for electric system automation. *Computer Network* 50(7):887–897.
2. Akyildiz I.F., Su W., Sankasubramaniam Y., Cayirci E. (2002) Wireless Sensor Networks: A Survey. *Computer Networks*, 38:393–422.
3. Howitt I., Gutierrez J.A. (2003) IEEE 802.15.4 Low Rate – Wireless Personal Area Network Coexistence Issues. In: Proc. IEEE Wireless Communications and Networking Conference – WCNC’2003, pp. 1481–1486.
4. Gutierrez J.A., Naeve M., Callaway E., Bourgeois M., Mitter V., Heile, B. (2001) IEEE 802.15.4: A Developing Standard for Low-Power Low-Cost Wireless Personal Area Networks. *IEEE Network*, pp. 12–19.
5. Huang H.C., Din J.W., Huang Y.M. (2006) An Implementation of Battery-aware Wireless Sensor Network Using ZigBee for Multimedia Service. In: Proc. International Conference on Consumer Electronics, pp. 369–370.
6. Zigbee Alliance HomePage, URL <http://www.zigbee.org/en/index.asp>.
7. Broch J., Johnson D., Maltz D. (1998) The dynamic source routing protocol for mobile ad hoc networks, URL <http://www.ietf.org/internet-draft/draft-ietf-manet-dsr-01.txt>
8. Perkins C., Royer E., Das S. (1999) Ad hoc on demand distance vector (ADOV) routing, URL <http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-03.txt>
9. Jin, McEachen K.S., Singh J.C., Gurminder (2006) RF Characteristics of Mica-Z Wireless Sensor Network Motes. In: Proc 49th IEEE International Midwest Symposium on Circuits and Systems – MWSCAS’2006, San Juan, PR., pp. 100–104.
10. Al-Karaki J.N., Ahmed E. (2004) Routing Techniques in Wireless Sensor Networks: A Survey. *IEEE Wireless Communications*, 11(6):6–28.

11. Riem-Vis R. (2004) Cold Chain Management using an Ultra Low Power Wireless Sensor Network. In: Proc. 2004 Workshop on Applications of Mobile Embedded Systems, Boston, USA, pp. 21–23.
12. Hsieh M.Y., Huang Y.M., Chao H.C. (2007) Adaptive Security Design with Malicious Node Detection in Cluster-Based Sensor Networks. *Computer Communications*, 30(11):2385–2400.
13. Teng W.G. (2005) Information Fusion for Self-Organizing Sensor Networks. In: Proc. 5th Annual Emerging Information Technology Conference, pp. 123–125.
14. Gaurav G, Mohamed Y (2003) Fault-Tolerant Clustering of Wireless Sensor Networks. In: Proc. Wireless Communications and Networking, pp. 1579–1584.
15. Younis O., Fahmy S. (2004) Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach. *IEEE Transaction on Mobile Computing*, 3(4):366–379.
16. Munir S.A., Ren B., Jiao W., Wang B., Xie D., Ma J. (2007) Mobile Wireless Sensor Network: Architecture and Enabling Technologies for Ubiquitous Computing. In: Proc. 21st International Conference on Advanced Information Networking and Applications Workshops - AINAW'2007, pp. 113–120.
17. Kottapalli V.A., Kiremidjian A.S., Lynch J.P., Carryer E., Kenny T.W., Law K.H., Lei Y. (2003) Two-tiered wireless sensor network architecture for structural health monitoring. In: Proc. 10th Annual International Symposium on Smart Structures and Materials, USA, pp. 8–19.
18. Shah R.C., Roy S., Jain S., Brunette W. (2003) Data MULEs: Modeling a Three-Tier Architecture for Sparse Sensor Networks. In: Proc. the First IEEE International Workshop on Sensor Network Protocols and Applications, pp. 30–41.
19. Kulkarni P., Ganesan D., Shenoy P., Lu Q. (2005) SensEye: A Multi-tier Camera Sensor Network. In: Proc. 13th Annual ACM international Conference on Multimedia, USA, pp. 229–238.
20. Heintzelman W.R., Chandrakasan A., Balakrishnan H. (2000) Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In: Proc. of the 33rd Annual Hawaii International Conference on System Sciences, pp. 3005–3014.
21. Romer K., Mattern F. (2004) Design Space of Wireless Sensor Networks. In: Proc. IEEE Wireless Communications, pp. 54–61.
22. Carlos F, Pablo H., Joaquín G., Jesús A. (2007) Wireless Sensor Networks and Applications: A Survey. *International Journal of Computer Science and Network Security*, 7(3):264–273.
23. The 29 Palms Experiments: Tracking vehicles with a UAV-delivered sensor network. URL <http://www-bsac.eecs.berkeley.edu/~pister/29Palms0103/>
24. Tatiana B., Wen H., Salil K., Branko R., Neil G., Travis B., Mark R., Sanjay J. (2006) Wireless Sensor Networks for Battlefield Surveillance. URL <http://www.cse.unsw.edu.au/~tbokareva/papers/lwc.html>.
25. Dickey R., Franklin T., Harmon J., Jennings R., Zimmerer A. (2004) Nuclear, Biological, and Chemical (NBC) Communications Network. In: Proc. IEEE Systems and Information Engineering Design Symposium, pp. 49–54.
26. Baldus H., Klabunde K., Muesch G. (2004) Reliable Set-Up of Medical Body-Sensor Networks. *Lecture Notes in Computer Science*, 2920:353–363.
27. Kappler C., Riegel G. (2004) A Real-World, Simple Wireless Sensor Network for Monitoring Electrical Energy Consumption. *Lecture Notes in Computer Science*, 2920:339–352.
28. Mainwaring A., Polastre J., Szewczyk R., Culler D., Anderson J. (2002) Wireless Sensor Networks for Habitat Monitoring. In: Proc. 1st ACM International Workshop on Wireless Sensor Networks and Applications, USA, pp. 88–97.
29. Tarik A., Yucel A. (2004) Adaptive Sensing for Environment Monitoring using Wireless Sensor Networks. In: Proc. Wireless Communications and Networking Conference, pp. 2347–2352.
30. Vandenberghe W., Latré B., de Greve F., Lamont K., Moerman I., Mertens M., Avonts J., Blondia C., Impens G. (2006) A System Architecture for Wireless Building Automation. In: Proc. 15th IST Mobile & Wireless Communications, Myconos, Greece, CD-ROM Proceedings.

31. Kim H.S., Song J.H., Lee S. (2007) Energy-Efficient Traffic Scheduling in IEEE 802.15.4 for Home Automation Networks. *Consumer Electronics, IEEE Transactions on* 53(2):369–374.
32. Bhambri L.P., Jindal C., Bathla S. (2007) Future wireless Technology-Zigbee. In: *Proc: National Conference on Challenges & Opportunities in Information Technology – COIT’2007*, RIMT-IET, Mandi Gobindgarh, pp. 154–156.
33. Demirbas M., Chow K.Y., Wan C.S. (2006) INSIGHT: Internet-Sensor Integration for Habitat Monitoring. In: *Proc: International Symposium on a World of Wireless, Mobile and Multimedia Networks – WoWMoM’2006*, pp. 26–29.
34. Xu N., Rangwala S., Chintalapudi K., Ganesan D., Broad A., Govindan R., Estrin D. (2004) A Wireless Sensor Network For Structural Monitoring. In: *Proc. the 2nd International Conference on Embedded Networked Sensor Systems*, pp. 13–34.
35. USC Information Sciences Institute, Dynamic Sensor Network HomePage (1999) URL <http://dsn.east.isi.edu/>
36. School of Engineering and Applied Sciences, Harvard University, The project of monitoring volcanic eruptions with a wireless sensor network, (2006), URL <http://www.eecs.harvard.edu/~mdw/proj/volcano/>

Wireless Sensor Network Transport Layer: State of the Art

Md. Abdur Rahman, Abdulmotaleb El Saddik and Wail Gueaieb

Abstract This chapter describes the essence of a generic transport layer of a Multi-hop Wireless Sensor Network (WSN). The transport layer of the Internet handles the congestion generated due to the network traffic and the end-to-end reliability of individual packets. Similar to the Internet, many WSN applications require a congestion control mechanism to regulate the amount of traffic injected within the WSN to avoid packet loss and to guarantee end-to-end reliable packet/event delivery. WSN researchers thus argue the presence of a transport layer for WSN similar to the Internet. Because of the resource constraint nature of sensor devices, researchers however admit that an Internet-scale transport layer will indeed be a matter of challenge. Literature reveals detailed analysis of the requirements and constraints of a WSN transport layer. The advancements in microprocessor technology, high speed and large memories, high speed networks, Ultra Wide Band frequency spectrums, very efficient sensor network Operating Systems and miniaturization of many heterogeneous sensor devices, to name a few, have led to the development of many transport layer protocols. This chapter addresses the unique characteristics of a WSN transport layer, classifies the attributes that characterize different functionalities offered by a transport layer, presents the most popular transport layer protocols based on the attributes found in each protocol, and finally, points out open research issues of this domain, which need further attention to overcome the aforementioned challenges.

Md. Abdur Rahman
Multimedia Communications Research Laboratory (MCRLab), University of Ottawa, Canada,
e-mail: rahman@mcrlab.uottawa.ca

Abdulmotaleb El Saddik
Multimedia Communications Research Laboratory (MCRLab), University of Ottawa, Canada,
e-mail: abed@mcrlab.uottawa.ca

Wail Gueaieb
Machine Intelligence, Robotics, and Mechatronics (MIRAM) Laboratory, University of Ottawa,
Canada, e-mail: wgueaieb@site.uottawa.ca

1 Introduction

A wireless Sensor network is realized as a collection of sensor nodes that are capable of sensing physical phenomena, locally processing the sensed data, and finally route the raw or aggregated data to a remote base station [1, 2] (see Fig. 1). One peculiar nature of WSN is that an individual sensor node might be resource constrained but a collection of sensor nodes can sense a large area with a greater degree of accuracy and deliver useful information to a remote location. Many sensor nodes capture information that is delicate or critical in nature. Therefore, many WSN applications require a sensor-to-base station data delivery guarantee in addition to a tolerable end-to-end delay. However, as sensor nodes intend to inject the captured sensory data within WSN, it faces congestion [3]. Due to the multi-hop nature of WSN, a different degree of congestion might be felt at different points of the network. One of the main sources of this congestion is the convergent nature of traffic toward the base station, which is sometimes called ‘sink’. As the traffic progress toward the base station, the degree of congestion increases, especially the nodes around the base station. Unless the congestion is detected and an appropriate avoidance technique is adopted, a significant amount of packet loss takes place due to lack of huge buffer space for the overwhelming number of packets. This further necessitates packet retransmission and causes a significant amount of energy loss and delivery delay. This pops up the idea of incorporating a transport layer for sensor networks similar to the one found in the Internet. Many WSN transport layer protocols have already been proposed by several researchers. However, before going into detail of a transport layer of a WSN, we first present the important terminologies required to understand the functionalities and requirements of a generic transport layer followed by a basic WSN transport layer model. What follows next is the major transport layer protocols proposed to date. Finally, we enumerate a number of open research directions.

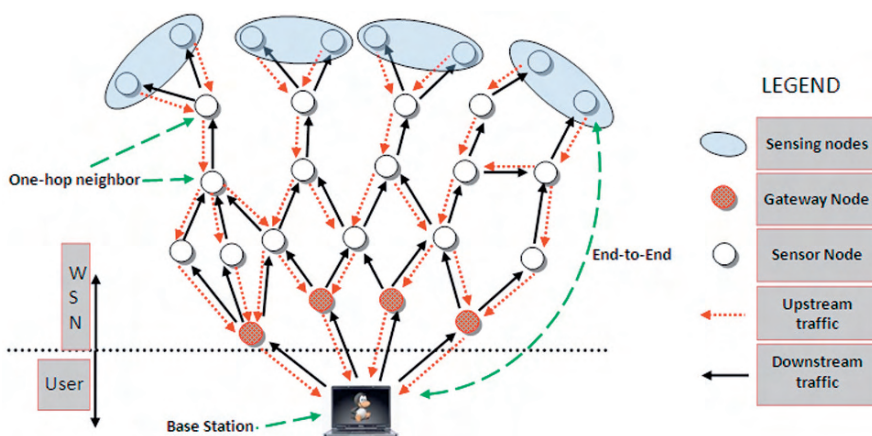


Fig. 1 Multi-hop WSN

1.1 Relevant Terminologies

1.1.1 Traffic Semantics

Sensory data flow handled by a transport layer protocol might be categorized in many ways. Based on the direction, they are named as *upstream* [3, 4, 5, 6, 7, 8, 9, 10] and *downstream* [9, 11, 12, 13, 14] sensory data traffic. When the sensory data flows from the sensing nodes to the base station, it is called upstream sensory data traffic and the reverse scenario is referred to as downstream data flow. Some literature refers to upstream data flow as *many-to-one*, *sensor-to-sink*, or *converge-cast* and to downstream data flow as *one-to-many*, *sink-to-sensor*, or *multicasting*. Another view is the traffic pattern experienced by any sensor node. The net traffic seen by any sensor node might consist of two sources [7, 15]. The first one is the sensing data captured by a sensor node itself and prepares to inject within the WSN. The second source is the neighbors of a sensor node where a sensor node simply receives the packets from the neighbors and routes them to its upstream or downstream node(s). The later type of traffic is sometimes referred to as *route-thru*, *en-route* or *transit* traffic. Another way of classifying the traffic is dense sources producing a high traffic rate, sparse sources generating low rates and sparse sources causing high rates [4]. Types of applications and the network topology also shapes the nature of traffic flowing within the network. For example, a traffic pattern might be bursty, continuous, time interval-based or query-based [6, 8, 9, 16]. Event-based applications generally produce bursty traffic. Some WSN applications need continuous delivery of captured sensory data. Some applications require timely dissemination of data and some applications want reactive responsive data from the sensor network based on the query sent.

1.1.2 Reliability Semantics

Reliability in a WSN [5, 12, 13, 16, 17] can be realized as *packet* reliability, *event* reliability, *end-to-end* reliability, *hop-by-hop* reliability, *upstream* reliability and *downstream* reliability. Another set of reliability semantics is presented in [11], where reliable data delivery can be assumed for the whole sensor network, some portions of a network, to a subset of sensor nodes so that the whole terrain to be sensed is covered, and to probabilistically covered (for example 80% of the nodes) sensor nodes. *Packet* level reliability ensures that a packet/fragment from a sensing node reliably reaches the base station. Due to a large number of redundancy, guaranteeing all the packets might not be efficient and hence, the concept of *event* reliability comes into existence, which makes sure that the captured event is reliably sent to the base station with a certain degree of accuracy. *End-to-end* reliability refers to reliable data delivery of sensed data from the source to the destination node or vice versa. To enforce *end-to-end* reliability, a sender node generally uses a *closed loop* feedback scheme where it waits for the reply message from the destination end point. On the other hand, *hop-by-hop* reliability works similarly to the MAC layer

protocols that uses an *open loop* non-feedback process. Depending on the type of application and the WSN architecture, *end-to-end* reliability might be unsuitable to offer and hence, *hop-by-hop* reliability is provided instead. A lack of any Internet-like unique addressing mechanism for WSN and frequent topology change makes *hop-by-hop* reliability an attractive option for the WSN.

Each approach has its own advantages and limitations. For example, despite their simplicity and robustness, *end-to-end* reliability approaches result in more in-network traffic. For real-time packet delivery, *hop-by-hop* packet delivery/recovery is preferable over the *end-to-end* approach [5, 12]. *Hop-by-hop* approaches offer a better solution to weaken congestion quickly with less on-going packets, but the biggest issue is that *hop-by-hop* loss recovery cannot assure message delivery in the presence of frequent network topology changes due to node relocation, addition and failure. Since less on-going packets can result in saved energy, the trade-off between end-to-end and hop-by-hop mechanisms is a design factor. *Upstream* reliability refers to the reliable delivery from sensor nodes to the base station while *downstream* reliability guarantees data delivery from the base station to all or a subset of sensor nodes. One interesting aspect of *downstream* reliability is that it requires 100% reliability guarantee without any failure. For example, if a new security patch is available for the OS, then it must be installed by all the sensor nodes without any exception. Any application might need one, a subset or all of them depending on the scenario. In order not to overwhelm the resource constrained sensor nodes, the downstream reliability mechanisms usually employ a *hop-by-hop* packet recovery using NACK (negative acknowledgement) messages while ACK (acknowledgement) messages are used for *end-to-end* packet recovery.

1.1.3 Loss Recovery

Another transport layer reliability paradigm is the mechanism of recovering the lost packets or fragments [5, 11, 12, 13, 18, 19]. In the case of *end-to-end* reliability guarantee, the loss recovery mechanism is initiated by the base station while in the case of *hop-by-hop* reliability each intermediate node can initiate the loss recovery process. This is achieved through in-network processing and caching the packets on their way to the base station. One design view suggests that all the sensor nodes between the source node and the base station caches each packet while another research view is that a selective set of sensor nodes takes part in caching including the source and the base station. The base station keeps the cache to maintain the integrity of the complete information. For the recovery of a lost segment, a transit node issues a NACK to its immediate source node and this process backtracks up to the original packet source to recover the packet from the cache.

1.1.4 Congestion Detection, Mitigation, and Control

High data rates, many-to-one network topology, huge bursts of event data and collision in the physical channel are the main source of congestion in a WSN

[3, 4, 5, 6, 7, 8, 9, 10]. The congestion pattern is different for *upstream* and *downstream* data flow. Because typical WSNs deploy a sheer number of sensor nodes and due to the large probability that many of them will be sensing events simultaneously, *upstream* data flow causes congestion. In some literature, this traffic scenario is called the *funneling effect* [10]. The degree of congestion varies from node to node as the traffic progresses toward the base station. Particularly sensor nodes closer to the base station route huge volume of packets, thereby facing huge traffic around the base station. On the other hand, *downstream* data flow typically consists of the query or program codes that are sent from the base station targeting the downstream sensor nodes to update them. Because the base station is typically assumed to have adequate processing and communication power, the base station might even reach all the sensor nodes with just one broadcast, i.e. all the sensor nodes are just one-hop away from the base station. This assumption relaxes the congestion requirement of *downstream* traffic to a certain degree. Typically the queries are sent with a pause and program codes are sent only when they are needed. Hence, downstream traffic does not typically contribute any significant congestion and as a result very little research work is concerned about the downstream congestion management.

Congestion detection protocols employ a mechanism whether or not a congestion occurred and at what location. A congestion can be detected through monitoring a sensor node's buffer and it's channel load. In some literature they are called *channel-sampling* based congestion detection and *queue-occupancy* based congestion detection respectively in which queue-occupancy based congestion detection shows superior results over its counterpart [8]. Using congestion mitigation technique, which bears similar meaning as congestion avoidance, sensor nodes limit their flow to their next-hop neighbors and help them to overcome the congestion. Several approaches of congestion mitigation are found in the literature. One of them is by overhearing the queue-occupancy state of a node's parent node, where the child node dynamically regulates its upstream or downstream data forwarding based on its parent's queue condition. Another approach is to forward the excess traffic to a secondary high speed network, if available, that has a long communication range and is able to forward the traffic to the base station quickly. Typical congestion control mechanisms adopt two popular approaches. First, dropping some packets at the congestion points by maintaining the content of the queue, and second, regulating the rate at which the sensing nodes inject the traffic within the WSN (such as Additive Increase Multiplicative Decrease (AIMD) in Transport Control Protocol (TCP) [20]).

1.1.5 Performance Metrics

In order to evaluate any transport layer protocol, several metrics have been proposed and adopted by researchers. They fall into two broad categories: reliability metrics [5, 12, 13, 16, 17] and congestion metrics [3, 4, 5, 6, 7, 8, 9, 10].

Reliability metrics are concerned with either reliable delivery to all the downstream nodes, upstream or downstream reliable delivery to sensors of a sub-region, reliable delivery to a sub-set of sensor nodes and probabilistic reliability (e.g. reliable

delivery of at least 80% of the sensor nodes of a WSN). *Event reliability* reflects how well an event is reported to the base station, which is defined by [13]

$$R(v) = \frac{\sum_{k=1}^K \text{Prob}(\text{success of } v_k)}{K} \quad (1)$$

where v is a message, K is the total number of events defined by the application, k is the event that needs to be delivered reliably and v_k is the message containing the event k . *Node reliability* for node i is defined as

$$R_n(i) = \frac{\text{number of packets of node } i \text{ received by sink}}{\text{total number of packets node } i \text{ generates}} \quad (2)$$

Following are the metrics that refer different attributes of congestion in a WSN. *Congestion degree* d is a congestion detection metric that is defined by [7, 15]

$$d(i) = t_s^i / t_a^i \quad (3)$$

where t_s^i is the mean packet servicing time and t_a^i is the mean packet inter-arrival time of node i . *Network efficiency* refers to how well the transport layer is capable of detecting congestion hot spots early, mitigating the congestion, and maintaining the required level of throughput at the base station. This takes into account the number of packets injected by the sensing nodes within a unit time and the number of them delivered to the base station. *Node efficiency* or the *average delivery ratio* [12] or *imbalance* [8] is calculated as

$$\frac{\text{number of packets received by node } i}{\text{number of packets received by } i\text{'s parent}} \quad (4)$$

Sink-received throughput measures how well a transport layer is capable of maintaining a required level of throughput at the base station. *Average delivery overhead* [12] is the total number of control messages (e.g. ACK or NACK) needed to successfully send a data packet from the source to the destination. The lower the value of this metric, the better the protocol. *Network fairness* allows each of N sensor nodes in a WSN to enjoy and experience an equal degree of network resources. Fairness ϕ is defined by [8] as

$$\phi(i) = \frac{(\sum_{i=1}^N r_i)^2}{N * \sum_{i=1}^N r_i^2} \quad (5)$$

where r_i is the packet delivery rate by node i . The *deadline miss ratio*, which is generally measured in percentage [9], is a metric that describes the efficiency of a real-time transport protocol and is defined as

$$\frac{\text{number of packets reach the sink within deadline}}{\text{total number of packets destined toward sink}} \quad (6)$$

The *packet loss ratio* is calculated as

$$\frac{\text{number of packets lost in the network}}{\text{number of packets generated by the sensing nodes}} \quad (7)$$

Opposite to *packet loss ratio* is *success rate*, which is defined as the number of packets that successfully reach the destination by bypassing the congestion. *Packet latency* is measured as the time taken by a packet to reach the destination (e.g. base station for upstream end-to-end communication or any next-hop node in the case of hop-by-hop communication) from the time the packet was created. *Average packet latency* is the median latency of packets that are observed within a time interval. Some literature defines *average packet latency* [12] as the average time elapsed between the transmission of the first packet of a message from the base station and the last packet received by the intended downstream node(s) within a time interval.

Apart from the above, some metrics characterize different attributes of a WSN transport layer. *Energy loss* per node and by the whole network are two metrics to evaluate how energy efficient a transport protocol is [16, 17]. For example, assuming dropped packets have a direct relation with energy wastage, the *energy loss* per node can be measured by [10]

$$E(i) = \frac{\text{number of packets dropped by node } i}{\text{total number of packets received by node } i} \quad (8)$$

whereas the *energy loss* by the whole network can be deduced by

$$E_{\text{network}} = \frac{\text{number of packets dropped by the network}}{\text{total number of packets received by the sink}} \quad (9)$$

The *residual energy* refers to the amount of energy remaining in a sensor node and is measured as

$$E_r = \frac{\text{remaining energy}}{\text{initial energy}} \quad (10)$$

Fidelity ratio is calculated as [10]

$$F = \frac{\text{packet throughput at the base station with protocol}}{\text{packet throughput at the base station without protocol}} \quad (11)$$

Network lifetime is the maximum time interval a WSN can operate without losing any of its functionality. *Redundancy* is a metric that measures how many times similar packets are received by a particular node within an interval that have caused unnecessary transmission, energy wastage and processing power by the node, e.g. in the case of hop-by-hop broadcasting or retransmission for loss coverage. Redundancy is elaborated in [14] and is measured by

$$\frac{(c + s)}{k} - 1 \quad (12)$$

where c is the number of times a similar packet is received, s is the number of times the received packet is further broadcasted and k portrays a threshold value that defines the tolerable redundancy.

However, there is a possibility that the above mentioned metrics exist with slightly different names because of the lack of any standard.

2 Major Transport Layer Protocols

The objective of an ideal transport layer is to govern congestion that arises from the variance of injected traffic within the network, recover packet loss due to congestion and queue overflow, to guarantee end-to-end reliability and Quality of Service (e.g. maintaining tolerable bandwidth, packet loss ratio and latency, depending on the application), and orderly delivery of packets, in case packets are fragmented at the transmitter end. In case end-to-end reliability cannot be provided due to some network constraints, a *hop-by-hop* reliability mechanism is provided instead. The MAC layer shares the responsibility of recovering packet loss due to bit error. However, it cannot recover any packet loss due to queue overflow that occurs in the case of congestion. Before deducing a model for the WSN transport layer, we present two Internet transport protocols, namely TCP (Transport Control Protocol) [20] and UDP (User Datagram Protocol) [21] that are two de-facto transport control protocols. However, due to several unique and challenging characteristics of WSN, neither of them can be directly adopted for WSN.

TCP is a connection-oriented protocol that sets up the connection (3-way handshake) between sender and receiver nodes before the actual packet communication starts. If implemented in WSN, where actual data might be only in the order of a few bytes, the 3-way handshake process will become a burden for such a small volume of data. Moreover, except for a few scenarios, a WSN is envisioned as a multi-hop wireless framework where each inter-hop link is characterized by its feeble and error-prone radio channels. Because TCP is an *end-to-end* protocol, the time to setup a TCP connection between two end nodes, that are a significant number of hops away from each other, might be very high. Therefore it is difficult for sensor nodes, especially for those that are far from the sink, to obtain enough throughput to support such WSN applications that require continuous data transmission. On top of that, the *end-to-end* approach has a longer response time in the case of congestion, which in-turn would result in a large number of segment drops. These segment drops would simply mean useless energy consumption. To guarantee reliability, TCP uses an *end-to-end* ACK and a retransmission strategy, which leads to a much lower throughput and longer transmission time.

Some unfavorable characteristics of UDP make it unsuitable for WSN, despite the connectionless paradigm it offers. UDP does not offer flow control and congestion control mechanisms. In the case of congestion, UDP simply drops packets, providing no scope of recovering the lost packets. Besides, UDP comes with no

ACK and thus only relies on the lower layer MAC algorithms or on some upper layers, including the application layer, to recover the lost packets.

On top of individual limitations, both TCP and UDP protocols do not come with inherent cross layer interaction mechanisms, especially with the lower layer protocols. However, the basic concept of TCP and UDP has shaped the subsequent design of transport layer protocols proposed to date.

Having covered the unique characteristics of WSN, we can now define the requirements of the WSN transport layer. In general, the transport layer should

- be coherent with the data flow model of the application such as event-driven, continuous, real-time and/or hybrid,
- provide congestion control and *end-to-end* reliability,
- provide upstream and downstream reliability and congestion control,
- be able to cope with the variable reliability model required by the application. For example, a temperature monitoring application might be more tolerable to packet loss than a covert military surveillance application. Some applications might need packet level reliability while others might look for event reliability,
- be fairly scalable as the sensor network density increases,
- be able to work seamlessly with other layers such as application, network and MAC layers,
- consume the least amount of energy, and
- minimize the usage of control messages without compromising the required level of data throughput.

Keeping the above sets of standards in mind, several transport control protocols have been developed for WSN [4, 5, 11, 12, 16, 22, 23]. Some of them are fairly simple while others fulfill a good number of the above requirements. As mentioned earlier, congestion control and/or reliability guarantee is an essential task of a transport control protocol, which can be offered in upstream, or downstream, or for both way traffic. As such, transport layer protocols vastly fall into three categories: upstream congestion control, upstream reliability guarantee, and downstream reliability guarantee. The major transport layer protocols proposed to date are presented in the following section.

2.1 *Congestion Detection and Avoidance(CODA)*

CODA [4] maintains an *upstream* congestion control mechanism. To do so, it introduces three schemes: congestion *detection*, open-loop *hop-by-hop* back-pressure, and closed-loop *end-to-end* multi-source regulation. CODA senses congestion by taking a look at each sensor node's buffer occupancy and wireless channel load. If they exceed a predefined threshold value, a sensor node will notify its neighbor source node(s) to decrease the sending rate through an open-loop *hop-by-hop* back-pressure. Receiving a back-pressure signal, the neighbor nodes simply decrease the packet sending rate and also replay the back-pressure continuously. CODA regulates

the multi-source rate by the closed-loop *end-to-end* approach, which works as follows. Before sending a packet, a sensor node probes the channel at a fixed interval and if it finds the channel busy more than a predefined number, it enables a control bit, called *congestion bit*, in the outgoing packet header to inform the base station that it is experiencing congestion. When the base station receives a packet with the *congestion bit* enabled, it sends back an ACK control message to the source node(s) informing them to decrease their sending rate. When the congestion is cleared, the sink actively sends an ACK control message to the source nodes to inform them to increase their data rate. CODA uses the AIMD-like mode employed in TCP protocol to regulate the data rate.

Although CODA was evaluated through both simulation and experimental implementation through a testbed, the testbed only consists of three sensing nodes and four routing nodes which does not reflect the traffic of most of the real life WSN applications. Despite its satisfactory performance, CODA shows poor congestion handling as the number of source nodes and data rate increases. It does not have any reliability mechanism, and the latency time of closed-loop multi-source regulation increases under heavy congestion.

2.2 *Event-to-Sink Reliable Transport (ESRT)*

ESRT [16, 17] aims at providing both *upstream event* reliability and congestion control while maintaining the minimum energy expenditure. ESRT can also reliably deliver multiple concurrent events to the base station. ESRT guarantees only the *end-to-end* reliable delivery of individual events, not individual packets from each sensor node. The notion of reliability is defined with respect to the number of data packets originated by any event that are reliably received at the base station. The base station node runs the ESRT algorithm to decide that the event is reliably detected at the base station or not. To do this, the base station tracks the event reporting frequency (f) of the successfully received packets originated by a particular event within a time interval and matches it with the required reliability metric. Five scenarios might occur. If the current calculated reliability at the base station falls below the required reliability and there is no congestion, ESRT increases the f abruptly. If there is no congestion and the reliability level is high, ESRT decreases f cautiously. In case congestion is detected and reliability falls, ESRT exponentially decreases the value of f . For the scenario where congestion is detected despite high reliability level, ESRT decreases reporting frequency to get rid of congestion without compromising the reliability. However, ESRT tries to operate on the optimum point where any event is reliably reported to the base station without causing congestion to the network. ESRT assumes that the base station has a high power radio and can reach all the sensor nodes in a single broadcast message. The base station broadcasts the newly calculated value of f to the whole sensor network. Upon receiving the event reporting frequency, each sensor node calculates its event reporting duration and checks at the buffer level at the end of each reporting interval to guess any possible

congestion. In case a sensor node faces congestion, it sets a congestion enable bit of the event report packet. When these packets arrive at the base station, the base station gets an overall view of the congestion level of the network. ESRT conserves energy by controlling the value of f . The evaluation of ESRT is done through analytical modeling and simulation.

However, ESRT has some performance problems. First of all, ESRT assumes that the base station is one-hop away from all the sensor nodes, which might not be applicable to many of the WSN applications. Second, ESRT floods the value of f to the whole network to override their event sensing rate, which is unfair because different portions of the network or different individual sensor nodes might face different traffic and therefore contribute different levels of congestion.

2.3 Reliable Multi-Segment Transport (RMST)

RMST [5] guarantees *upstream packet* reliability using in-network processing. It adopts a cross layer synergy by working in co-operation with the underlying routing protocol at the network layer and MAC protocol at the link layer in order to guarantee *hop-by-hop* reliability. RMST uses the term *fragmentation/reassembly*, which simply means the packets originating from a source node (called RMST entity) are fragmented and then reassembled at the base station. Fragmentation is necessary to adjust the size of the maximum transmission unit (MTU) permissible by the transit nodes. The notion of reliability adopted by RMST is the reliable delivery of fragments originating from any particular RMST entity to the base station. RMST introduces two modes of operation: cached and non-cached. In caching mode, the nodes between the source and base station cache the fragments and any RMST node can initiate recovery for missing fragments along the path toward the source. In the case of non-cached mode, only the source and the base station maintain the cache and the base station monitors the integrity of an RMST entity in terms of the received fragments. RMST uses selective NACK-based protocol to detect a fragment loss and sends NACK from the detecting RMST node to the source node. Each RMST entity receiving the NACK first looks at its cache to find out the missing segment. In the negative case, it forwards the NACK to the RMST entity down the hierarchy toward the source node. RMST is evaluated via *simulation*.

Despite the above mentioned features, RMST has some drawbacks. RMST is only suitable for those applications that need to send large size sensory data such as JPEG image that takes advantage of fragmentation at the source and reassembly at the base station. RMST might not be suitable for reliably delivering fragments from multiple RMST entities to the same base station. It cannot ensure the orderly delivery of fragments to the base station. More number of fragments will cause more contention for the channel i.e. more in-network data flow will happen. Moreover, RMST does not provide any real-time reliability guarantee or congestion control.

2.4 Pump Slowly Fetch Quickly (PSFQ)

PSFQ [12] is designed to provide *downstream* reliability where the control message from the base station is sent to the downstream sensor nodes at a relatively slow pace and allows any intermediate sensor node, which experiences packet loss to quickly recover any missing segment from immediate neighbors. This protocol is suitable for timely dissemination of code segments to a group of specific target sensor nodes for re-tasking their jobs. PSFQ employs a *hop-by-hop* error recovery mechanism in which intermediate nodes also cache fragments and share responsibility for loss detection and recovery.

It introduces three operations to maintain reliability: *pump* operation, *fetch* operation, and *report* operation. During *pump* operation, the base station slowly broadcasts a packet containing control scripts to its neighbors every T unit of time interval. The *fetch* operation is triggered as soon as a sequence number gap is found by any downstream node. In this mode, a sensor node halts its regular data routing operation to its downstream nodes and issues a NACK message to its immediate upstream neighbors to recover missing fragments. PSFQ supports the term called *loss aggregation* in which case the *fetch* operation deals with more than one packet loss. Finally, during the *report* operation, the base station makes any specific sensor(s) feedback data delivery status information to it, where the report message is designed to travel from the target node back to the base station on a *hop-by-hop* basis. PSFQ is evaluated both in terms of simulation and experimental implementation through a testbed.

However, PSFQ has several disadvantages. PSFQ cannot recover the loss of every single packet due to congestion because it uses only NACK. Both pump and the fetch operation is performed through broadcast, which might be expensive in terms of energy usage. The slow nature of pump operation in PSFQ results in large delay. PSFQ does not allow any out-of-order delivery of packets, which poses a greater challenge on cache management by the intermediate nodes. It is only intended for re-tasking the sensor node applications and thus might not be suitable for upstream data reliability. It does not provide a congestion control mechanism.

2.5 GARUDA

Similar to PSFQ, GARUDA [11] is a framework designed to provide *downstream* reliable delivery of control codes, and query-metadata. The control code is used to re-program sensor nodes, for example, new algorithms for human face recognition for the image sensors. Query-metadata is sent to the sensor nodes to store it so that, later on, when the actual query is made, the sensor nodes can match it with this metadata. For example, for a remote surveillance application where mobile robots are deployed to automatically detect human body and then look for a wanted face, base station can send a particular image containing the face of a person that needs to be tracked by the mobile robots. Individual sensors can save the image and when

capturing the face image from the sensed environment, each mobile robot can match the captured facial image and respond accordingly. GARUDA makes *core* nodes to cache the packets and *non-core* nodes recover any lost packet from them.

It uses three phases to propagate the data reliably. During the first phase, GARUDA uses a technique called Wait-for First Packet (*WFP*) pulse broadcasting from the base station to guarantee the delivery of the first packet of a message to all the sensor nodes. This is used to construct the core nodes and any loss in *WFP* pulses is recovered by *NACK*. During the second phase, GARUDA elects the core sensors based on sensors with $\text{HopCount}=3*i$ where i is a positive integer and the upper bound of i depends on the number of hops in between the sink and downstream leaf node(s). The third phase is initiated by a two-phase loss recovery approach: initially all the *core* nodes gather the lost fragments on the way back to the base station while the *non-core* sensors recover their lost fragments from the core sensors using out-of-sequence *NACK*. This approach ensures the highest availability of the lost fragments to the non-core sensors and the least channel contention and congestion during each phase.

However, the approach followed by GARUDA might not be suitable for upstream data reliability. In case of a very large WSN, the core construction and loss recovery might be very lengthy. GARUDA only offers reliable transfer of the very first packet without guaranteeing the rest of the packets of a particular message. It does not provide any congestion control mechanism and is evaluated through simulation, not on experimental implementation through a testbed.

2.6 Tiny TCP/IP

Tiny TCP/IP proposed in [18, 24] tends to modify the TCP/IP protocol suite to make it viable for WSN and provides reliability that is a blend of *end-to-end* and *hop-by-hop* reliability. The protocol assumes that each sensor node knows its spatial location a priori and falls into any of the pre-defined subnets. Each sensor node obtains the first two octets from the subnet and calculates the last two octets based on its spatial location within the subnet. It proposes four modifications of the existing TCP/IP protocols: spatial IP address assignment, shared context header compression, application overlay routing, and distributed TCP caching (DTC). Sensor nodes of the same IP subnet do not need to transmit a full IP header. Hence, the IP header can be compressed and shared among the sensor nodes of the same subnet. Local IP broadcasting of UDP datagrams is used to form an application layer overlay network on top of the physical sensor network. Finally, DTC provides the packet reliability using a distributed approach. Tiny TCP/IP proposes a novel idea of TCP packet caching within the in-network sensor nodes to minimize the burden of the *end-to-end* retransmission of fragments in case packet loss occurs. Figure 2 shows the packet loss recovery process where intermediate nodes 5 and 7 cache packets 1 and 2 respectively and hence, in case both the packets are lost, node 5 supplies packet 1 while node 7 retransmits packet 2 to the receiver. In the worst case, a

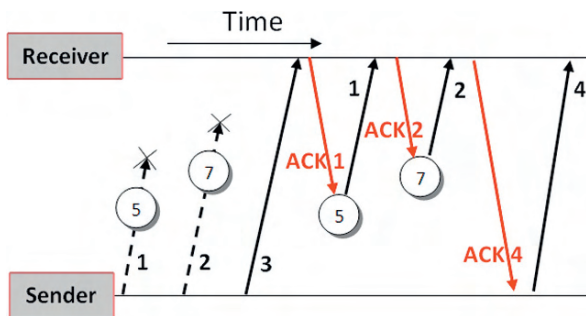


Fig. 2 Distributed TCP caching used in Tiny TCP/IP

receiver fetches the lost packet from the sender if the lost packet is not cached by any intermediate node. The protocol is evaluated through both simulation and an actual WSN.

The proposed protocol seems to have some performance issues. The assumption of static spatial subnet IP makes is unsuitable for many of the mobility-supported WSN applications. The protocol's reliability performance depends on the efficiency of caching the last seen packets. Which node will cache which packets thus makes a complex design issue of this protocol. The protocol does not explicitly define any congestion control mechanism. Finally, it does not explicitly address the design challenges of *upstream* or *downstream* reliability.

2.7 Sensor TCP(STCP)

STCP [6] is a generic *end-to-end upstream* transport protocol for a wide variety of WSN applications. STCP provides both congestion *detection* and *avoidance* and a variable degree of *reliability* based on the application requirement. STCP uses three types of packets: session initiation, data, and ACK. The session initiation packet is meant to synchronize any sensor node with the base station, which constitutes a number of flows (assuming each sensor node is capable of originating multiple flows such as *event-driven*, *continuous* etc.) originating from it, type of data flow, transmission rate and required reliability. STCP data packets take an important role in maintaining the congestion information. STCP employs most of its functionalities at the base station. The base station uses NACK for applications requiring continuous *end-to-end* sensory data flow and hence, clock synchronization is maintained between the base station and the source nodes. In the case of event-driven sensory data flow applications, source nodes use ACK to make sure that the base station has successfully received the packets. Each packet is kept in the source node's cache until it gets an ACK from the base station. Intermediate nodes detect congestion based on queue length and notify the base station by setting a bit in the data packet headers. STCP was evaluated through *simulation*.

STCP assumes that all the sensor nodes within the WSN have strict clock synchronization with the base station, which might be a cause of performance problem with STCP. Sensing nodes waiting for the ACK reply from the base station will cause long latency in large scale multi-hop WSN.

2.8 *SenTCP*

SenTCP [23] is an open-loop *hop-by-hop* congestion control protocol intended for *upstream* traffic flow. SenTCP measures the degree of congestion in every intermediate sensor node by taking a look at the average local packet-servicing time, local packet inter-arrival time, and the buffer occupancy. In the face of congestion, SenTCP makes each intermediate sensor issue a feedback signal to its neighbors, which carries the local congestion degree and the buffer occupancy ratio. Finally, SenTCP employs a mechanism to process the received feedback signal to adjust the local data sending rate. This use of *hop-by-hop* feedback control regulates the congestion quickly and reduces packet dropping, which in turn conserves energy and increases the throughput.

Nevertheless, SenTCP only provides congestion control without any loss recovery and does not guarantee reliability. Until now, the efficiency of SenTCP is tested via *simulation* and hence, its suitability need to be verified by implementing on a physical testbed.

2.9 *Trickle*

Trickle [14] facilitates WSN reprogramming by providing *downstream* nodes to intelligently infer any new code availability and subsequently pushing the actual code in a *hop-by-hop* fashion. Trickle uses the concept of *polite gossip* to propagate metadata regarding any updated code that needs to be pushed downstream. Trickle focuses on metadata propagation rather than an actual code propagation inside the network. When a sensor node detects any older metadata from its neighbors, it updates its neighbors by broadcasting the appropriate code. Conversely, if any sensor node receives any newer metadata from its neighbors, it broadcasts its own metadata, which in turn makes the receiving sensor node with the new code to broadcast that code. Trickle is evaluated through simulation and an experimental implementation through a testbed. The empirical results show that Trickle imposes an overhead of 3 packets/hour and can reprogram the entire network in 30 seconds. Although Trickle guarantees the delivery of metadata about the code, it does not guarantee reliable delivery of the code itself. Also, trickle does not provide any mechanism of querying the current code version from any one or a set of sensor nodes. This makes the base station unaware of the current status of the WSN.

2.10 FUSION

FUSION [8] provides an *upstream* congestion control mechanism that fuses three techniques: *hop-by-hop* flow control, rate limiting of source traffic in the transit sensor nodes to provide *fairness* and a prioritized MAC protocol. Using the first technique, i.e. *hop-by-hop* flow control, a sensor node performs congestion *detection* and congestion *mitigation*. Congestion is detected through both *queue-occupancy* and *channel sampling* techniques. A node signals local congestion to its neighbors by setting a congestion bit in the header of every outgoing packet. It discourages any sensor node to send to such a neighbor who is already over running its queue. The second technique tries to maintain the *fairness* of allocating resources in the en-route sensor nodes so that a packet traversing a handsome amount of hops gets proper treatment. The third technique is to help a sensor node under congestion to drain its output queue to ameliorate the congestion by allocating prioritized access to the physical channel. It works as follows: a sensor node experiencing congestion makes its back-off window one-fourth the size of a normal sensor's back-off window, so that the sensor node experiencing congestion has more probability of winning the contention race. The efficiency of FUSION was tested with a physical WSN testbed composed of 55 sensor nodes. The congestion handling capacity of FUSION was tested for both *event-based* and *periodic* data traffic.

Frequenting the wireless radio for channel probing is a source of energy wastage. FUSION lacks any packet recovery mechanism and hence, does not provide any reliability measure.

2.11 Asymmetric and Reliable Transport (ART)

ART [13] provides *upstream end-to-end event* reliability, *downstream end-to-end* query reliability and *upstream* congestion control. ART selects a subset of sensor nodes called essential nodes (*E-nodes*) that can cover the whole area to be sensed in an energy efficient way. ART forms a sub-network consisting of those *E-nodes* and only those *E-nodes* take part in reliable data transfer to the upstream and downstream nodes and lost fragment recovery. ART offers four attractive features. First, non-essential nodes do not face *end-to-end* communication overhead. Second, congestion control mechanisms can be decentralized to regulate the traffic flow efficiently. Third, less number of nodes take part in lost message recovery and finally, ART uses distributed energy aware congestion control. Because ART provides a reliability guarantee in both downstream and upstream, it used both ACK and NACK mechanisms. For reliable *query* propagation, it adopts two measures. The first measure is connectionless and reactive where the base station simply sends the query fragments without worrying about any loss. It is the responsibility of the receiving *E-nodes* to detect a query fragment loss by taking a look at the sequence order and as a recovery measure sends back a NACK to the base station. The second measure resembles connection oriented communication where the base station pro-actively handles the loss detection using the time out mechanism. A timeout event without

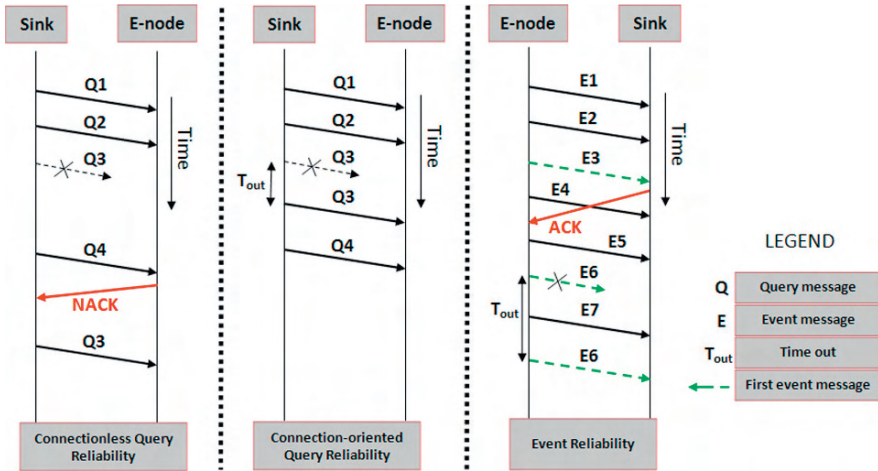


Fig. 3 ART query and event reliability mechanism: left-connection less NACK-based query loss detection; middle:ACK-based query loss detection; and right-ACK-based event loss detection mechanism

getting any ACK for a particular query fragment makes the base station resend the fragment again. End-to-end event reliability is assumed to be achieved if the first message containing the event information, which is sent by the *E*-nodes, is reliably received by the base station. For event reliability, *E*-nodes are responsible for detecting the event-message loss and to recover it. Each *E*-node enables a control bit to notify the base station that this message portrays the first event-message, which enforces the base station to reply back with an ACK. Figure 3 portrays the query and event reliability mechanism. Congestion control is handled by the *E*-nodes and the presence of congestion is assumed if a timeout happens without receiving any ACK from the base station. In this case, the *E*-node persuades its neighboring non-essential nodes to restrain from sending any data until the congestion is cleared. The performance of ART is evaluated through simulation.

However, the view of detecting congestion by lost ACKs is not an efficient solution as the ACK can be lost due to many reasons including link loss. ART does not explicitly mention its suitability of upstream data communication patterns other than event-based scenarios. Because congestion control and the two-way reliability mechanism is maintained by only *E*-nodes, any packet loss due to congestion at non-essential nodes will go unnoticed and their recovery is not guaranteed.

2.12 Congestion Control and Fairness (CCF)

CCF [25] provides *hop-by-hop upstream* congestion control using a many-to-one distributed and scalable algorithm that not only eliminates congestion but also ensures the fair delivery of packets to the base station. It allows the same number

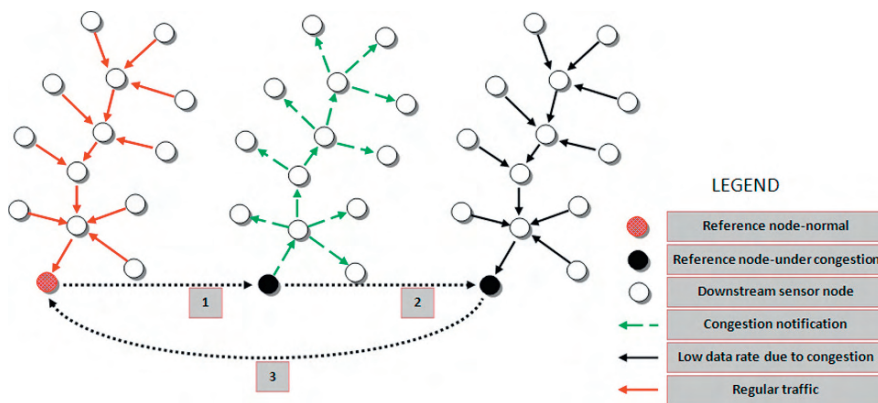


Fig. 4 CCF: 1- a node under regular traffic facing congestion informs the downstream nodes to limit their data flow; 2-constrained data flow under congestion; 3-congestion is cleared and all the downstream nodes resume their regular data flow

of data packets from sensor nodes that are many hops away as well as from sensor nodes a few hops away. CCF formulates the congestion control of any particular sensor node by calculating the number of its downstream nodes, average rate at which packets can be sent by it, per-node data packet generation rate by its parent, and the downstream propagation rate. In order to provide fairness for each child, CCF proposes two concepts: per-child packet queues, i.e. each sensor node maintains one indexed queue for each of its children along with its own queue, and per-child subtree size. CCF incorporates control information within the data packets, thus eliminating the injection of additional packets to the downstream sensor nodes. When any sensor node experiences congestion, it informs the downstream nodes to reduce their data transmission rate and vice versa (see Fig. 4).

Since CCF implements the congestion control algorithm in the transport layer, it is independent of the underlying network and MAC layers. CCF is evaluated through simulation and in a real WSN environment. However, reserving equal resources for each sensor node to provide equal opportunity might be inefficient for many scenarios. For example, some sensor nodes might be capturing events more often than others such as a video sensor capturing 10 frames per second needs a higher bandwidth and channel access than that required by a static sensor. Another shortcoming of CCF is that it does not provide any reliability mechanism.

2.13 Priority-based Congestion Control Protocol (PCCP)

Wang et al. [7, 15] proposes a *hop-by-hop* node priority-based *upstream* congestion control protocol for WSN. PCCP refutes the congestion control protocols that argue in favor of providing equal fairness (e.g. CCF) to each sensor node in a multi-hop

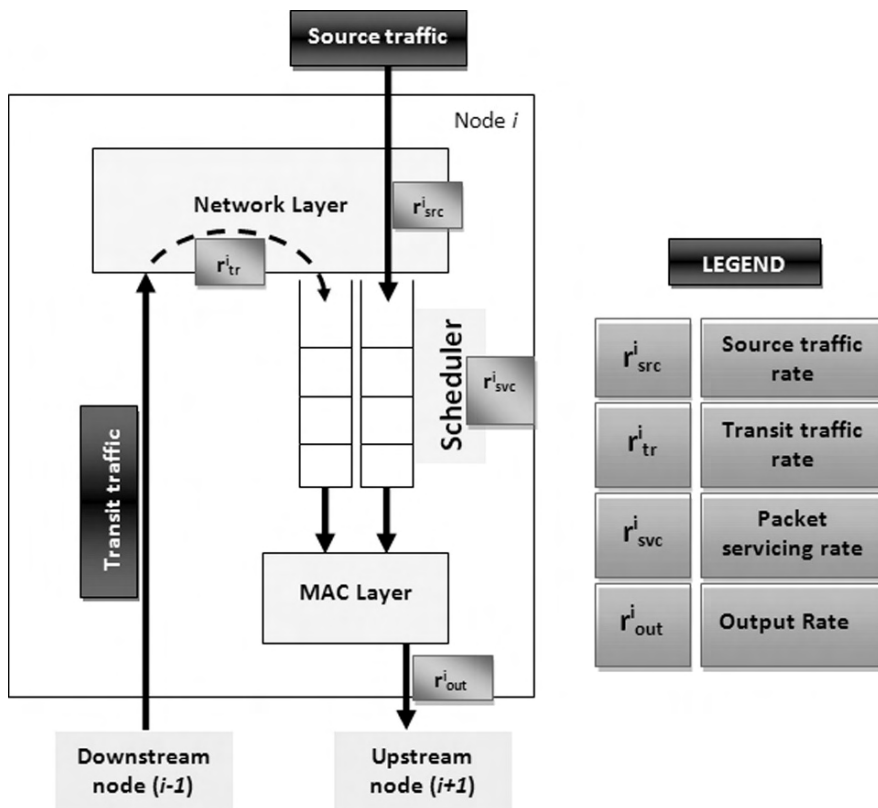


Fig. 5 Scheduler proposed by PCCP for congestion control

WSN by attaching a weighted fairness to each sensor node. PCCP offers a different degree of priority indexes such that a sensor node with a higher priority index enjoys a higher bandwidth and also sensor nodes that inject more traffic get more bandwidth. PCCP further defines the priority index for both self generating traffic and transit traffic, based on which the queue length for source and the transit traffic is allocated (see Fig. 5). PCCP infers the degree of congestion through packet inter-arrival time and packet service time and then imposes *hop-by-hop* congestion control depending on the measured congestion degree and the priority index. PCCP uses implicit congestion notification by piggybacking the congestion information in the header of data packets, thus avoiding additional control packets. PCCP allows the application layer to dynamically override the priority index of any sensor node(s) of any particular region. This feature might be required by many applications of WSN. PCCP has been evaluated through *simulation*. Despite many attractive features of PCCP, it has some limitations. PCCP does not come with a packet loss recovery mechanism. Also, PCCP lacks the notion of reliability guarantee.

2.14 Siphon

Siphon is an *upstream* congestion control protocol that aims at maintaining application fidelity, congestion detection, and congestion avoidance by introducing some virtual sinks (VS) with a longer range (IEEE 802.11 Wi-Fi) multi-radio (such as Stargate [26]) within the sensor network [10]. VSs can be distributed dynamically so that they can tunnel traffic events from regions of the sensor field that are beginning to show signs of a high traffic load. At the point of congestion, these VSs divert the extra traffic through them to maintain the required throughput at the base station. The siphon algorithm mainly aims at addressing the VS discovery, operating scope control, congestion detection, traffic redirection, and congestion avoidance. The VS discovery works as follows: the physical sink sends out a control packet periodically with a signature byte embedded in it. The signature byte contains the hop count of the sensor nodes that should use any particular VS. Each ordinary sensor node maintains a list of neighbors through which it can reach its parent VS. Finally each VS maintains a list of its neighbor VSs. Each VS has a dual radio interface: a long range one to communicate with other VSs or with a physical sink (if applicable), and a regular low-power radio to communicate with the regular sensor nodes. In the case of congestion, a sensor node enables the *redirection bit* in its header and forwards the packet to its nearest VS. When the VS finds the redirection bit enabled, it routes the packets using its own long range communication network toward the physical sink, bypassing the underlying sensor network routing protocols. Siphon uses a combination of *hop-by-hop* and *end-to-end* congestion control depending on the location of congestion. If there is no congestion, it uses *hop-by-hop* data delivery model. In case of congestion, it uses *hop-by-hop* data delivery model between source nodes and the VS at point of congestion and an *end-to-end* approach between the VS handling the congestion and the physical sink. Siphon has been evaluated using packet-level simulation and experimental implementation through a testbed.

The optimality of Siphon is dependent on the optimality of the number of VSs. Although Siphon addresses the congestion detection and avoidance mechanism, it does not have any packet recovery mechanism due to congestion. Also, Siphon does not address the reliability issue of the transport layer.

2.15 Reliable Bursty Convergecast (RBC)

RBC [19] is suitable for *event-driven* bursty *upstream* traffic and provides real-time packet reliability through *hop-by-hop* loss recovery. Every sensor node maintains a priority queue and makes two assumptions for the real-time packet transport: first, a sensor node is capable of guessing whether its neighbor has received and forwarded its supplied packet or not by listening to the channel, and second, the sensor nodes maintain precise time synchronization. The above two assumptions waive the transport layer from maintaining in-sequence packets. RBC uses a windowless block Acknowledgment [26] (see Fig. 6) scheme to reduce the packet and ACK

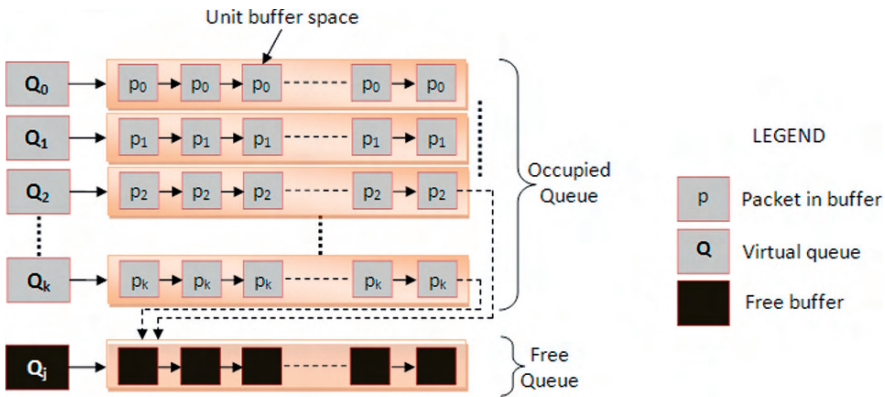


Fig. 6 A virtual queuing model used in RBC to support uninterrupted traffic flow through a sensor node

loss and employs differentiated contention control to rank the sensor nodes based on the queuing condition. Among the neighboring nodes, the node with the highest rank accesses the channel first. This improves the contention scenario, which in turn improves overall congestion. As shown in the figure, Q_0 is the virtual queue that holds packets that have the highest priority of channel access where packets within Q_0 are serviced with First In First Out rule. In fact, packets in the Q_0 buffer are sent zero (0) times and the probability that the packets kept in Q_0 have been received by the receiver is also zero (0). Virtual queues Q_0, Q_1, \dots, Q_k are ranked such that the rank of Q_k is higher than Q_j if $k < j$. Packets are ranked based on the ranks of virtual queues and are stored accordingly. The occupied buffer space hold the packets that need to be sent or to be acknowledged and the free virtual queue is to hold the newly arrived packets. RBC offers a set of rules of when to remove a packet from the buffer and how to handle incoming packets while facing congestion. RBC was evaluated by experimenting with an outdoor sensor network testbed.

RBC has several challenges to be worked out. It will be interesting to see how efficient the proposed real-time *hop-by-hop* packet reliability works for continuous data flow. The lack of energy model makes it unsuitable for many sensor network applications that needs strict energy optimization. Another challenge will be to extend the queue model for multimedia sensory data communication.

2.16 RAP

RAP [9] provides a real-time *upstream* and *downstream* communication protocol for large scale WSN. RAP supports both *periodic* and *event-based* data flow. The protocol has a *downstream* query propagation service and a suite of cross layer network services for *upstream* query-result delivery to the base station. Cross layer

includes a transport layer location address protocol (LAP), location-based routing protocol, a velocity monotonic packet scheduling layer and a prioritized MAC layer. The cross layer helps in reducing the *end-to-end* deadline miss ratio by most of the packets. RAP provides fairness by giving higher priority to the packets originating from sources deeper in the network than packets originating from sources close to the base station. LAP is a connectionless transport protocol similar to UDP. LAP only provides an *upstream* unicast communication between each sensor node and the base station. Each sensor node maintains a priority FIFO queue and inserts an incoming packet, which also carries a required velocity and priority, to the appropriate queue. The packets that have missed the deadline already are being dropped from the queue by any sensor node to conserve bandwidth and buffer space.

RAP has several shortcomings. Until now, the evaluation of this protocol is presented through simulation. RAP does not provide any packet loss recovery mechanism for the dropped packets.

Figures 7, 8, and 9 summarizes the aforementioned protocols based on the classification made earlier.

Congestion Control						
Protocol	Upstream	Congestion Detection	Congestion Avoidance	Loss Recovery	End-to-End	Hop-by-Hop
ART	Y	Y	Y	Y	Y	
CCF	Y	Y	Y	Y		Y
CODA	Y	Y	Y		Y	Y
ESRT	Y	Y	Y	Y	Y	
FUSION	Y	Y	Y			Y
PCCP	Y	Y	Y			Y
RAP	Y	Y	Y		Y	Y
SenTCP	Y	Y	Y			Y
Siphon	Y	Y	Y			*
STCP	Y	Y	Y		Y	
Reliability						
Protocol	Upstream	Downstream	Type	Loss Recovery	End-to-End	Hop-by-Hop
ART	Y	Y	Event/Query	Y	Y	
ESRT	Y		Event		Y	
GARUDA		Y	Code/Packet	Y		Y
PSFQ		Y	Packet	Y		Y
RBC	Y		Event/Packet	Y		Y
RMST	Y		Packet	Y		Y
STCP	Y		Packet/Event		Y	
Tiny TCP/IP	Y		Packet	Y		*
Trickle		Y	metadata	Y		Y

* Uses a combination of both *end-to-end* and *hop-by-hop* congestion control/reliability

Fig. 7 Classification of the protocols presented in this chapter based on congestion control and reliability

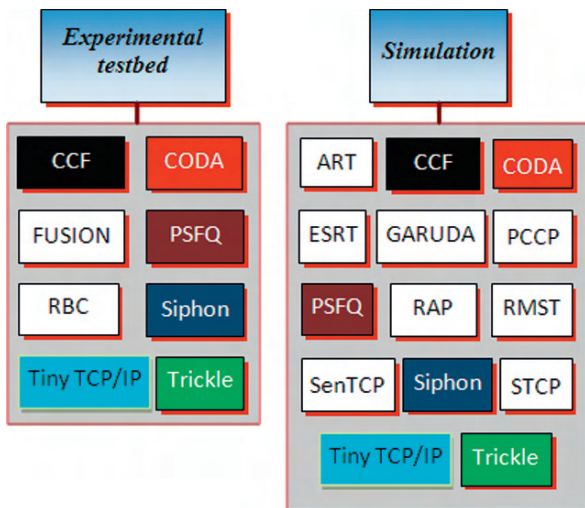


Fig. 8 Classification based on evaluation type

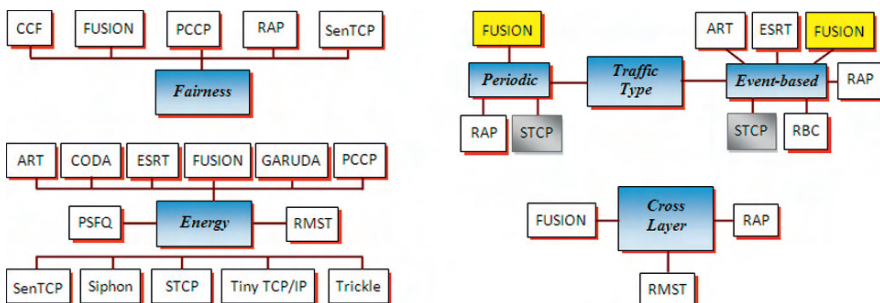


Fig. 9 Some miscellaneous classification

3 Conclusion

WSN is evolving in a rapid manner and as a result new dimension of applications are emerging. This requires the transport layer to be adaptive and be ready to take on the new challenges. Many aspects of WSN that influence the underlying transport layer protocol heavily are still unexplored by researchers. So far, among the protocols presented in this book chapter, only STCP, ART and ESRT provide both congestion control (detection and avoidance) mechanisms and reliability guarantee. Lost packet recovery is addressed by few of them. Practical deployment of any transport protocol in a large scale WSN is very challenging until now. Although some of the protocols are tested with a physical testbed, many of the protocols are only evaluated via simulation only.

One particular application worthy of mentioning is multimedia over WSN. With recent advancements in hardware and software technologies many multimedia

sensors capable of capturing audio, video, and still images are now being deployed. Multimedia sensors need a different quality of service from the underlying transport protocol, which is mostly unexplored by the transport protocols developed to date. For example, a video sensor capturing 10 frames per second needs a different processor capability, bandwidth and buffer space in each of the en-route nodes and needs to be delivered with certain jitter and end-to-end delay. This imposes several challenges including a new retransmission strategy, an enhanced buffer management technique, and redefining priorities for channel access.

A transport layer aimed at maintaining a constant data rate at the base station is another challenging research area for a multi-hop WSN. At the time of congestion, packets are dropped by the sensor nodes, which requires the retransmission of packets from the source or from its parents to maintain a satisfactory level of reliability. This generates latency and jitter at the base station. This issue is particularly acute for the multimedia applications where each media might have its own QoS metrics. For example, if a WSN is composed of many heterogeneous multimedia sensors then the transport layer has to provide different loss recovery, speed, and reliability to different sensory data streams. We believe that this problem will be of the center of future research.

A mainstream of research focuses on the independence of the transport layer protocols from underlying routing or MAC layer protocols to make it generic. This is looked at differently by others who stress the importance of cross layer synergy. For example, the transport layer can work with the MAC layer to regulate the back-off window size to prioritize the channel access of the node under congestion. Another case might be utilizing the network layer optimization. For example, underlying routing protocol might be chosen such that it uses the shortest path or multi-path to the destination or network layer header to re-route the traffic from the point of congestion. If the routing protocol notices any route failure due to one of the en-route nodes being unavailable, it can inform the transport layer that the packet loss is due to route failure and not due to congestion in order to prohibit packet loss recovery immediately. In summary, cross layer optimization will open many new research directions in the near future.

References

1. Rahman Md. A, Miah S, Gueaieb W, El Saddik A (2007) SENORA: A P2P Service oriented framework for collaborative multi-robot sensor network. *IEEE Sensors Journal*, Special Issue on Intelligent Sensors, 7(5):658–666
2. Mukhopadhyay S C, Gupta G S (2007) Sensors and robotic environment for care of the elderly. In: *Proc. IEEE International Workshop on Robotic and Sensors Environments (ROSE'07)*, Ottawa, Canada, pp. 68–73
3. Ghasemaghaei R, Rahman ASMM, Rahman M A, Gueaieb W, El Saddik A (2008) Ant colony-based many-to-one sensory data routing in wireless sensor networks. In: *Proc. ACS/IEEE International Workshop on Wireless Internet Services (WISE'08)*, Doha, Qatar, April 1–4
4. Wan C -Y, Eisenman S B, Campbell A T (2003) CODA: Congestion detection and avoidance in sensor networks. In: *Proc. the First International Conference on Embedded Networked Sensor Systems (SenSys'03)*, Los Angeles, CA, USA, pp. 266–279

5. Stann F, Heidemann J (2003) RMST: reliable data transport in sensor networks. In: First IEEE International Workshop on Sensor Network Protocols and Applications, Anchorage, AK, USA, pp. 102–112
6. Iyer Y G, Gandham S, Venkatesan S (2005) STCP: A generic transport layer protocol for wireless sensor networks. In: Proc. IEEE ICCCN, San Diego, CA, USA
7. Wang C, Sohrawy K, Lawrence V, Li B, Hu Y (2006) Priority-based congestion control in wireless sensor networks. In: Proc. IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pp. 22–31
8. Hull B, Jamieson K, Balakrishnan H (2004) Mitigating congestion in wireless sensor networks, In: Proc. ACM Sensys 04, Baltimore, MD, USA
9. Lu C, Blum B, Abdelzaher T, Stankovic J, He T (2002) RAP: A Real-Time Communication Architecture for Large-Scale Wireless Sensor Networks, In: Proc. IEEE RTAS
10. Wan C -Y, Eisenman S B, Campbell A T, Crowcroft J (2005) Siphon: over-load traffic management using multi-radio virtual sinks in sensor networks. In: Proc. ACM SenSys 05, San Diego, California, USA, pp. 116–129
11. Park, S -J, Vedantham R, Sivakumar R, Akyildiz I F (2004) A scalable approach for reliable downstream data delivery in wireless sensor networks. In: Proc. International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), Tokyo, Japan, pp. 78–89
12. Wan C -Y, Campbell A T, Krishnamurthy L (2002) PSFQ: A reliable transport protocol for wireless sensor networks. In: Proc. ACM International Workshop on Wireless Sensor Networks and Applications, Atlanta, GA, USA, pp. 1–11
13. Tezcan N, Wang W (2007) ART: an asymmetric and reliable transport mechanism for wireless sensor networks. *International Journal of Sensor Networks*, 2(3-4):188–200
14. Levis P, Patel N, Culler D, Shenker S (2004) Trickle: A self regulating algorithm for code propagation and maintenance in wireless sensor networks. In: Proc. First Symposium Networked Sys. Design and Implementation (NSDI)
15. Wang C, Li B, Sohrawy K, Daneshmand M, Hu Y (2007) Upstream congestion control in wireless sensor networks through cross-layer optimization. *IEEE Journal on Selected Areas in Communications*, 25(4):786–795
16. Sankarasubramaniam Y, Akan O B, Akyildiz I F (2003) ESRT: event-to-sink reliable transport in wireless sensor networks. In: Proc. 4th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc '03), New York, NY, USA, pp. 177–188
17. Akan O, Akyildiz I (2005) Event-to-sink reliable transport in wireless sensor networks. *IEEE/ACM Transactions on Networking*, 13(5):1003–1016
18. Dunkels A, Voigt A T, Alonso J, Ritter H, Schiller J (2004) Connecting wireless sensor networks with TCP/IP networks. In: Proc. Second International Conference on Wired/Wireless Internet Communications (WWIC2004). Frankfurt, Germany: Springer Verlag
19. Zhangy H, Aroray A, Choiz Y R, Goudaz M G (2005) Reliable bursty convergecast in wireless sensor networks. In: Proc. ACM MOBIHOC 05, Urbana-Champaign, IL, USA
20. Postel J (1981) Transmission control protocol. Information Sciences Institute, Tech. Rep., 1981, RFC-793.
21. Postel J (1980) User datagram protocol. Information Sciences Institute, Tech. Rep., RFC-768.
22. Sundaresan K, Anantharaman V, Hsieh H -Y, Sivakumar A (2005) ATP: a reliable transport protocol for ad hoc networks. *IEEE Transactions on Mobile Computing*, 4(6):588–603
23. Wang C, Sohrawy K, Li B (2005) SenTCP: A hop-by-hop congestion control protocol for wireless sensor networks. In: Proc. IEEE INFOCOM, Miami, USA
24. Dunkels A, Alonso J, Voight T (2004) Making TCP/IP Viable for wireless sensor networks. In: european workshop on wireless sensor networks (EWSN), Berlin, Germany
25. Ee C -T, Bajcsy R (2004) Congestion control and fairness for many-to-one routing in Sensor networks. In: Proc. ACM Sensys 04, Baltimore, MD, USA, pp. 148–161
26. [Online]. Available: Stargate datasheet: <http://www.xbow.com/Products/productsdetails.aspx?sid=85>
27. Brown M G G, Miller R (1989) Block acknowledgment: Redesigning the window protocol. In: Proc. ACM SIGCOMM, pp. 128–134

Part IV
Sensors for Tracking and Navigation

Real Time Tracking and Monitoring of Human Behavior in an Indoor Environment

Maki K. Habib

Abstract This chapter reports the development of a real time 3D sensor system and a new concept based on space decomposition by encoding its operational space using limited number of laser spots. The sensor system uses the richness and the strength of the vision while reducing the data-load and computational cost. The chapter presents the development and implementation of an intelligent 3D Fiber Grating (FG) based vision-system that can monitor and track human being status in real time for monitoring purposes to support wide range of applications. The 3D visual sensor is able to measure three-dimensional information with respect to human, objects and surrounding environment. The sensor system consists of a CCD camera, a laser spot array generator (constitutes: laser diode and driver, lens, fiber gratings and holder), and a processing unit with alarm facilities and interfacing capabilities to a higher-level controller and decision-making along with a user-friendly interface. The system works by projecting a two-dimensional matrix of laser spots generated through two perpendicularly overlaid layers of FGs. Then, the spot array generator projects the laser spot array on the front scene within the active view of the CCD camera and the reflected laser spots from the scene play an essential role in detecting and tracking targets. It is possible to adjust or translate the position of the laser spot generator with respect to the CCD camera to have better and balanced resolution. In addition, it is possible to rotate the FG in order to obtain a better laser spot pattern that can facilitate processing and enhance accuracy. Furthermore, multi-laser spot generators can be configured with one CCD camera to widen the operational coverage of the developed sensor system. This chapter introduces and illustrates the structure of the developed sensor system, its operational principles, performance analysis and experimental results.

Maki K. Habib

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan, Currently with the American University in Cairo, Egypt, e-mail: maki@ieee.org

1 Introduction

Intelligent monitoring and tracking systems must have significant sensing capabilities and should be flexible. This includes reasoning about sensing data [1]. They must be able to track moving and non-moving targets as a function of time, to report their status and approach a proper decision accordingly [2, 3, 4]. Automation realized in many fields by using robots. Robots need visual information to recognize objects in the working environment [5, 6, 7]. Visual information for mobile robots is most important to sense and recognize the environment during movement and to provide intelligence. Many approaches in practice use visual recognition based on two-dimensional data. The development of a three-dimensional visual approach that is fast and reliable has been sought for long time [6, 7, 8, 9, 10]. Therefore, it is necessary to have an efficient sensor system (hardware and software structure) that can fulfill these tasks properly and in real time.

Sensors for motion-tracking applications such as detecting the existence of a person or an object, locate, track, and decide their status have been an active area of research and development [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. Among the popular sensors that have been actively studied and used for this purpose include: laser range finders, sonar sensors, infrared sensors and vision sensors. Systems using vision sensors have been actively studied in many fields. Vision sensors are rich source of information that can support wide range of applications beyond object/person detection and tracking. Vision systems are capable of advanced positioning and control and offer high resolution and details that benefit tracking, identification, activity analysis, and mood beside numerous advantages for achieving such tasks. However, they have their own drawbacks, such as, high data-load, intensive analysis, algorithm complexity, heavy computation time, and at high cost. In addition, depth images are noisy and unreliable in area of low visual texture. One of the traditional approaches in mobile robot research has been the use of stereo-vision systems to extract range information from pairs of images. Applying such approach time real time navigation in a dynamic environment faces the difficulty of the intrinsic computational expense of extracting three-dimensional 3D information from stereo pairs of images and this leads to limits the number of points that can be tracked.

In order to compensate such drawbacks, researchers have been developing other alternative approaches to detect, track and monitor human while aiming to reduce the computational requirements and cost, examples of such approaches are, the use of acoustic localization [24] and infrared tracking sensors [22, 23]. The alternative approaches to vision lack the ability to present the required details that can describe the features of the source due to their limited resolution in general while some are bounded by their physical sensing properties. Tracking systems based on infrared motion sensors achieve low data-loads but in general have poor spatial resolution [23]. However, the alternative approaches lack to present the required details that can describe the features of the source due to their limited resolution in general while some bounded by their physical sensing properties. As the technology is evolving with better features, compactness and cut in price, there is a need to look

into the possible consideration of improving the way to utilize the strength of vision for real time tracking and monitoring.

As the technology of vision sensors is evolving with better technical features, compactness and cut in price, there is a need to look into a better way to utilize the strength of vision to facilitate real time obstacle detection and tracking, along with other range of applications. This article tries to contribute to this by presenting the development of 3D based active vision system with effective feature that encodes coarsely the working space through a projected 2D laser spots, and can deliver in short time 3D range information. In addition, a new concept that decompose spatially the depth of robot's trajectory into parallel virtual planes have been developed and integrated with the sensor to facilitate fast detection and support real time obstacle detection and avoidance.

2 Motivation and Objectives

The key points that have initiated the development of a real time detection, monitoring, and tracking system can be summarized as follow,

- a. The aging population of all developed countries puts a double strain on health-care resources: more healthcare services are required while comparatively fewer working age staff is available to provide the required services. In healthcare, there is a need to monitor patient behavior in a hospital room, people with chronic illness at home (bedroom, toilet, bathroom, etc.), elderly people, patients with diseases like Parkinson and Alzheimer, and early age children, while keeping their privacy (not to see them directly on a monitor through a camera). The monitoring task is mainly required to enhance safety and to extend urgent help to overcome dangerous situations at places that are difficult to access due to privacy or isolation. In general, there is an urgent need to react quickly and deliver timely help for people in need for it.
- b. Intelligent and autonomous robots works in real physical environments are essential due to their immediate applicability in a variety of tasks. The ability for an autonomous mobile robot to interact, to move safely, navigates and adapt intelligently in uncertain, complex and dynamic environments represents the core functionality any autonomous mobile robot should be endowed with. Such robots should have significant and flexible sensing capabilities to reach a specified target and to act in real time while avoiding obstacles and achieve an assigned task. The problems associated with the available sensors are computation time, cost, and reliability.
- c. Humans and intelligent robots living in a real world in which 3D visual information is a natural requirement to fulfill the requirements of interactive and real time tasks reliably.

Accordingly, the objectives of this work focus on developing an intelligent 3D visual system that fulfils reliably real time task requirements for the purposes indicated

by the motivation points above. Such a system must be fast in terms of processing speed and decision-making, compact and cost effective. The author achieved these objectives by developing a real time monitoring and tracking system-using FG based vision sensor.

3 Fiber Grating and its Operational Principles

Diffraction gratings are useful in many optical systems. To enable high-end applications, the gratings should have high diffraction efficiency and produce multiple beams of uniform intensity. However, in an ordinary diffraction grating with a slit array a large amount of light concentrates on the zero-th diffracted order and the other higher order intensities weaken rapidly. This is due to the unmatched relation between the slit widths and the light wavelength. Therefore, the diffraction efficiency turns out to be very low. In order to obtain uniform intensity, the slit widths must be as small as the wavelengths. Accordingly, a new simple type of grating called a Fiber Grating has been developed [25]. This FG produces multiple beams of uniform intensity with high efficiency [25].

An FG is a high efficiency diffraction grating that is composed of hundreds optical fibers, cut into appropriate lengths (normally 10 mm) grouped in an array, as a monolayer with no air-gaps existing between adjacent fibers. When an FG projector constitutes two-overlaid (crossed) fiber sheets arranged at a right angle and irradiated by laser light, a sphere-like lens formed at each intersection of two fiber gratings. Spherical waves interfere with one another and generate a tetragonal 2D array of laser spots. The generated 2D laser spots array are used to encode the working space by projecting it onto a plane ahead of the sensor system as shown in Fig. 1. The author has selected a fiber of 20 μm in diameter and 10 mm in length with accuracy of 0.02 μm [25, 26]. Figure 1 illustrates the description a one dimensional FG sheet [26].

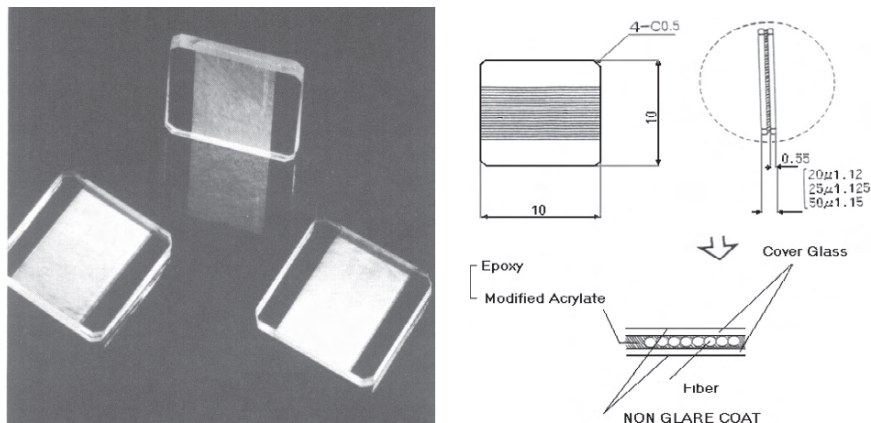


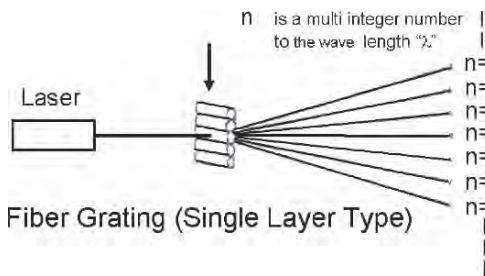
Fig. 1 Shows an example of FG with one dimensional FG sheet [25]

A sheet of fibers acts as an array of micro-cylindrical lenses; when a coherent plane wave such as a laser beam is projected perpendicularly onto the grating the laser beam is first condensed and then diverges spherically. Each optical fiber focuses the incident light just behind the grating at its focal point. Normally, each focal point functions as the arrays of light point source that emit coherent spherical wave with the same phase as the others and having a wide angle of uniform intensity [25]. Every spherical wave interferes one with another and form a diffractive spot array of a uniform intensity.

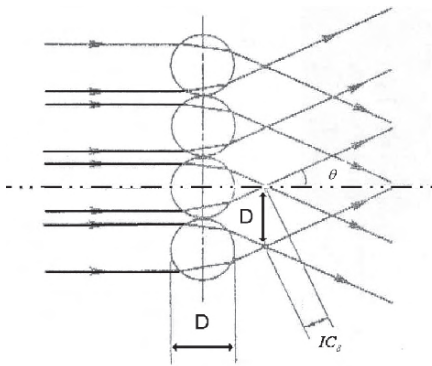
The diffraction pattern produced by single optical fiber grating in a single array represents a one-dimensional diffraction spots. Figure 2a, shows the interference condition for the FGs. ‘*n*’ is a multi integer number to the wave length ‘*λ*’. The angle of the output beams ‘*θ*’ has direct proportional relation with ‘*n*’. It increases when ‘*n*’ increases. Figure 2 (b and c) show the interference condition with its relevant parameters. The mathematical formulation for the Interference Condition is as follow [26]:

$$IC_{\delta} = D * \sin \theta = n\lambda$$

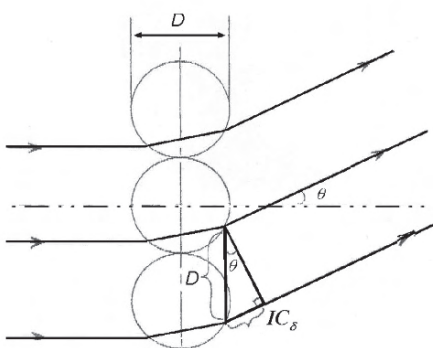
IC_δ represents the Interference-Condition
D is the diameter of the optical fiber.



(a)



(b)



(c)

Fig. 2 (a, b and c) illustrates the inference condition of the FG [26]

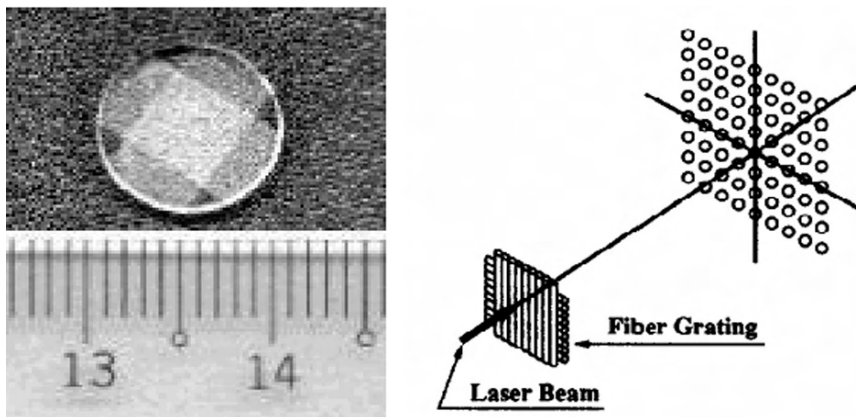


Fig. 3 2D laser spot pattern generated through two overlaid FG sheets irradiated by a laser beam

A 2D pattern of bright diffraction spots can be obtained with two overlaid identical FG sheets (Fig. 3 shows two overlaid (crossed) FG sheets), in which the two gratings are crossed at right angles [11, 12, 13, 21, 25, 26].

The experimentation shows that the distance between the source of the laser beam and the FG has no effect on spot density when it is within 2 m. When this distance is long enough for the laser beam to be scattered or absorbed in the air, the spot density will decrease accordingly [26]. However, for indoor application, the longer distance will have less effect and it is possible to compensate it through the sensitivity of the sensor, i.e., CCD camera in our case.

4 The FG Based 3D Vision Sensor System and its Hardware Configuration

The developed sensor system hardware consists of a fiber grating based vision sensor, a frame memory, an image processing and software supporting decision making, alarm indicators, and interfacing capabilities with a higher-level controller along with a user friendly interface when it is used as an independent module. The fiber grating based vision sensor consists of a bright laser-spots array projector, laser driving circuit, and a CCD camera. The laser-spots array projector consists of a semiconductor laser light source that is compact and lightweight in design (laser diode of 830 nm wavelength, and 30 mW output power), collimating lenses, and crossed fiber grating. Typically, the wavelength generated by the solid-state device in the range 750-850 nm, which is well-known for its use in digital signal storage and retrieval.

To complement these features, small diameter collimating lenses were used between the laser diode and the FG sheet to reduce laser power losses and to prevent

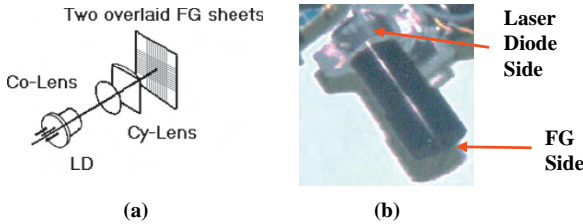


Fig. 4 (a and b) internal schematic of laser spot projector and the implemented laser projection unit

the laser beam diameter from expanding, i.e., to keep the divergence angle to a minimum (See Fig. 4a). Fine-tuning the lens distance can collimate the laser beam. The FG is composed of two identical and orthogonally overlaid optical fibers sheets. Each optical sheet consists of one hundred or more optical fibers (for the purpose of this work a fiber of $20\mu\text{m}$ in diameter and 10 mm in length was selected). One unit used to integrate and mount the laser diode, laser diver, the collimating lenses, and the FG as shown in Fig. 4b.

The FG vision sensor that represents the core of the tracking and monitoring system is small, light in weight, and easy to mount. The field of view of the CCD camera is chosen to be wide enough to cover the projected 2D laser spots matrix ahead of it (the focal length of the lens: 8.0 mm, with angle of vision in the vertical direction 41 degrees, and in the horizontal direction 51 degrees). To enhance the spectral response of the CCD camera the author used the camera without an infrared cutting filter. In addition a band-pass filter is selected, attached to the lens of the camera and used to reduce the influence from unwanted light (such as ambient light) to enable the extraction of spot related laser light. The filter selected is an optical band pass filter that can transmit more than 90% of light of wavelength 760–900 nm. The spot array is infrared light and is invisible to humans. The position of the laser spot generator can be translated (as a location) with respect to the axes of the CCD camera and the FG can be rotated (as orientation of the FG itself). Synchronization between the radiation timing of emitted laser light from a pulse laser and the shutter speed reduces the effect of ambient light. Keeping the irradiation time very short can reduce the actual power obtained after irradiating the FG.

The controller of the developed sensor system has a frame memory that receives the video signal from the CCD camera. The frame memory quantizes the image, and has the ability to store temporarily 2 images, 512×512 pixels, and 256 Gray scale levels. The processing power of the system is contained in an on-board computer based on a Pentium 3–750 MHz. (It is possible to use a better processor can for faster performance.) With the on-board computer there are I/O and communication ports used to manage situation-based alarms with details supported by live images when needed. In addition, the on-board computer has the capability to transmit alarm-based information and to communicate with higher-level controllers to facilitate efficient decision-making. There is also a laser driver circuit and a power supply to feed the overall circuits and driver's requirements. To keep the system within the

required safety standards, i.e., to have the average output of laser power within a certain limit, the irradiation time of the laser was kept as small as possible while considering other processing constraints.

5 Optical Arrangement and Working Principles

Figure 5 shows the layout of the optical arrangement for the FG vision sensor. The physical coordinate system is based on the lens of the CCD camera and the U-V coordinate is based on the image plane. The optical axis of spot array projector is parallel to that of the CCD camera, and the FG is located at $(0, -d, 0)$. The image plane is located at $(0, 0, l)$. With this arrangement, the FG vision sensor is treated as one unit, and it can be positioned and oriented as required with respect to the plane of projection [11, 12, 27]. Laser spots play an essential role in detecting and tracking targets by simply acquiring their reflection as an image through the CCD camera.

With reference to Fig. 5, a 3D coordinate system is defined and the physical coordinates of any disturbed spot on an object can be expressed by (X_s, Y_s, Z_s) , s is for spot. This information is determined from the new (u, v) coordinate values of disturbed spots on newly acquired image plane. The value of Z_s is limited within the measurement space as it is bounded by the maximum detectable object height by

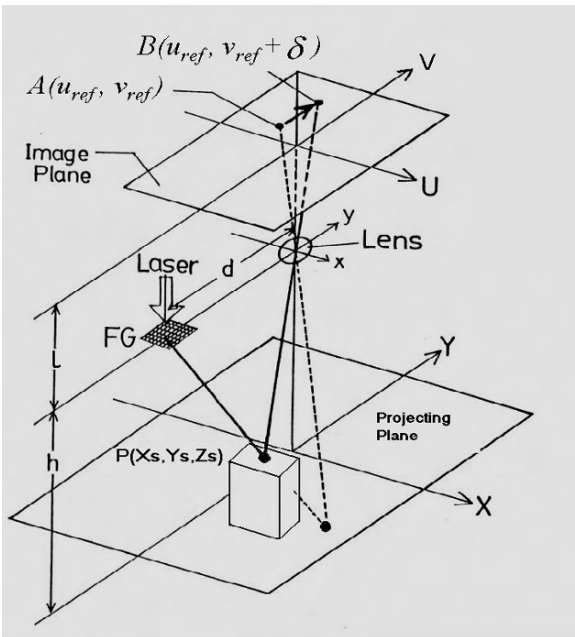


Fig. 5 The optical arrangement of the developed FG vision sensor [28]

the developed sensor. It is possible to clarify the measurement space of a detectable object height or range in relation to the optical arrangement as follows: When projecting a laser spot on a person or on an object, the position of the projected spot translates from its original reference in a particular direction on the image plane from $A(u_{ref}, v_{ref})$ to $B(u_{ref}, v_{ref} + \delta)$. The number of pixels that represent spot's translation (shift in position) due to the presence of an object is determined from the newly calculated coordinate values of the spot (at point B) coordinate values with respect to its reference coordinate on the image plane. Accordingly, the new 3D physical location of a disturbed spot (X_s, Y_s, Z_s) can be calculated using the following equations,

$$Z_s = \frac{h^2 \delta}{dl + h\delta} \quad (1)$$

$$X_s = \frac{u_{ref}}{l}(h - Z_s) \quad (2)$$

$$Y_s = \frac{(v_{ref} + \delta)}{l}(h - Z_s) \quad (3)$$

where h is the height that represents the distance between the center of the CCD lens and the reference plane surface. δ is the shift quantity of the translated spot on the image plane calculated in distance. l is the lens focal length and represents the distance from the lens to the image plane. d is the center distance from the laser-spot generator to the lens of the CCD camera. u_{ref} and v_{ref} are the reference image plane coordinates of a spot (i.e., the original coordinates of a spot before moving). The origin of the image coordinate frame is transformed to the center of the image plane and the corresponding physical zero of the coordinates on the image plane is set at $v = 255$ and $u = 255$, and all of the calculation is made relevant to it.

It is important to mention that the parameters d , l and h influence directly the performance of the FG based vision sensor system. The position of a disturbed spot within the image plane translates only along a specific path and direction, and a maximum number of pixels bound it. The number of pixels describing the translation path of a spot defines the maximum possible detected height of a target or a range/depth to that target with respect to the sensor unit and consequently affects the resolution of such detection. Hence, in order to have better resolution, there is a need to have a longer path for spot's movement, i.e., more number of pixels describing the path, an efficient FG alignment technique has been decided to generate and project an effective spot pattern structure. The adopted FG alignment is achieved by rotating the FG before fixing it to the spot generator at an angle (ϕ) to have a longer path described by more pixels for spot movement. Table 1 shows the experimental data that compares a spot's movement limit in pixels for different type lenses.

The same principle will stay valid when installing the FG vision sensor at a sidewall or on a mobile robot and properly oriented with respect to the plane of projection.

Table 1 FG Alignment Comparison

Lens Focal Length mm	Movement limit of spot's path using horizontal alignment (zero angle) pixels	Movement limit of spot's path by rotating the FG with an angle ϕ Pixels
4.8	14	36
6.5	19	46
8.0	23	54

6 FG-CCD Sensor Installation, Spot Pattern Projection and Effective Visible Area

The FG-CCD sensor when fixed at a ceiling with spot projection on the floor surface represented by a plane perpendicular to the Z-axis. This will have the following disadvantages,

- Limit in the visible area. Hence, on order to cover a wider area there is a need to use wide-angle lenses and this leads to nonlinearity problem, which complicates the searching process for a spot translated along its path. In addition, the use of a wide angle lenses affect negatively the resolution of the sensor.
- Limit the area at which the full height of human/object is detectable.
- Difficult to track people from the top.

In the case of a FG-CCD sensor fixed to a sidewall or to a side-stand with spot projection on the plane making an angle with Z-axis. In addition, it can be fixed horizontally on a mobile robot with projection on the plane perpendicular on the trajectory of the robot. This will have the advantages,

- Wider view with suitable depth is achieved and this view can be covered by a single laser-spot array generator.
- Overcome the problem of lens nonlinearity as it become possible to achieve the required view using lens with larger focal length.

It is important to keep in mind that the effective area of the projected spots does not increase by using a wide angle lens. For this and when wide effective coverage is needed it is required to use multiple FG heads (laser-spot array generator) to cover such area with a single camera. It is not necessary for d values that describe the locations of laser-spot array generators to be the same with respect to the CCD. Proper distribution around the CCD and a suitable selection of d values for each of them will lead to proper coverage of a wide area with good contributed resolution by each laser-spot array generator. This means, laser-spot array generators aiming at a far section of the targeted scene with respect to the CCD camera are positioned within a larger d value while laser spot generators aiming at a closer section of the targeted scene are positioned within a smaller d value from the CCD. Laser-spot array generators can be positioned and configured along any of the axes of the CCD camera as needed and in any numbers while considering the impact on the requirements of real time processing and the active view of the CCD camera.

To specify the visible area, in which the lens attached with the CCD camera can view and the area constraints in which an object with specific height can be detected completely, it is necessary to analyze the effective view of the lens in association with the detectability of the FG vision sensor. The experimental results (In case of top projection) show that as the targeted detectable height increases, i.e., the object or human are getting closer to the sensor, the effective area that shows full detectability of that height decreases. This is valid for all types of lenses. However, the overhead view has reliable separation of people while it cannot present details features of a targeted object/human. Overhead FG-CCD sacrifices on ground coverage when ceiling is low and when ceiling is high, its resolution will be affected.

In the other cases in which tracking details of a target are required, no physical ceiling is available, or a wider effective area that shows full detectability of a targeted object height. In this case, a side view is a good alternative. In the case of side projection with an angle, Fig. 6 shows a good setup for the side view installation with the centerline of the CCD view go roughly through the edge between the front vertical wall and the floor. In this case, the sensor can detect a better level of maximum height including human beings. This will be shown later in the experimental results.

For monitoring a person in a room, and to cover a view of 3 m width and 5 m depths, the author selected a side view installation with the following parameters:

- Lens with 8.0 mm focal length attached to the CCD camera.
- 12 mm distance (d) for proper pixel resolution related to the application, and
- 230 cm height (h) between the targeted surface and the CCD camera.

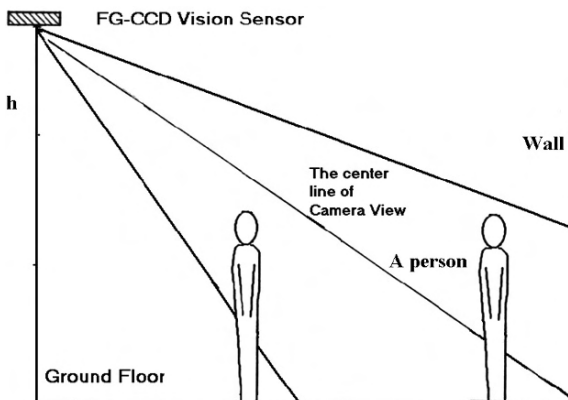


Fig. 6 The selected slide view for monitoring and tracking a person in a room

7 Operational Software Development

The developed system software consists of three main parts: the generation of a reference spot frame and real time tracking. The following sections describe these parts.

7.1 Reference Spots Frame Generation

The operation of the FG vision sensor to detect and track an object and to measure its 3D information was observed by detecting the translation movement of the spots disturbed by the object relative to a reference spots frame, i.e. the reference spots frame plays a main role at the functional level of the sensor. For this, it is necessary to describe how to generate the reference spots frame and to know what are the necessary steps needed for having reliable and accurate spot reference data. A reliable reference spot frame is necessary for accuracy in detecting spot translation movement. An efficient algorithm has been developed for fulfill this purpose, and is described as follows:

- a. Acquire an image without projecting the laser beam on the Fiber Grating, i.e. the laser switched OFF.
- b. Acquire an image while the laser is ON, i.e. a 2D laser spots pattern is projected. Steps (a) and (b) are performed in the frame memory (see Fig. 7a).
- c. Subtract the image acquired in (a) from the image acquired in (b) and store the results in the memory of the processing unit.
- d. Isolate and minimize noise effect, such as, laser speckle, by smoothing the image obtained from (c). This is done using a 3×3 averaging filter. Speckle arises when laser light falls on a non-specular reflecting surface and is caused by interference at the image plane in a camera from coherent light reflected by a non-specular (rough, at least on the scale of a wavelength of light) surface. Depending on the laser and surface, the speckle patterns can be quite dramatic [27].
- e. Apply dynamic floating threshold technique to enhance spot's area, reject background noise, minimize its effect, and optimize searching and deciding a spot. This has been done because it is necessary for the threshold level to be determined adaptively due to differences in the brightness of the background at the center of the image and that at its boundary. This is due to the zeroth order of reflection and the illumination effect of the surrounds. For this purpose the smoothed image from (d) is used again with a mask window of 21×21 . The center of this window on the image represents the pixel of interest. The output of

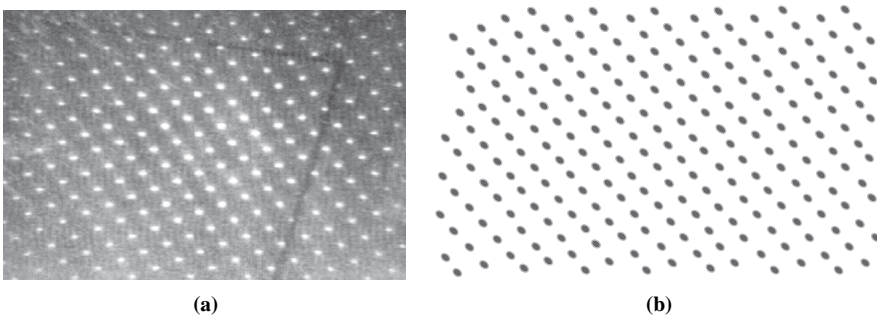


Fig. 7 (a and b) Illustrates an example of laser spot projection and the calculated reference spots frame [28]

this operation is used as a threshold to decide whether to keep the intensity of the pixel of interest (in gray level value) or to reset it to zero. The output image from this stage includes only the Gray level of the pixels that pass the threshold check and have the possibility to be a part of a spot's area.

- f. Extract the area of each spot from the image obtained in (e).
- g. Calculate the position in U-V coordinates of the identified spot along with its brightness using the information obtained in (f). Since the area of each spot is not uniform, the position of a spot ($u_{\text{ref}}, v_{\text{ref}}$) is calculated by determining the center of gravity for the extracted spot's area. The input information for this purpose is the (u, v) coordinates and the brightness of each pixel within the identified spot's area. The brightness of each reference spot is calculated by having the average brightness value of the top three pixels within the identified spot's area.
- h. Save the calculated position ($u_{\text{ref}}, v_{\text{ref}}$) and brightness of each spot obtained in (g) (see Fig. 7b) in a file called reference spots frame data file.

During the generation of the reference spots frame, it is necessary to avoid having any moving object within the active view of the sensor. The reference frame is calculated off line, and can be calculated at the initialization procedure of the system before starting real time operation as needed. The generation of the reference spot frame is done only once and there is no need to change it until there is a change in the arrangement of the sensor itself. The reference spot data is considered to be at zero height, i.e. Z_s value of all spots at the reference surface is zero. Image filtering and floating threshold is done only for calculating reference spot frame, i.e., no filtering is required during the real time processing. The performance of the sensor was tested in an indoor environment. The time required to generate the reference spot data is 1.2 min with the tested system processing capability.

7.2 Real-Time Processing

During real time processing images are captured at constant intervals and the system tries to detect and track an object and calculate its 3D information by recognizing disturbed spots and measuring the translation of these spots in pixels on the image plane. Such tracking needs to be done in a short time especially when the object is moving in real time. To achieve this, two scan algorithms were developed which simplify the way to identify the disturbed spots and to calculate the information necessary to specify the status of the object. The two scan algorithms are described below.

During real time sensing images are captured at constant intervals. For each scan images are acquired one with the laser OFF and the other with the laser ON. The image with the laser OFF is subtracted from that with the laser ON.

7.2.1 Short (fast) scan

The purpose of the short scan is to check whether any of the reference spots have been translated/moved from its reference positions $(u_{\text{ref}}, v_{\text{ref}})$. The check is performed based on the spot's reference brightness as follows:

- Select a spot (spot i) from the reference spots frame. All spots within the reference frames have their $(u_{\text{ref}}, v_{\text{ref}})$ on the image plane along with their reference intensities.
- Using the selected spot's reference coordinates $(u_{i\text{-ref}}, v_{i\text{-ref}})$ access directly the same position at the currently acquired image and read its intensity $B_{i\text{-Current}}$ (brightness).
- Compare the intensity of the reference spot $B_{i\text{-ref}}$ selected in (a) with the intensity $B_{i\text{-Current}}$ for each pixel around the same spot reference coordinate inside a newly acquired image and within an area of 5×4 mask window with its center occupied by the coordinate of i spot as indicated by the generated spot's reference frame data file. If the difference in intensities between the reference spot and that of any of the pixels within area covered by the mask window is within a specified threshold, the spot is assumed to be unmoved from its reference position.

7.2.2 Long Scan

For laser spots which have moved, all the pixels within the maximum moving path limit of a spot's translation movement are searched for the purpose of specifying the number of pixels that the spot has been translated and accordingly to calculate its new physical (X_s, Y_s, Z_s) coordinates using equations (1), (2) and (3).

The search is conducted along spot's i moving path covering an area of two pixels to each side of the moving path and for the whole of its limit while starting from the second pixel on the path.

The intensity of each pixel $B_{n,m}$ within the searched area of spot's i is checked in relation to its reference intensity to conclude spot's displacement. The dimension of the searched area is $5 \times (\text{Spot-Path-Limit} - 1)$, i.e., $n = 5$ and $m = (\text{Spot-Path-Limit} - 1)$.

In real time processing, the time required to process one image frame and approach a conclusion about it, is nearly equal to 100 msec using the processing capabilities with the sensor system that have mentioned in Sect. 4.

Figure 8 displays an image acquired through the CCD camera after having an object on the reference plane. The disturbed spots due to the presence of the object have been identified, and their displacement has been calculated and highlighted on the same figure. The 3D position of each spot is successively calculated by triangulation using equations (1), (2) and (3). The value of each of the moved spot gives the range between the relevant object surface and the optical sensor. In addition, the 2D shape data extracted from the intensity image can be easily obtained. Accordingly, the position of the object is calculated using the center of gravity and moment-invariant techniques.

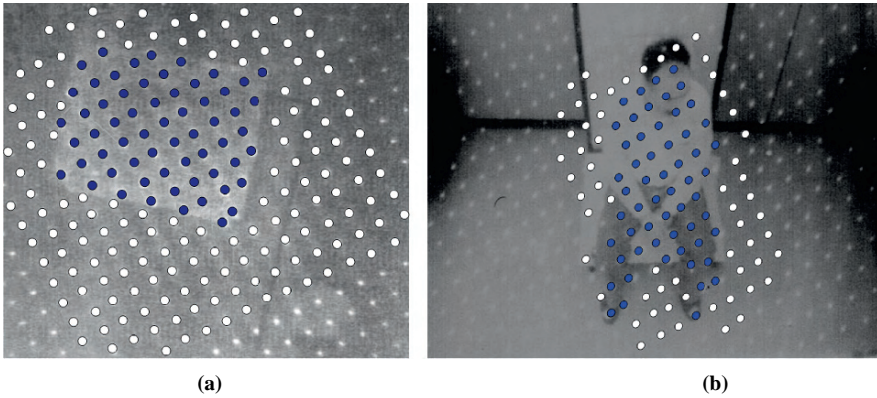


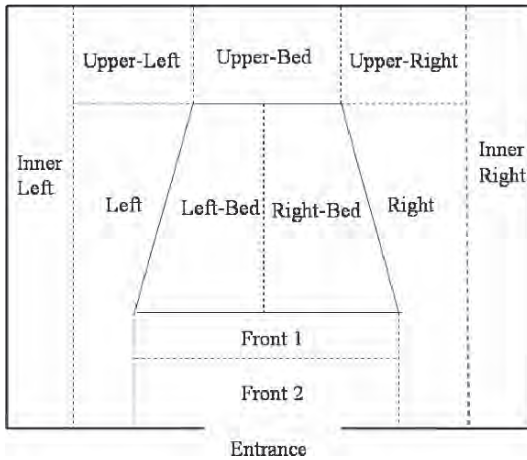
Fig. 8 (a and b) illustrate two example for overhead projection and side projection with disturbed laser spots due to the presence of an objects and human respectively

8 Application and Experimental Results

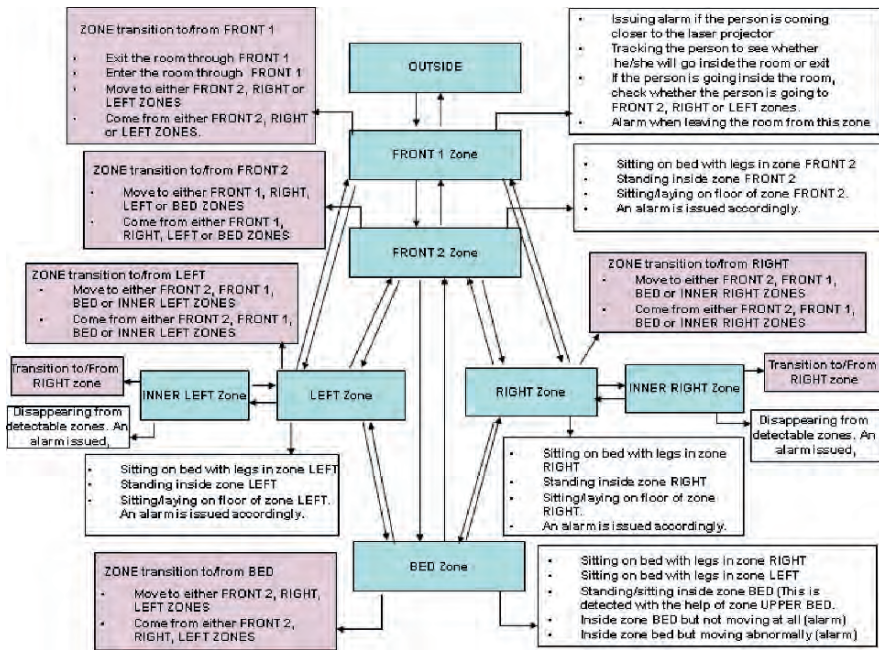
The selected application for the developed sensor system aims to track and monitor a person in a room environment and in real time, and to issue an alarm when detecting unsafe behavior. An alarm is issued when one of the following cases is detected:

- a. The person is not moving at all in his bed within a specified time;
- b. The person is a wake in his bed without sleeping within a specified time,
- c. The person is leaving his bed;
- d. The person is laying down on the ground;
- e. The person is leaving the room;
- f. The person is moving abnormally within a specified time; Etc.

To facilitate efficient and fast tracking and monitoring functions, the room environment has been divided into 11 zones that support smooth transition for tracking the movement of a person between them (see Fig. 9a). These zones are: Front 1, Front 2, Right, Upper Right, Right-Bed, Left-Bed, Upper-Bed, Left, Inner Left and Upper-Left. During the generation of the reference frame, the spots inside each of the configured zones are specified and located automatically. It is important to mention that the upper zones (left, bed and right) are not specified for tracking movement rather to help in specifying human status within other zones. The system is dedicated to track one person (the patient in this case) and monitors his/her behavioral status with full identification of that person. The frequency for the change in moving spots has been used to differentiate between stationary objects and human beings. In addition, when an authorized service person wants to enter the room, he/she can suspend the operation of the sensor during the time of the service and enable it again later. Fig. 9b shows the transition between zones for tracking purposes. In tracking a person movement, the number of detected moved spots within any of the zones is used as an indication for tracking and this helped to avoid the calculation of the physi-



(a) The partition of a room environment into zones. For tracking and monitoring of a person



(b) The transition movement between zones.

Fig. 9 (a and b) shows the division of a room environment into zone and the requirements for transition between zones

cal coordinates for any of the relevant moved spots and hence enhanced processing speed in decision-making. In addition, with such approach the tracking will be independent on the key parameters of the system, i.e., l , h and d that are required to calculate the physical coordinates of a moved spot.

The performance of the sensor was tested successfully under ordinary room illumination within an indoor environment including the case when the fluorescent light is ON, to detect, track and monitor a person in a room environment. However, the

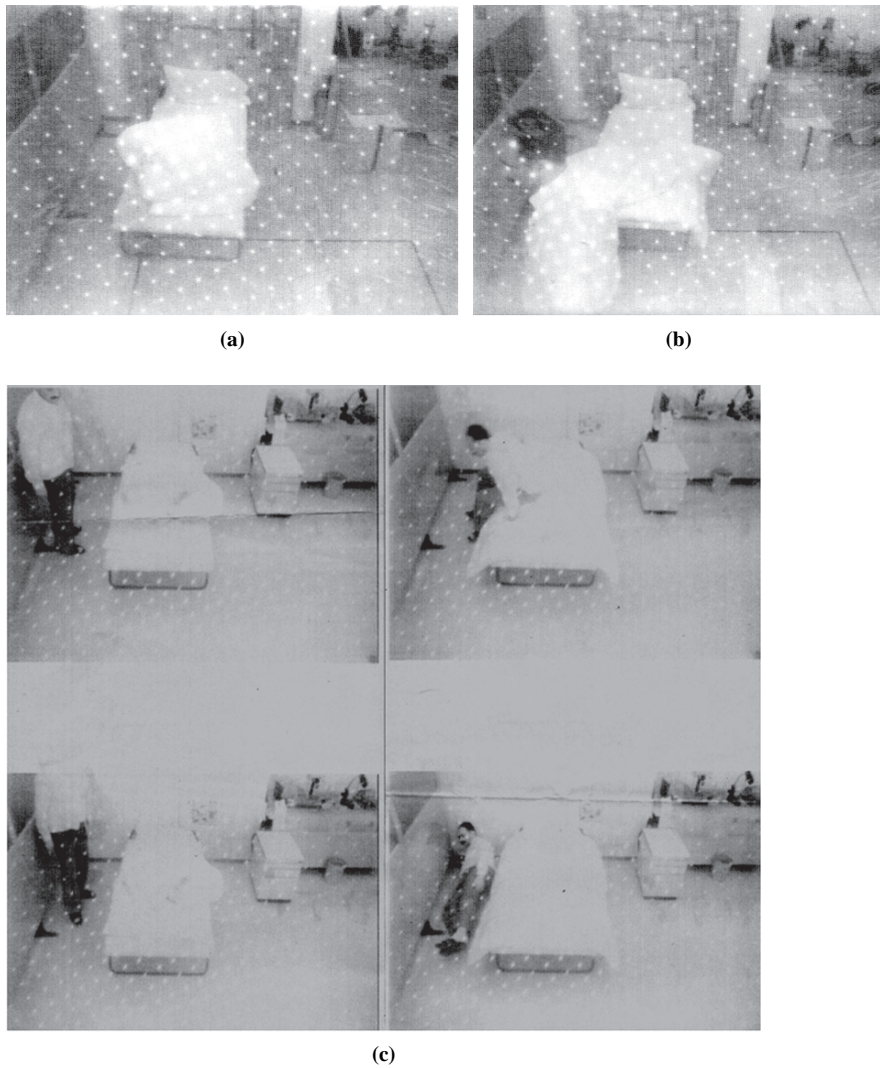


Fig. 10 (a, b and c) experimental examples illustration the use of the developed sensor system in a room environment

sensor system performance is influenced when direct sunlight enters through any of the windows within the operational indoor environment.

During the testing, no miss alarms have been issued, but a few false alarms have been identified in relation to the type of the behavioral status. While this can be tuned and minimized at a later stage, its occurrence is better than having miss alarms in which we may fail to react to dangerous situations. The false alarms at this stage require additional attention by the operator for clarification. Figure 10 shows examples of the experimental test at a room environment.

In real-time processing, the time required to process one image frame and approach a conclusion is nearly equal to 100 ms for the worst case scenario using the processing unit mentioned in Sect. 4. This speed is enough to track indoor moving activities (no running) of a normal person.

9 Conclusion

This paper presented the development and the use of a 3D fiber grating based vision sensor system. The sensor is efficient, small in size, light in weight, fast to generate 3D information, compact and cheap. The developed sensor system was tested and accurately measured the range to an object/human, recognized shape, and position of an object, and detect, track and monitored moving human/objects. The sensor is suitable for real time applications because it obtains 3D information within a short time. The sensor can be fixed on a moving device, such as an autonomous mobile robot, or it can be fixed in a static place for monitoring purpose. The key parameters that influence the range resolution and the performance of the system are given as follows, and they should be decided after clarifying the requirements of the targeted application. h is the height/depth, and it represents the distance between the center of the CCD lens and the surface of the reference plane; d is the center distance from the laser-spot generator to the lens of the CCD camera; and l is the lens focal length, and it represents the distance from the lens to the image plane.

The experimental analyses shown the possibility to cover wider and deep scene using one CCD camera with properly configured multiple number of laser-spot array generators around it. The problem of nonlinearity appeared within the image plane when choosing lenses with small focal length. While there is advantage in using a small focal length lens to widen the visible area by the CCD, it adds nonlinearity problem that affect the searching process for a moved spot. For this, when using a small focal length lens, it is necessary to overcome such nonlinearity/distortion by developing algorithm to compensate results on the distortion of the optical aberration.

The author is currently working to expand the applications and enhance sensor resolution and accuracy for wider applications.

References

1. M. H., Lee, "Intelligent Robotics", Halsted Press and Open Univ. Press, 1989, Chap. 1, pp. 1–12.
2. R. D., Kalfter, T. A., Chemielewski, and M., Negin, "Robotic Engineering and Integrated Approach", Prentice Hall, 1989, Chap. 6, pp. 440–506.
3. M. K., Habib, and S., Yuta, "Map Representation of a Large in-door Environment with Path Planning and Navigation Abilities for an Autonomous Mobile Robot with its Implementation on a Real Robot", *Automation in Construction*, Vol. 1, No. 2, 1993, pp. 155–179.
4. S., Suzuki, M. K., Habib, J., Iijima, and S., Yuta, "How to Describe the Mobile Robot's Sensor-Based Behavior", *Robotics and Autonomous Systems*, 7, pp. 227–237, 1991.
5. M., Kondo, S., Tachiki, M., Ishida and K., Higuchi, "Automatic measuring system for body fit on the automobile assembly line", *IEEE International Conference on Robotics and Automation (ICRA'1995)*, Vol. 1, 1995.
6. J., Ishikawa, K., Kosuge and K. Furuta, "Intelligent control of assembling robot using vision sensor", *IEEE International Conference on Robotics and Automation (ICRA'1990)*, 1990, Vol. 3, pp. 1904–1909.
7. S. Y., Chen, W. L., Wang, G., Xiao, C. Y., Ya and Y.F., Li, "Robot perception planning for industrial inspection", *IEEE Region 10 Conference TENCON*, 2004.
8. J., Lee, "Applying 3-D vision to robotic manufacturing automation", *Proceedings of Rensselaer's Second International Conference on Computer Integrated Manufacturing*, Troy-USA, May 1990, pp. 99–104.
9. S., Li, I., Miyawaki, K., Ishiguro, and S., Tsuji, "Finding of 3D structure by an active-vision-based mobile robot", *IEEE International Conference on Robotics and Automation (ICRA'1992)*, 1992, Vol. 2, pp. 1812–1817.
10. M. A., Garcia and A., Solanas, "3D simultaneous localization and modeling from stereo vision", *IEEE International Conference on Robotics and Automation (ICRA'2004)*, 2004, Vol. 1, pp. 847–853.
11. K., Nakazawa, S., M., Nakajima, and H., Kobayashi, "Development of 3D shape measurement system using Fiber Grating", *Trans. IEICE*, Vol. j69-D, No. 12, 1986, pp. 1929–1935.
12. M. K. Habib, "Development of 3D Fiber Grating Based Vision Sensor", *The second ACCV'95*, Singapore, 1995, pp. II 269–274.
13. J., Yamaguchi, H., Gou, and M., Nakajima, "Finding Intruders System using Hologram Disk", *Technical Digest of the 8th Sensor Symposium*, 1989, pp. 83–86.
14. C. E., Smith, C. A., Richards, S. A., Brandt, and N. P., Papanikolopoulos, "Visual Tracking for Intelligent Vehicle-Highway Systems", *IEEE Transactions on Vehicular Technology*, Vol. 45, No. 4, Nov. 1996, pp. 744–759.
15. C., Setchell, and E. L., Dagless, "Vision-based road-traffic monitoring sensor", *IEE Proceedings-Vision, Image and Signal Processing*, Vol. 148, No. 1, Feb. 2001, pp. 78–84.
16. M.-Y. Kim, K.-W. Ko, H.-S. Cho and J.-H. Kim, "Visual Sensing and Recognition of welding environment for intelligent shipyard welding robots", *Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)*, Vol. 3, 2000, pp. 2159–2165.
17. T. Arjuna Balasuriya; Ura, "Vision-based underwater cable detection and following using AUVs", *Oceans '02 MTS/IEEE* Vol. 3, 2002, pp. 1582–1587.
18. S., Grange, E., Casanova, T., Fong, and C., Baur, "Vision-based sensor fusion for human-computer interaction", *2002. IEEE/RSJ International Conference on Intelligent Robots and System*, Vol. 2, 2002, pp. 1120–1125.
19. F., Wallner, R., Graf, and R., Dillmann, "Real-time map refinement by fusing sonar and active stereo-vision", *Proceedings of the 1995 IEEE International Conference on Robotics and Automation*, Vol. 3, 1995, pp. 2968–2973.
20. J.-H., Kim and J. C., Myung, "SLAM with omni-directional stereo vision sensor", *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, Vol. 1, 2003, pp. 442–447.

21. K., Terada, J. Yamaguchi and M. Nakajima, "An Identification of human faces using bright-spots matrix projection", Proceedings of the 1993 International Joint Conference on Neural Network, Nagoya-Japan, 1993, pp. 2093–2096.
22. U. Gopinathan, D. J. Brady, and N. P. Pitsianis, "Coded apertures for efficient pyroelectric motion tracking", The international Electronic Journal of Optic, Vol. 11, No. 18, September 08, 2003 , pp. 2142–2152.
23. A. Armitage, D. Binnie, J. Kerridge, and L. Lei, "Measuring pedestrian trajectories with low cost infrared detectors: preliminary results.", Pedestrian and Evacuation Dynamics, Greenwich: CMS Press London, 2003, pp. 101–110.
24. P. Aarabi, "The fusion of distributed microphone arrays for sound localization", EURASIP Journal of Applied Signal Processing, Vol. 4, 2003 pp. 338–347.
25. H., Machida, J., Nitta, A., Seko, and H., Kobayashi, "High Efficiency Fiber Grating for Producing Multiple Beams of Uniform Intensity", Applied Optics, Vol. 23, No. 2, pp. 330–332, 1984.
26. Machida Endoscope Co. Ltd., Tokyo, Japan, 1996 direct communication.
27. T., Asakura, and N., Takai, "Dynamic Laser Speckle and their Application to Velocity Measurements of Diffuse Object", Applied Physics, Vol. 25, 1981, pp. 179–194.
28. M. K., Habib, "Fiber Grating Based Vision System for Real Time Tracking, Monitoring and Obstacle Detection", IEEE Sensor Journal, Vol. 7, No. 1, Jan. 2007, pp. 105–121.

Dynamic VRML-Based Navigable 3D Map for Indoor Location-Aware Systems

Wan-Young Chung and Chi-Shian Yang

Abstract Three Dimensional Navigation Viewer (3DNV), a convergence of location-aware application and three-dimensional (3D) graphics technology are developed for a 3D visualization of location-aware information. The system allows visualization of situational information in a complete, 3D model of indoor environments equipped with instantly updated route results, synchronized with physical world. The approach is validated via indoor context-aware technologies, Cricket and Received Signal Strength Indication (RSSI). The overall results provide a valuable insight into the novel integration approach between 3D graphics standard, Virtual Reality Modeling language (VRML) and indoor location-aware systems.

Keywords 3D · 3D navigation viewer · 3D Graphics · VRML · cricket · RSSI · Indoor Location-Aware Systems

1 Introduction

Recent technological advances have made it feasible to track location of people, computers, and practically any other objects we concern about. Today, there exist a number of deployed indoor location-aware applications as location-awareness becomes an essential feature of software applications.

When navigate in an unfamiliar environment people tend to rely on maps. Even though two-dimensional (2D) map display is a well known visualization technique, it has been found out that the problem using the traditional 2D map is the inability to understand the spatial relationships between the physical world objects and to match them with 2D map. Location-based information: distances, landmarks, and

Wan-Young Chung
Department of Computer & Information Engineering, Dongseo University, Busan 617-716, Korea,
e-mail: wychung@dongseo.ac.kr

Chi-Shian Yang
Department of Ubiquitous IT, Graduate School of Design & IT, Dongseo University, Busan
617-716, Korea

objects are easier to perceive via 3D map [1] as they represent the world as we see in real life. Furthermore, it includes features that are not possible with 2D map, for example positioning services and dynamic information gathered from tracking devices. As 3D map provides a more realistic way to display situational information, by making it works in real-time, it offers the additional ability to continually update the 3D scene, synchronized with user's navigation in physical world. This allows user even better visualization than simply studying a 2D plan.

Hence, our motivation is to bring elaborated overview of nowadays VRML visualization technology and its association with indoor location-aware systems. VRML [2] is a modeling language used to describe interactive 3D worlds. It offers a higher-level abstraction with variety of nodes that serve different purposes. With the combination of these nodes, not only physical attributes and spatial positions, real and nature behaviours of shared objects can be represented as well.

Specifically, this paper describes a dynamically constructed 3D map that delivers and conveys location information of a target continuously and instantly upon his navigation in physical world.

2 Indoor Location-Aware Systems

To fully utilize 3DENV as a location-based navigation viewer in three-dimensional, it must be augmented with a location-support system. A location-aware system must be based on a well-defined location module as knowledge of locations of targets and equipments is prerequisite for the support of context-aware applications. With the knowledge of positions of target, 3DENV is able to present results about user's navigation in 3D world continuously.

2.1 Cricket-Based Location-Aware System

Cricket-based indoor location tracking system [3] consists of two types of devices: beacons and listener to locate targets in physical world (see Fig. 1). The stationary beacons positioned around the indoor environments transmit both radio-frequency (RF) and ultrasound signals. The listener which is collocated with target listens to packets from each beacon to determine its distance from beacons using Time Difference of Arrival (TDOA) of ultrasound and RF signals. The listener subsequently sends the distance information to server where position of users is computed.

2.1.1 Location Estimation

Cricket-based tracking system provides space and coordinates (x, y, z) (in centimetres) information. The space information can be gathered from beacons that demarcate boundaries while location is determined using the distance information from the reference beacons and their coordinates by solving triangulation algorithm (see (1)).

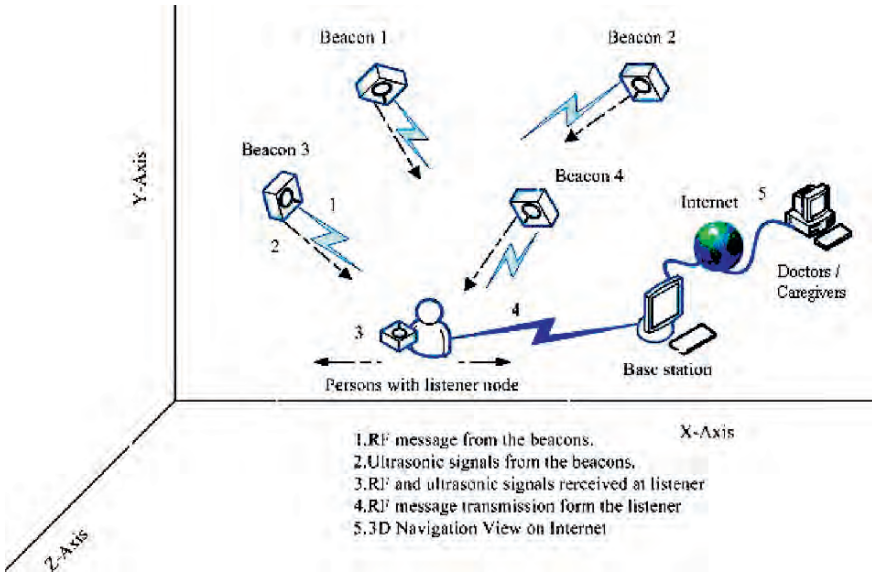


Fig. 1 System architecture of Cricket-based indoor location tracking system

The measurement error for each distance from the beacons was neglected as the system assumes the error is small. Nonlinear optimization technique is applied when the number of beacons is more than three.

$$\begin{aligned}
 d_1^2 &= (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 \\
 d_2^2 &= (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 \\
 d_3^2 &= (x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2
 \end{aligned}
 \tag{1}$$

2.2 RSSI-Based Location-Aware System

Given a model of radio signal propagation in a building or other environment, Received Signal Strength Indication (RSSI) [4] can be used to estimate the distance from a transmitter to a receiver, and thereby estimate the position of a mobile node.

2.2.1 Received Signal Strength Indicator (RSSI)

Second last byte of a packet read from CC2431 contains the RSSI value that is measured after receiving eight symbols of the actual packet. It can be erroneous when a large number of nodes are talking on the same channel at the same time as it will reflect not only the signal power belonging to the received data but also the intensity of the received signal strength at the moment. It is typically in the range of -40 to -90 dBm, where -40 dBm is the highest value.

2.2.2 Theoretical Signal Propagation

RSS is a function of the transmitted power and the distance between sender and receiver. It decreases with increased distance (see (2)).

$$\text{RSSI} = -(10n \log_{10} d + A) \quad (2)$$

n : signal propagation constant, also named propagation exponent.

d : distance from sender.

A : received signal strength at a distance of one meter.

2.2.3 CC2431 Location Engine

The location algorithm used in CC2431 location engine is based on RSSI values [5]. CC2431 uses the RSSI value combined with the physical location on reference nodes to calculate its own position.

Reference node is a node which has static location and it must be configured with x and y value (in meters) that correspond to the physical location. Reference node with highest RSSI value is used for calculation. Blind node initializes all communication required by broadcasting messages to reference nodes that are in the range. It gathers data from each of the reference nodes and communicates with its nearest reference nodes, collecting x , y and RSSI for each of these nodes and calculates its position based on this parameter input using location engine hardware. Average RSSI value for each node is calculated and fed into the engine. After calculation the new position is sent to 3DNV for visualization.

3 Modeling Approach Using VRML

Modeling technique used in this implementation can be divided into three distinct stages presented by workflow scheme as in Fig. 2.

The first stage is getting data for the building and indoor environments to be constructed. The efficient 3D world generation is greatly depending on the data required by the next stage. Construction with complex procedure will definitely complicate the process of the task as more data is required. Besides, locations of objects inside the particular building within their 3D environment are necessary in which manual input is required.

To construct indoor environment from the data gathered at the second stage, floors and objects are modelled and placed in a 3D world. The amount of details required in the environments need to be carefully considered before constructing the environments.

The final part is to export data in a format that can be rendered and subsequently render the data. Every object inside the building must be translated to its accurate position in the modelled environment. This requires additional input data to shift an object to its location.

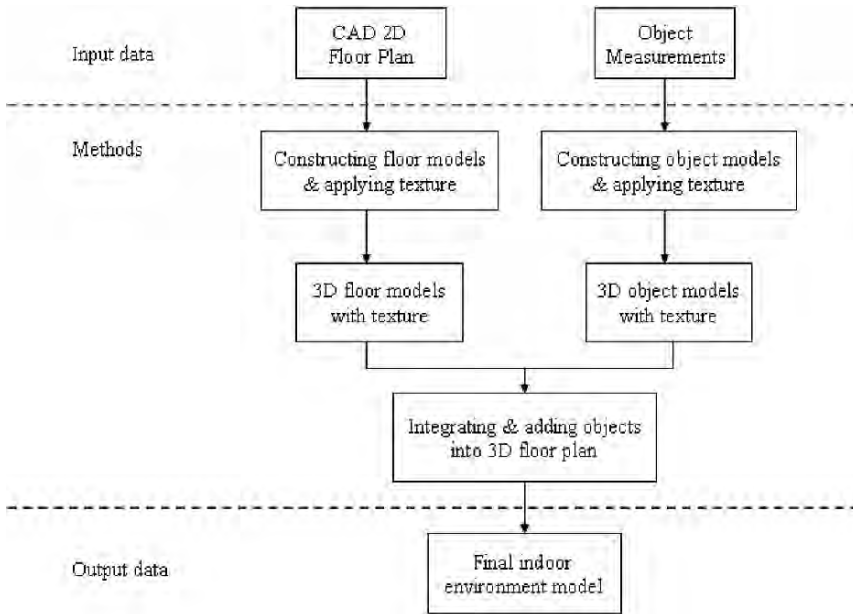


Fig. 2 A workflow scheme of methods used in modeling 3D indoor environment

3.1 Acquiring Data

The starting point of any creation process is to gather information of the architectural and other background information of the environment to be modelled. Automatically extracting data from plans would greatly reduce the effort required to generate input data for the construction process. However, from [6] it can be seen that this was a complex procedure which has not been achieved satisfactory.

Although the chosen method of generating the input data required manual input, it is simple and quick to produce the basic structures. This method is robust as it does not rely on poorly structured and error prone input data. Creating objects in a 3D world involves manual input as it is unlikely a single 2D plan would contain all the information needed to construct a 3D world.

The core data source of constructing 3D map for indoor environments is the computer-aided design (CAD) floor plan (see Fig. 3). Basic geometry was created from this CAD file as the template for the construction of our 3D indoor environment model.

Gathering data for objects inside the environment is not simple as it is recommended to do the measurement manually which is particularly time consuming. This is due to the influence of accuracy of measurements on location positioning in 3D world.

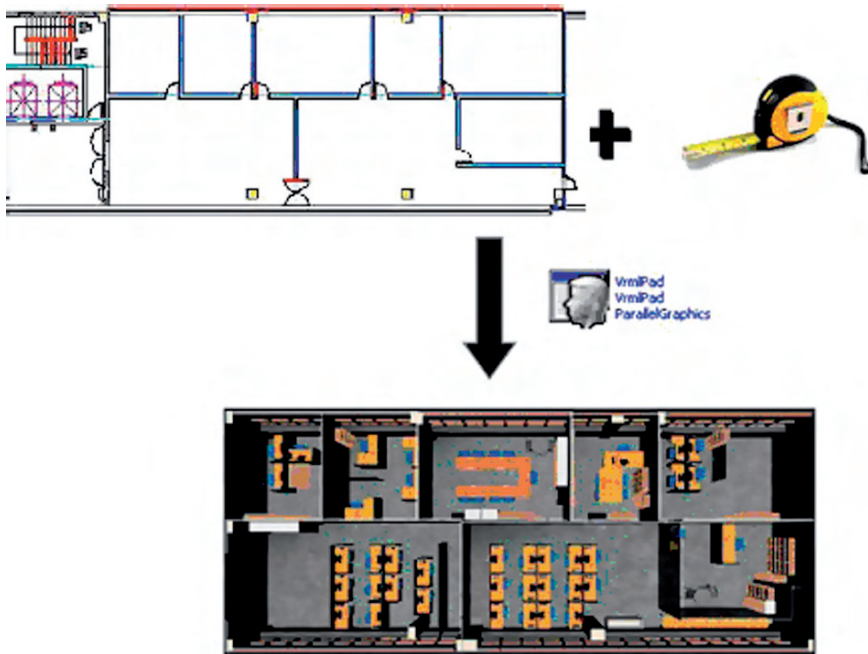


Fig. 4 Basic stage of workflow

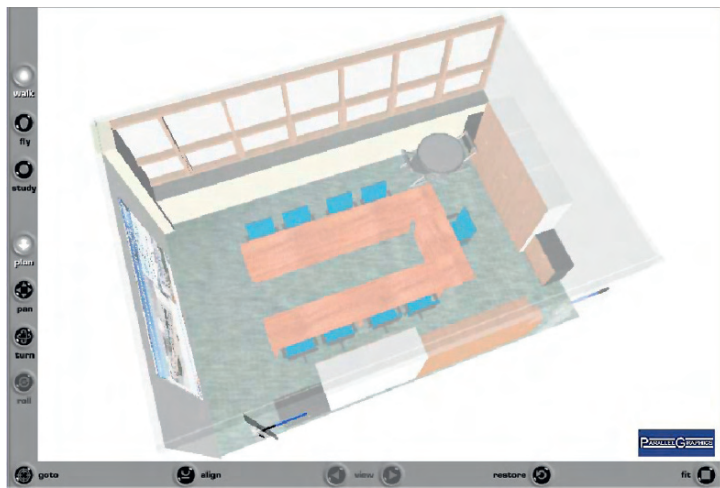


Fig. 5 Final product of modeling 3D indoor environment, Cell U803 viewed via Cortona3D Viewer

4 Visibility Computations and Determination

A huge number of calculations need to be done in various stages of rendering pipeline before rendering a single polygon. This is why it is rather difficult to achieve sufficient rendering speed with complex models which consist of several millions of polygons. This section presents portal culling algorithm, improves performance of the rendering process by render only cells and objects that are visible to target (see Fig. 6).

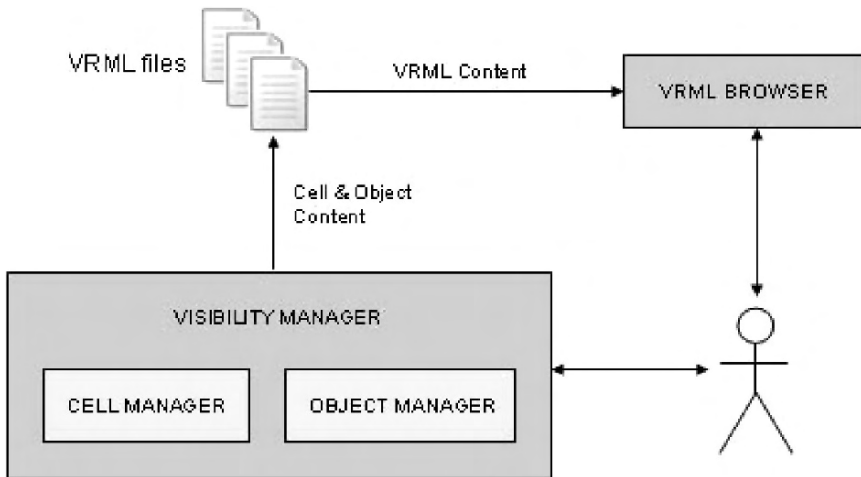


Fig. 6 Overall architecture of proposed portal culling algorithm

4.1 Portal Culling Algorithm

Our system considers portal culling algorithm [9, 10, 11]. It works on representation of the scene made of cells connected to each other by portals. The fundamental idea is that the viewer is inside a cell, and objects belonging to other cells can only be seen through portals. Culling can therefore be applied to all objects falling outside the portal areas.

There are two-type of visibility. Following [12], they can be called cell-to-geometry visibility and cell-to-cell visibility. Cell-to-geometry visibility is responsible to determine which objects are visible from a given cell and cell-to-cell visibility tells which cells are visible from some point within a cell.

Based on the visibility concepts, 3DNV supports dynamic rendering of cells by considering all of its portals from the target's current cell and checks whether they are visible: if portal is visible, the cell connected to the target's cell by the portal will be visible as well. For object retrieval, 3DNV considers position and viewing direction of user. The system seeks through the hierarchical trees describing the objects inside a cell and compares its bounding box values with user's visible frustum

vertices. Only objects reside in user's visible frustum will be sent to graphics pipeline for rendering.

4.1.1 Cell Manager

Cell manager is responsible to manage the rendering of cell. It determines the cell where the user is, and other cells that are visible from the user's position. It indicates VMRL files which files to render, load and flush, according to user's movement in the indoor environment, and VRML browser renders these files accordingly.

Cells are dynamically loaded into main memory when they are visible and flushed from it when they have not been visible. To determine user's cell and visible cells, Cell Manager exploits: (1) the position of user (2) the viewing direction of user and, (3) the boundaries of all the cells in the model.

The following presents the code of culling algorithm. When *manageCell* is called, it renders the environment by calling *renderCell* function. This function descends the tree in VRML file that represents the cell (*cell.VRML*) and renders the geometry found in the tree. The function then proceeds by considering the validity of user's viewpoint. To decide which visible cells to be loaded, user's viewing direction is essential. VRML files representing the visible cells will be rendered according to the angle of viewing direction. As soon as visible cells are determined, they will be rendered and placed on the correct position in the scene.

```
manageCell(cell,viewpoint)
begin
    renderCell(cell.VRML);
    if isViewpointValid(viewpoint) then
        angle = viewpoint.orientation.angle;
        switch (Angle)
            case '0':
                renderVisibleCell(north, cell.VRML);
            end case;
            case '90':
                renderVisibleCell(east, cell.VRML);
            end case;
            case '180':
                renderVisibleCell(south, cell.VRML);
            end case;
            case '270':
                renderVisibleCell(west, cell.VRML);
            end case;
            case 'others':
                break;
            end case;
        end switch;
    end if;
end manageCell;
```

(1)

4.1.2 Object Manager

This section explains how to conservatively estimate the cell-to-geometry visibility using a simple search in the visibility tree. Based on the appropriate cell-to-geometry algorithm, the entire space is traversed. Tree hierarchy is formed representing objects inside a cell in VRML files with a bounding box for each object's node.

In computer graphics and computational geometry, a bounding box is a type of bounding volume that completely contains the union of the objects in the set. It is a rectangular box in 3D space, with sides parallel to the coordinate planes, that contains an object. In the system, it facilitates the implementation of portal culling algorithm, producing a list of objects that necessary to be displayed.

Visible Volume Area (VVA) of user should be computed before hand. VVA is defined as part of user's viewpoint where the rendered objects in the cell can be seen. It considers two parameters: cell and viewpoint of the user. To compute potentially visible objects defined by VVA of user, calculation must be performed. Depth first search is performed to check each encountered objects its bounding box with the user's VVA. This is to decide whether the object is inside or outside of the VVA. Coordinates of each vertices of bounding box are compared to the boundaries of VVA. If the result of comparing the coordinate of bounding box with the boundaries of VVA is greater than threshold value 4, it succeeds the visibility test. Node geometry of the object is rendered as it is proved to be visible to user.

5 3D Navigation Viewer (3DNV)

The core idea of 3DNV is to develop a navigable 3D map for indoor location-aware systems. It supports interaction of indoor location-aware systems with 3D world and provides users with means to retrieve situational information. For indoor navigation, 3DNV aims to offer high quality, real time, and customized location information during navigation.

5.1 System Architecture of 3DNV

3DNV was designed in a way, which combines VRML representation of currently navigated indoor environments through the use of indoor location-aware data to offer a simple mechanism to control the interactivity and to associate the information to modelled virtual worlds. It is composed by two modules:

- Application Interface Module permits target to view 3D world and her status information. This module receives data and presents them to users.
- 3D World Module contains files storing the information of 3D models. In our strategy, it also exploits Cortona3D Viewer [8] for visualization of 3D world and

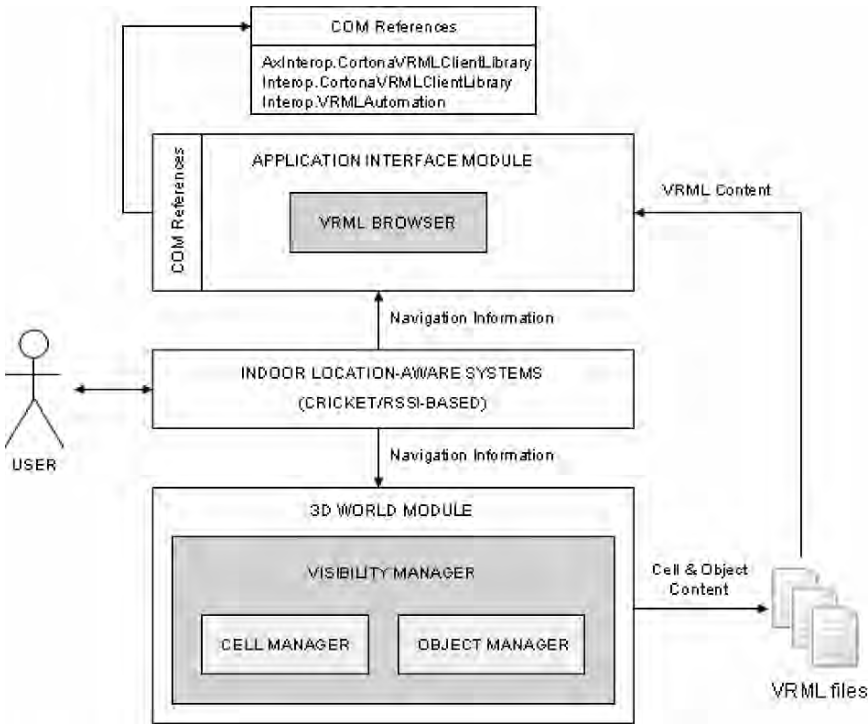


Fig. 7 System architecture of 3D NV

in charge of updating the target’s viewpoint and position based on contextual information received and efficiently manages the rendering of large-scale environments.

The fundamental elements which compose of these modules are illustrated in Fig. 7, and a detailed description on how these system modules have been implemented is provided in following subsections.

5.1.1 Application Interface Module

Figure 8 is the screenshot of 3D NV for desktop environment, showing how the interface appears to user and which features it actually offers. It can be identified into three main parts. In default mode, target’s spatial information is demonstrated in two virtual 3D views: top-down view and local view. Top-down view is a view of an object from high locations, specifically means looking straight down, perpendicular to the surface below whereas in local view only the selected objects are displayed, which can make rendering easier in complex scenes.

Part (1) is the local view 3D VRML world shows the actual visualized world based on user’s viewpoint and orientation. The frame marked with (2) contains the

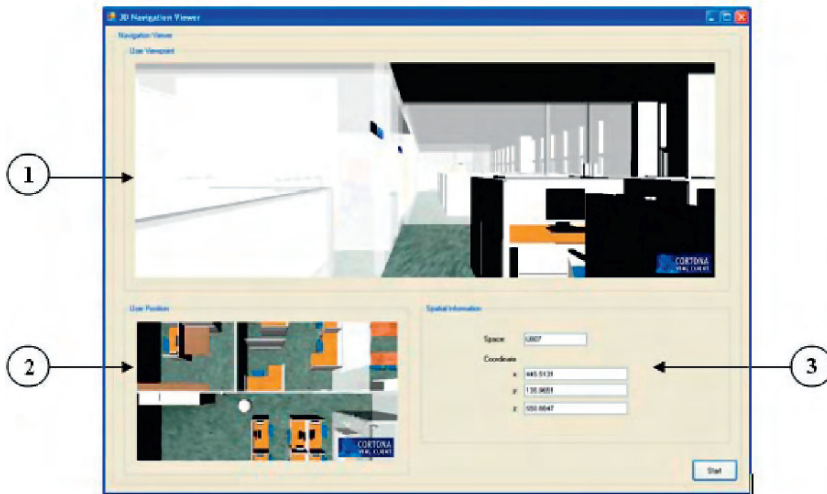


Fig. 8 Screenshot of 3DNV

top-down view map of current indoor environment with a marker to indicate the position of the target in the virtual world. Part (3) displays the Cartesian coordinate values of target to assist him to be acquainted with his position reference in the environment. When user moves in the physical world, the 3D worlds change accordingly.

5.1.2 3D World Module

As its name suggest, its main functionality is to manage the process of adapting and associating information to the virtual environment. By means of this purpose, 3D world module carries out two different processes:

- Adaptation of contents associated to the virtual world. System has to react in real time to readapt the associated information and interactivity levels to adjust the camera position and user's position indication. One of the techniques to allow 3D world module to communicate with VRML is by utilizing the VRML Automation Interface [13] from Cortona SDK. The embedding of Cortona3D Viewer into an application using Cortona SDK as programming interface makes it possible to communicate between VRML and the force feedback SDK in the application.
- Generation of VRML world on-the-fly by portal culling algorithm. When target is tracked entering one of the cells in the physical world, which is subsequently mapped into the VRML view, the VRML component issues a notification to determine which objects are visible through current viewpoint of user. To minimize the data transmitted, only the visible geometries are sent through the graphics pipeline for rendering.

6 System Implementation and Evaluation

3DENV system has been implemented in C# for desktop environment. VRML is used as the description language for the 3D world and Cortona3D Viewer to display the 3D world on the application.

To investigate effects of both tracking systems, experiments were conducted under the same environment conditions with equal system conditions. Experimental test-bed was located on the highest floor of an eight-story building. A specific cell had been chosen for the experiment, cell U803. Figure 9 shows the scenarios for the experiments in which throughout the trial, user has to walk along the same route in same order. During navigation, scene analysis was performed per pace and the time needed for execution of the system and scene retrieval on standard PC hardware with 3 GHz processor was observed.

Performance of culling approach depends on the cell location and the surrounding geometry. Therefore, in order to test the functionality and suitability of the concept of the algorithm, a variety of viewing direction and position is necessary. Figure 10 illustrates the scene rendered with and without culling algorithm. The result shows almost the entire objects outside the VVA of user are culled away by the algorithm.

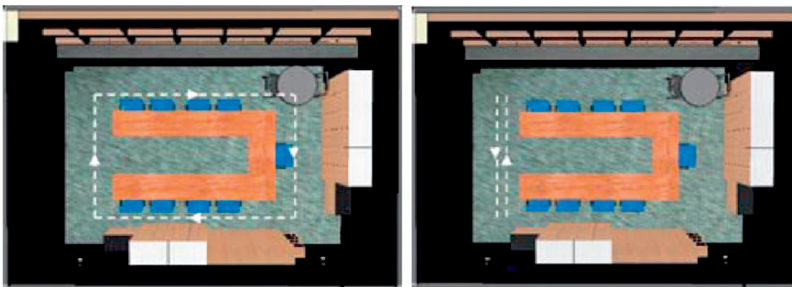


Fig. 9 Scenarios considered in the experiments. Dotted line shows the intended walking trial

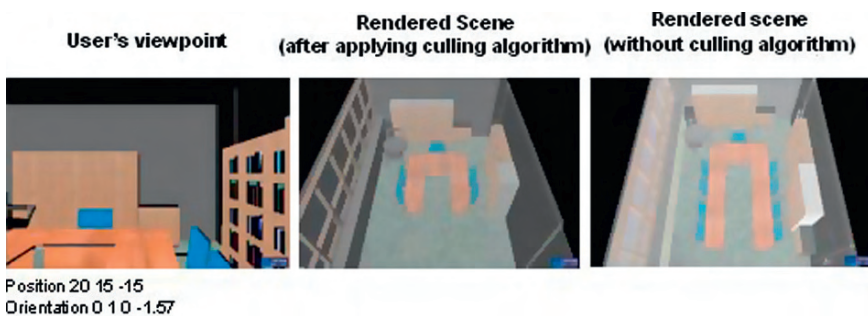


Fig. 10 Visibility culling using portal culling algorithm. Only the potentially visible set of objects is rendered according to user's location information

Table 1 Average rendering time of the 3D world both with and without culling on desktop environment (Position 10 10 -15, Orientation 0 1 0 0)

Cells	Rendering time & document weight without culling (s, KB)	Rendering time & document weight with culling (s, KB)	Speedup (%)
U801	1.7 / 0.86	0.3 / 2.11	82.35
U802	1.7 / 0.88	0.3 / 2.08	82.35
U803	1.8 / 0.99	0.5 / 3.16	72.22
U804	1.7 / 0.98	0.3 / 2.48	82.35
U805	1.7 / 0.91	0.3 / 2.23	82.35
U806	1.5 / 0.97	0.3 / 2.22	80.00
U807	1.7 / 1.02	0.3 / 2.39	82.35
U808	1.4 / 0.63	0.3 / 2.44	78.57

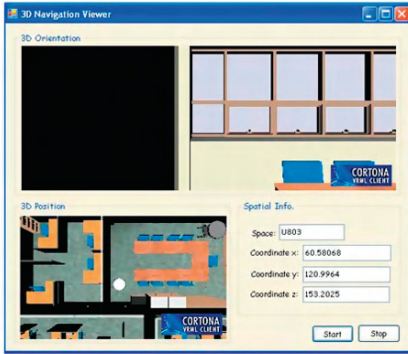
Table 1 reports real experienced rendering time of the scenes with and without culling algorithm. From the assessment, time taken for rendering is reduced approximately 80 percent for a Pentium 4 computer equipped with CPU 3 GHz, 504 MB of RAM memory and Intel 82915G/GV/910GL Express Chipset at 1280×1024 resolution. In general, interactive rates are achieved in all situations. Ultimately, the large 3D model can be explored with acceptable rendering time via portal culling algorithm.

Indoor location systems are found to affect significantly the location positioning accuracy in 3D world and have been considered during the experiments. Ideally, a location service should provide correct and complete location information. The accuracy of coordinate in Cricket-based tracking system was approximately 7~12 cm with each node installed at 45 degree angle from the ceiling with distance between beacons less than 400 cm whereas 2–3 m in RSSI-based tracking system.

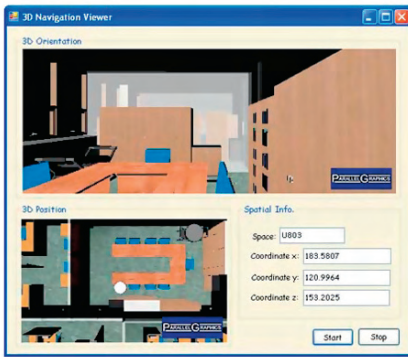
The results obtained proves location mapping is sufficiently accurate after evaluations of the indication position in 3D environment with the one estimated by indoor tracking system. Nevertheless, there is still an occasional low accuracy of the positioning due to poor precision of the received location data. Therefore, in some occasions, the visualized 3D representation did not adequately correspond to the actual location of the user in the physical world. The precision of positional data must be improved to provide a sufficient level of navigation assistance to users by the means of 3D representations.

Another related problem concerned is the viewpoint. Orientation estimation is difficult when user stops moving. To defect such insufficiency, a hybrid approach must be deployed utilizing a balance between both hardware integrated indoor tracking devices and electronic magnetic compass with 3D graphics technology to achieve the best results.

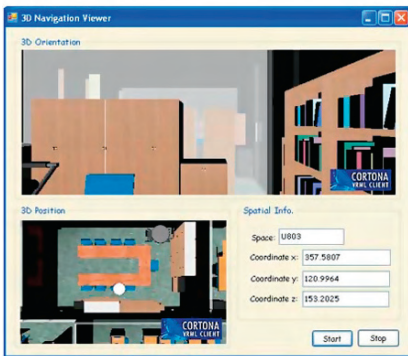
With respect to the 3D representation, the frame rate was greatly influenced by the time interval between subsequent indoor location-aware data and the complexity of the scene. In the system continuous location information can be displayed through 3D scene every 4–5 s. If user walks faster than those periods per pace in the area of a sensor, the system tends to response slow compare to the speed of movements



(a)



(b)



(c)

Fig. 11 3DNV displays and updates camera of the world which shows the viewing direction of user (*local view*), and indicates user's position (*top-down view*) in the 3D maps

in which it waits for location information from tracking systems and processes the data accordingly before deciding cell that user is currently in and render the visible objects based on the information. Apart from the speed of target, the size and shape

of locations and the coherence of the movement are parts of the underlying factors that affect the rate of location change.

Figure 11 depicts the results of 3DENV, updating 3D scenes and marking target's position and viewing direction immediately when he travels in the scene. As target navigates in the cell U803, his location information is updated accordingly.

7 Conclusion

Three dimensional graphics technology was applied to realize visualization of indoor location-aware systems. The developed system gives a good indication of how suitable 3D map could assist navigation in a complex and unknown indoor environment. Apart of its intuition, it fulfils the basic users' navigation requirements with a better visualization and navigational performance. Generally, it proves 3D representation has some benefits over the 2D maps as people could recognize the real world objects from 3D map without any additional help.

References

1. Rakkolainen I, Timmerheid J, Vainio T (2000) A 3D City Info for Mobile Users. In: Proc. 2000 International Workshop in Intelligent Interactive Assistance and Mobile Multimedia Computing, Rostock, Germany, pp 115–12
2. The Web3D Consortium, The Virtual Reality Modeling Language, ISO/IEC Std 14772-1:1997 and ISO/IEC 14772-2:2004, <http://www.web3d.org/x3d/sp-ecifications/vrml/ISO-IEC-14772-VRML97/>
3. Chung WY, Singh VK, Lim H (2006) Passive and Cost Effective Indoor Location Tracking System for Ubiquitous Healthcare. *J KIMICS* 10:1119–1123
4. Lee Y, Cha H (2006) A Light-weight and Scalable Localization Technique Using Mobile Acoustic Source. In: Proc. IEEE CIT, p 235
5. CC2431 Data Sheet, <http://www.chipcon.com>
6. Drury T (2001) Generating a Three-Dimensional Campus Model. Advanced Undergraduate Project, Massachusetts Institute of Technology
7. ParallelGraphics, VRMLPad, <http://www.parallelgraphics.com/products/vrml-pad/>
8. Cortona3D Viewer, <http://www.cortona3d.com/cortona>
9. Airey J (1990) Increasing Update Rates in the Building Walk-through System with Automatic Model-space Subdivision and Potentially Visible Set Calculations. PhD. Thesis, North Carolina University
10. Teller SJ, Sequin CH (1991) Visibility Preprocessing for Interactive Walkthroughs. In: Proc. 18th Annual Conference on Computer Graphics and Interactive Techniques, New York, USA, pp 61–70
11. Leubke D, Georges C (1995) Portals and Mirrors: Simple, Fast Evaluation of Potentially Visible Sets. In: Proc. Symposimm on Interactive 3D Graphics, New York, USA, pp 105–06
12. Marvie JE, Bouatouch K (2004) A VRML97-X3D Extension for Massive Scenery Management in Virtual Worlds. In Proc. 9th International Conference on 3D Web Technology, New York, USA, pp 145–53
13. VRML Automation Interface, <http://www.parallelgraphics.com/products/sdk/>

Part V
Ultrasonic Sensor

Ultrasonic Sensing: Fundamentals and its Applications to Nondestructive Evaluation

Ikuo Ihara

Abstract This chapter provides the fundamentals of ultrasonic sensing techniques that can be used in the various fields of engineering and science. It also includes some advanced techniques used for non-destructive evaluations. At first, basic characteristics of ultrasonic waves propagating in media are described briefly. Secondly, basic concepts for measuring ultrasonic waves are described with introductory subjects of ultrasonic transducers that generate and receive ultrasonic waves. Finally, specialized results demonstrating the capabilities of using a buffer rod sensor for ultrasonic monitoring at high temperatures are presented.

Keywords Ultrasonic sensing · transducers · nondestructive evaluation

1 Introduction

Ultrasonic sensing techniques have become mature and are widely used in the various fields of engineering and basic science. Actually, many types of conventional ultrasonic instruments, devices and sophisticated software are commercialized and used for both industrial and medical applications. One of advantages of ultrasonic sensing is its outstanding capability to probe inside objectives nondestructively because ultrasound can propagate through any kinds of media including solids, liquids and gases except vacua. In typical ultrasonic sensing the ultrasonic waves are travelling in a medium and often focused on evaluating objects so that a useful information on the interaction of ultrasonic energy with the objects are acquired as ultrasonic signals that are the wave forms variations with transit time. Such ultrasonic data provides the fundamental basis for describing the outputs of ultrasonic sensing and evaluating systems.

In this chapter the fundamentals of ultrasonic sensing techniques are described. What is ultrasound, how to produce and capture ultrasound, what kinds of methods

Ikuo Ihara

Department of Mechanical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan, e-mail: ihara@mech-nagaokaut.ac.jp

and equipments can be used to measure ultrasound, and what kinds of information can be obtained from ultrasonic measurements? These questions are addressed in the following sections and the answers to the questions are briefly explained from the viewpoint of industrial applications. In addition, some specialized results using a buffer rod sensor that is an effective means for high temperature ultrasonic measurements are introduced to demonstrate its applicability for non-destructive evaluations and monitoring. For further studies on ultrasonic sensing, it is recommended to refer to some books, [1, 2, 3, 4, 5, 6, 7] for basic theories of ultrasound propagations, [8, 9, 10, 11, 12] for transducers and instruments, and [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] for ultrasonic measurements, evaluations, applications and others.

2 Fundamentals of Ultrasound

2.1 Ultrasonic Waves in Media

It is known that frequency range of sound audible to humans is approximately 20–20,000 Hz (cycles per second). Ultrasound is simply sound that are above the frequency range of human hearing. When a disturbance occurs at a portion in an elastic medium, it propagates through the medium in a finite time as a mechanical sound wave by the vibrations of molecules, atoms or any particles present. Such mechanical waves are also called elastic waves. Ultrasound waves or ultrasonic waves are the terms used to describe elastic waves with frequency greater than 20,000 Hz and normally exist in solids, liquids, and gases. A simple illustration of the ultrasonic waves produced in a solid is shown in Fig. 1, where distortion caused depending on whether a force is applied normal or parallel to the surface at one end

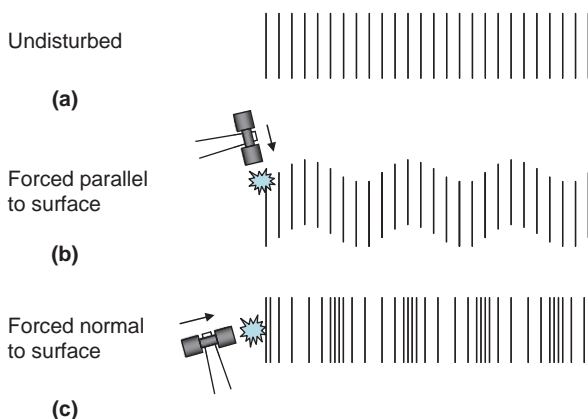


Fig. 1 Schematics of ultrasonic waves in a bulk specimen: (a) equilibrium state with no disturbance, (b) waves relating to shear (transverse) vibrations, (c) waves relating to longitudinal vibrations

of the solid can result in producing compression or shear vibrations, respectively, so that two types of ultrasonic waves, i.e. longitudinal waves or transverse waves, propagate through the solid. The energy of the wave is also carried with it.

In a continuous medium, the behaviour of ultrasonic waves is closely related to a balance between the forces of inertia and of elastic deformation. An ultrasonic wave moves at a velocity (the wave velocity) that is determined by the material properties and shape of the medium, and occasionally the frequency. The ultrasonic wave imparts motion to the material when it propagates. This is referred to as particle motion, to distinguish it from the wave motion. This particle motion is usually specified as a particle velocity v . It is noted in ultrasonic measurements that the particle velocity is much smaller than wave velocity. Also, one can understand that no ultrasonic wave propagates in vacua because there are no particles that can vibrate in vacua.

The balance between inertia and elasticity develops into a linear relationship between stress σ and particle velocity v , $\sigma = zv$. The proportional factor z is called the specific acoustic impedance of an ultrasonic wave [6, 7, 8, 9, 10, 11, 12, 13]

$$z = \sigma/v = \rho c \quad (1)$$

where, ρ is the density, and c is the wave velocity. The acoustic impedance characterizes the ability of a material to vibrate under an applied force and can be considered as the resistance of the material to the passage of ultrasonic waves. There is an analogy between impedance in electrical circuits and the acoustic impedance. The acoustic impedance is useful for treating the transmission of ultrasonic waves between two media, just like that the electrical impedance is effective to characterize a resistance in an alternating electric current circuit. For example, the transmission of an ultrasonic wave from one medium to another becomes maximum when the acoustic impedances of the two media are equal. The concept of using the acoustic impedance plays an important role in determining of acoustic transmission and reflection at a boundary of two media having different material properties and therefore, the acoustic impedance is an important parameter in designing ultrasonic sensors and sensing systems.

In Fig. 1, ultrasonic waves propagating across the material is simply shown in terms of the displacement of the layers from their equilibrium position and its amplitude. At a fixed position in the material, the displacement changes sinusoidally with time t , where the time required for the wave to propagate the distance between successive maxima is the period T . At any time, the amplitude of the displacement decreases periodically with increasing propagation distance because of its attenuation by the material. The distance between successive maxima in the amplitude variation is equal to the wavelength λ .

2.2 Features of Ultrasonic Waves

It is important to understand the behaviour and properties of ultrasonic waves in media, to design ultrasonic sensors and develop ultrasonic sensing systems. Some basic features of ultrasonic waves are introduced here.

2.2.1 Types of Wave (Modes of Propagation)

What types of ultrasonic waves can exist? The answer to this question can basically be given from solutions of the wave equations that predict wave behaviours by showing that material properties and body shape dictate the vibrational response to the applied forces that drive the wave motion. Details of wave types obtained by solving wave equations and their characteristics are shown in [1, 2, 3, 4, 5, 6, 7]. In short, there are two types of ultrasonic waves: bulk (fundamental) waves that propagate inside of an object, and guided waves that propagate near the surface or along the interface of an object [4, 5, 6, 7].

Waves that propagate wholly inside an object, independent of its boundary and shape, are called bulk waves. Two types of bulk waves can exist in an isotropic medium: longitudinal (or dilatational, compression, primary), and shear (or distortional, transverse, secondary) waves as shown schematically in Fig. 1. As mentioned in Sect. 2.1, ultrasonic wave propagations are usually described in terms of the direction of particles motion in relation to the direction in which the wave propagates. The longitudinal waves can be defined on this basis as waves in which the particle motion is parallel to the direction of the wave propagation. The shear waves are defined as waves in which the particle motion is perpendicular to the direction of the propagation. Both waves can exist in solids because solids, unlike liquids and gasses, have rigidity that is a resistance to shear as well as compressive loads. However, the shear waves cannot exist in liquids and gasses because of no resistance to shear loads in such media.

When the influences of the boundaries or shape of an object are considered, other types of waves called the guided waves are produced. There are three types of guided waves depending on geometry of an object: surface acoustic waves (SAWs), plate waves, and rod waves.

SAWs are defined as waves that propagate along a free surface, with disturbance amplitude that decays exponentially with depth into the object. There are many kinds of SAWs such as Rayleigh, Scholte, Stoneley, and Love waves and the wave propagation characteristics of SAWs strongly depend on material properties, surface structure, and nature at the interface of the object. When an SAW propagates along a boundary between a semi-infinite solid and air, the wave is often called Rayleigh wave in which the particle motion is elliptical and the effective penetration depth is of the order of one wavelength. Among many types of SAWs, Rayleigh wave is the most common and well-known wave so that many researchers often call any SAWs Rayleigh wave.

When an ultrasonic wave propagates in a finite medium (like a plate), the wave is bounded within the medium and may resonate. Such waves in an object of finite size are called plate waves if the object has a multilayer structure, and called Lamb waves if it has a single layer. Also, when a force is applied to the end of a slender rod, an ultrasonic wave propagates axially along it. Wave propagations in rodlike structures such as a thin rod and hollow cylinders have been studied extensively. Further information on the guided waves and their characteristics can be obtained in [4, 5, 6, 7, 20]. In general, the wave propagation characteristics of guided waves

strongly depend on not only material properties but also the plate thickness, the rod diameter, and the frequency. The frequency dependence of the wave velocity of guided waves is called frequency dispersion. While the frequency dispersion often makes wave propagation behaviour complicated, it also provides unique materials evaluations using guided waves. It is noted that similar types of bulk and guided waves can exist for anisotropic materials and in general, their behaviours become much more complicated than those for isotropic materials [5, 6, 7].

2.2.2 Velocity

Ultrasonic velocity is probably the most important and widely used parameter in ultrasonic sensing applications. Each medium has its own value of the velocity that usually depends on not only propagation medium but also its geometrical shape and structure. The theoretical values can be obtained from wave equations and typically determined by the elastic properties and density of the medium. For example, the wave equations for an isotropic solid give the following simple formulae for the longitudinal and shear wave velocities

$$v_l = \sqrt{\frac{E}{\rho} \cdot \frac{1-\nu}{(1+\nu)(1-2\nu)}} \quad (2)$$

$$v_s = \sqrt{\frac{E}{\rho} \cdot \frac{1}{2(1+\nu)}} = \sqrt{\frac{G}{\rho}} \quad (3)$$

where, v_l and v_s are the longitudinal and shear wave velocities, respectively, E is Young's modulus, ν is Poisson's ratio, G is shear modulus and ρ is the density. For most of solid materials the longitudinal wave velocity is faster than the shear wave velocity because the shear modulus is lower than the Young's modulus. It is noted that Poisson's ratio is not a dominant factor affecting the velocities. As a rule of thumb, the velocity of the shear wave is roughly half the longitudinal wave. Although the velocities can be determined theoretically if material properties such as the elastic moduli and density are known precisely, these material properties are not always available for the determination because they change depending on mechanical processing and heat treatments. Therefore, it is important and necessary to make a calibration measurement for the velocities when one wants to know the correct values for velocities.

2.2.3 Attenuation

When an ultrasonic wave propagates through a medium, ultrasonic attenuation is caused by a loss of energy in the ultrasonic wave and other reasons. The attenuation can be seen as a reduction of amplitude of the wave. There are some factors

affecting the amplitude and waveform of the ultrasonic wave, such as ultrasonic beam spreading, energy absorption, dispersion, nonlinearity, transmission at interfaces, scattering by inclusions and defects, Doppler effect and so on. To characterize the ultrasonic attenuation quantitatively, attenuation coefficient α is defined as follows

$$A = A_0 \cdot e^{-\alpha x} \quad (4)$$

where A is the peak amplitude of the wave at propagation distance x , A_0 is the initial peak amplitude. The attenuation coefficient α is experimentally determined from the variation of the peak amplitude with the propagation distance, and it can be given in decibel per metre (dB/m) or in neper per metre (Np/m). In general, the attenuation coefficient highly depends on frequency. Since this frequency dependence reflects microstructures of materials, it can be used for characterizing microscopic material properties relating to chemical reactions and mechanical processes. Further information on the attenuation can be obtained in [7, 9, 10, 12, 13].

2.2.4 Wavelength

Wavelength λ is the distance over which one spatial cycle of the wave completes and the following expression can be given,

$$\lambda = v/f \quad (5)$$

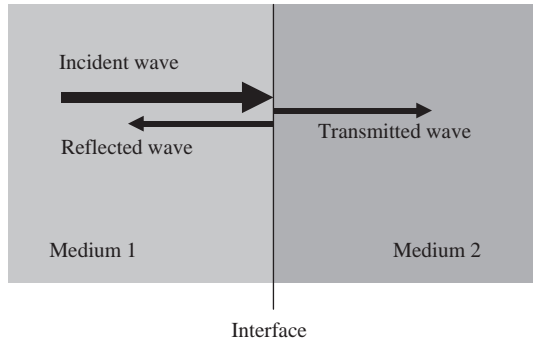
where v is the ultrasonic velocity and f is the frequency. Wavelength is a useful parameter in ultrasonic sensing and evaluations. In ultrasonic detection of a small object, the smallest size that can clearly be detected must be larger than half a wavelength at the operating frequency. If the critical size of an object to be detected is known, such prior information on size is helpful for selecting an appropriate frequency for measurements.

2.2.5 Reflection and Transmission

When an ultrasonic wave perpendicularly impinges on an interface between two media as shown in Fig. 2, a part of the wave is reflected back to the medium 1 and the remainder is transmitted to the medium 2. The ratio of the amplitude of the reflected wave A_R to that of the incident wave A_I is called reflection coefficient R , and the ratio of the amplitude of the transmitted wave A_T to that of the A_I is called transmission coefficient T . Considering a balance of stresses and a continuity of velocities on both sides of the interface, the reflection and transmission coefficients, R and T can be given as follows

$$R = \frac{A_R}{A_I} = \frac{z_1 - z_2}{z_1 + z_2} \quad (6)$$

Fig. 2 Normal reflection and transmission at an interface between two media



$$T = \frac{A_T}{A_I} = 2 \cdot \frac{z_1}{z_1 + z_2} \tag{7}$$

where subscripts 1 and 2 refer to the medium 1 and 2, respectively, and z is the acoustic impedance defined as (1). It can be seen from these equations that the maximum transmission of ultrasonic wave occurs when the impedances of the two media are identical, and most of ultrasonic wave is reflected when the two media have very different impedances. The reflection and transmission at interface play an important role in designing ultrasonic sensing systems and understanding experimental results with the ultrasonic systems.

2.2.6 Refraction and Mode Conversion

When an ultrasonic wave obliquely impinges on an interface between two media as shown in Fig. 3, several things happen depending on the incident angle of the wave as well as the material properties of the two media. One of important things is refraction in which a transmitted wave has a different angle from the incident. The refraction is basically caused by the velocity difference on either side of the interface. The refraction angle can be calculated from Snell’s law [19] if the velocities of the two media and the incidence angle are known.

Another important phenomenon is mode conversion that is a generation of one type of wave from another type in refraction as shown in Fig. 3. For example, a longitudinal wave incident on an interface between liquid and solid is transmitted partially as a refracted longitudinal wave and partially as a mode converted shear wave in the solid. Mode conversion can also take place on reflection if the liquid shown in Fig. 3 is a solid. It is noted that any types of waves can be converted to another type, e.g. from a shear wave to a longitudinal wave, and from a longitudinal wave to a surface wave. The angles of reflection and/or refraction by mode conversion can be calculated from Snell’s law.

Figure 4 shows a simulation result for refraction and mode conversion, calculated by a finite difference method. We can see that an incident plane wave (longitudinal wave) of 10° in water is refracted at the refraction angle of 43° in steel and simultaneously converted to shear wave at refraction angle of 22° .

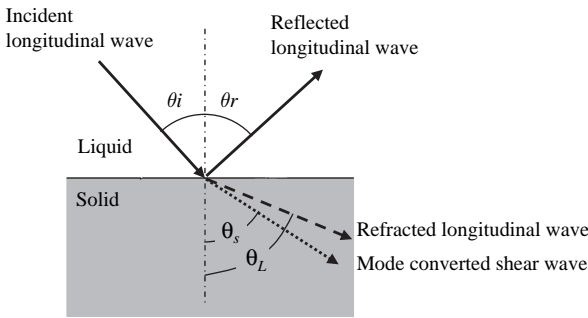


Fig. 3 Schematics of reflection, refraction and mode conversion at an oblique interface

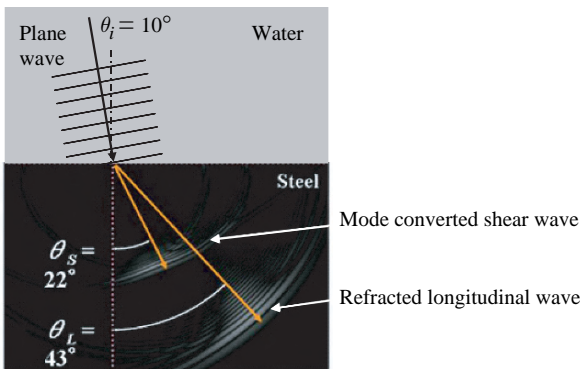


Fig. 4 A simulation result for refraction and mode conversion

3 Measurement of Ultrasound

3.1 Generation and Detection of Ultrasonic Waves

3.1.1 Transducers

Ultrasonic sensors are often called transducers. The function of the transducers is to convert electrical energy into mechanical energy which directly corresponds to ultrasonic vibration, and vice versa. The most common way of generating and detecting ultrasonic waves utilizes the piezoelectric effect of a certain crystalline material such as quartz. Since the piezoelectric effect is reciprocal, it produces a deformation (a mechanical stress) in a piezoelectric material when an electrical voltage is applied across the material, and conversely, it produces an electrical voltage when a deformation (a mechanical stress) is applied to the material. Thus, the piezoelectric materials can be used for generating and detecting ultrasonic waves that are related to the mechanical stresses. Appropriate cuts and directions of quartz are utilized for two types of waves, longitudinal and shear, as shown in Fig. 5. Nowadays,

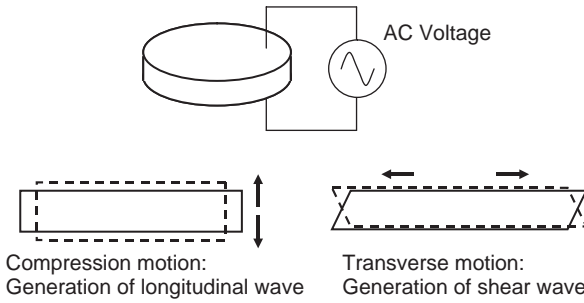


Fig. 5 Response of a piezoelectric plate to an alternating voltage

many piezoelectric materials besides quartz are available, such as barium titanate (BaTiO_3), lead metaniobate (PbNb_2O_3) and lead zirconate titanate (PZT), etc. The size and shape of piezoelectric transducers have to be precisely designed depending on the desired frequency. For industrial applications, solid-state transducers are usually used, because of their robustness. A piezoelectric transducer consists of a piezoelectric element, electrical connections, backing materials, front layers and a casing. The typical construction is shown in Fig. 6. The front layer is to protect the piezoelectric element against external stresses and environmental influences, and also must function as an impedance matching layer with which the transfer of ultrasonic energy to the target medium is optimized. The backing material functions as a damping block that alters the resonance frequency of the piezoelectric element and deletes unwanted ultrasonic waves reflected from the back wall. The electrical line is connected AC or DC voltage supplies that are often operated at the resonant frequency of the piezoelectric element.

Depending on applications, other types of transducers can be available. Piezoelectric polymers that can exhibit the piezoelectric effect, often called PVDF (polyvinylidene fluoride), have some advantages owing to polymer characteristics

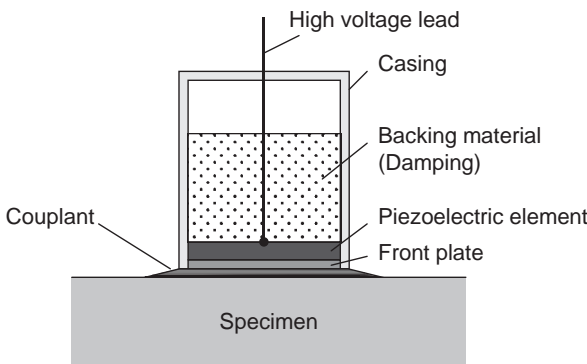


Fig. 6 Typical construction of a piezoelectric transducer and its use in measurement of a solid specimen

such as its low acoustic impedance and softness. Magnetostriction effect that occurs in ferromagnetic materials is also utilized as transducers in industries.

It should be noted that the piezoelectric and magnetostrictive effects generally decrease with a rise in temperature and disappears at the Curie temperature. This is a crucial limitation in use of the ultrasonic transducers. When ultrasonic measurements are conducted at high temperatures near the Curie temperature, precautions are necessary so that the ultrasonic transducer does work properly. One of methods for high temperature measurements and its applications are presented in Sect. 4. It is also noted in the use of the transducers mentioned above that it is necessary to use some coupling medium for making an effective ultrasonic energy transmission between the transducer and specimen, as shown in Fig. 6. Gels, liquids or grease are often used as a coupling medium. It is extremely difficult to conduct the ultrasonic measurements without such coupling medium because of any air gap or large acoustic impedance between the transducer and specimen surface. This is another disadvantage of using contact-type transducers. Further information on transducers can be obtained in [10, 11, 12, 13].

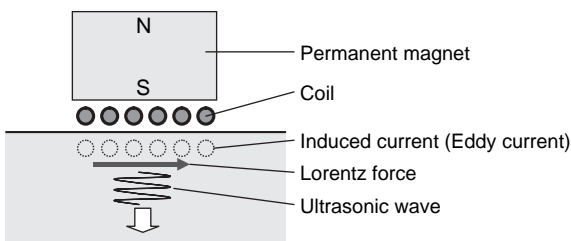
3.1.2 Non-contact Techniques

Non-contact ultrasonic measurements are of great practical interests in the many fields of engineering. There are three kinds of non-contact methods for generation and detection of ultrasonic waves: optical method, electromagnetic method, and air-coupled method. Although each method has advantages and disadvantages, they have the potential to be powerful diagnostic tools for advanced ultrasonic sensing.

Optical methods for measuring ultrasonic waves are called laser-ultrasonics in which ultrasonic waves are generated and detected by using lasers. Laser generation of ultrasonic waves can be recognized as exciting the waves with an optical hammer. When a high energy pulsed laser beam is irradiated onto a specimen surface, an interaction of the laser beam with the specimen occurs in one or both of two distinct processes, thermoelastic and ablative. By controlling the laser irradiation conditions, it is possible to generate any types of ultrasonic waves such as longitudinal, shear and guided waves at a desired frequency. To detect ultrasonic waves, a laser beam is illuminated onto the specimen surface for the duration sufficiently long to capture the ultrasonic signal of interest. Ultrasonic waves are then detected by measuring surface displacements caused by ultrasonic disturbance, using an laser-assisted interferometer or other device. Mickelson, Confocal Fabry-Perot or Photorefractive Two-wave Mixing interferometers are often utilized. The ability of laser-ultrasonics to operate at large standoff distances provides big advantages in industrial applications such as materials process monitoring at high temperatures. Further information on laser ultrasonics can be obtained in [21].

Electromagnetic acoustic transducer (EMAT) is an alternative technique for generating and receiving ultrasonic waves, with which the ultrasonic measurements are conducted without any coupling medium between the transducer and specimen. The EMAT consists of a stack of coils and magnets to generate and receive ultrasonic

Fig. 7 Schematic of generation of an ultrasonic wave using an EMAT



waves in an electrically conductive material as shown in Fig. 7. When a coil that is placed near to the surface of a specimen is driven by a pulse current with a desired ultrasonic frequency, eddy currents will be induced by electromagnetic induction in near surface region of the specimen. Since a static magnetic field is present, the eddy currents will experience Lorentz forces F of the following form

$$F = J \times B \tag{8}$$

where J is the induced eddy currents and, B is the static magnetic field. Interactions of the Lorentz forces with the specimen produce high frequency vibrations resulting in generating ultrasonic waves. Since these processes are reciprocal, the same mechanisms work to allow the ultrasonic energy to be converted into electromagnetic energy, so that the EMAT works as a receiver as well as a generator. The EMAT eliminates the problems associated with the coupling medium because the electro-mechanical conversion takes place directly within the electromagnetic skin depth of the specimen surface. Thus, EMATs allow non-contact ultrasonic sensing for moving specimens, rough surfaces, in vacuum and also in hazardous locations. Further information on EMATs can be obtained in [22, 23].

Another method for non-contact ultrasonic sensing is air-coupled ultrasonics. In air-coupled ultrasonics, air is used as a coupling medium between the transducer and specimen. Although air-coupling is very attractive, it has some difficulties because of high attenuation coefficient of air and high impedance mismatch between a transducer and air. To overcome such problem, a specially designed transducer with an optimal impedance matching layer is required for air-coupled ultrasonic measurements. Some piezoelectric-type air-coupled transducers have been commercialized and used for non-contact inspections. However, most of them have relatively low and narrow band frequency response with which it may not be sufficient to be used in a wide variety of applications. Recently, micro-electromechanical systems (MEMS) technology has applied to ultrasonic sensors. A capacitive type air-coupled transducer, consisting of a metallized insulating polymer film placed upon a contoured conducting backplate, is developed using semiconductor manufacturing techniques [24]. This provides effective air-couple measurements with a higher and wider band frequency, in the range 100 kHz to 2 MHz. Utilizing such advantage, a novel noncontact method for characterizing surface roughness of materials by air-coupled ultrasound is developed [25].

3.2 Basics of Instrumentation

Figure 8 shows a block diagram of a basic construction of an ultrasonic measurement system used to generate and detect ultrasonic waves in a specimen. The synchronization generator gives trigger signals with high repetition rate (e.g. 1000 repetitions per second) to the pulse generator (pulser). Using these triggers, the pulser provides electrical voltage to the transducer so that the transducer generates ultrasonic waves at the same repetition rate. The reflected ultrasonic waves through the specimen are received by the same transducer and the resulting voltage of the received waves goes to the display through the amplifier. The computer is often used to analyze the acquired ultrasonic data.

Figure 9 shows typical configurations for transducers used to launch and receive ultrasonic waves for ultrasonic measurements. Pulse-echo configuration with a single transducer shown in Fig. 9(a) is most commonly used to measure reflected waves from a flaw or the opposite side of the specimen. Through-transmission with a two transducers shown in Fig. 9(b) is probably the second most commonly used

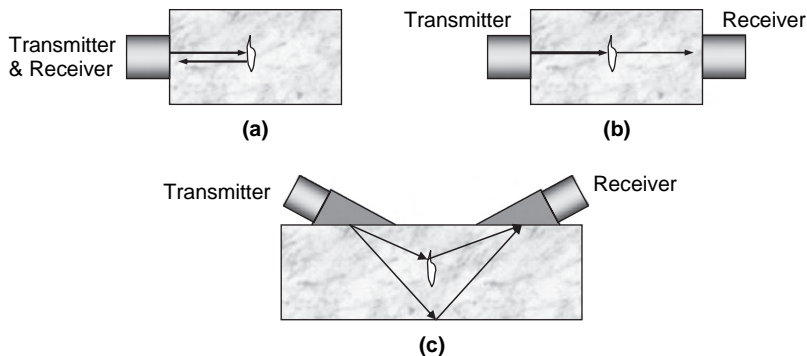
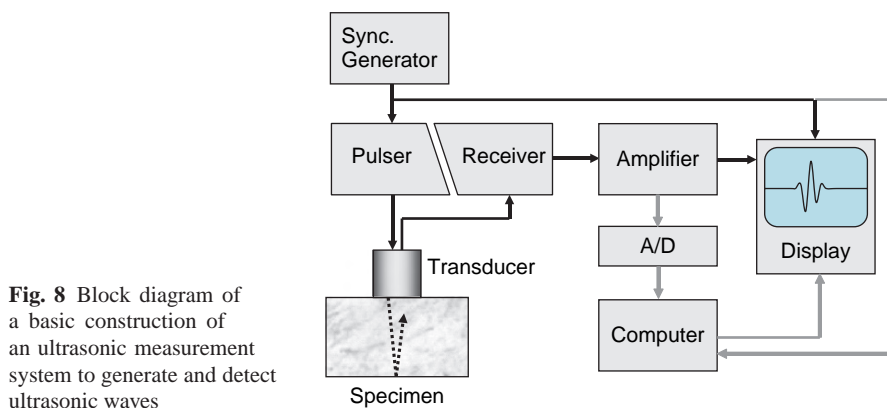


Fig. 9 Typical configurations of transducers used in ultrasonic measurements

configuration. The third one is so-called pitch-catch configuration in which two transducers are placed on the same side of the specimen as shown in Fig. 9(c). This can be useful in the cases that the back wall is not parallel to the front wall or there is difficulty to use normal incidence ultrasonic beams.

In general, an ultrasonic transducer operating at a high frequency radiates a narrow ultrasonic beam into a medium, which results in sensing over a narrow spatial region. To cover a wider region in ultrasonic sensing, scanning techniques are often used. Another powerful solution to probe a wide area is to use transducer arrays that are typically composed of number of individual transducer elements. A one-dimensional (linear) array or a two-dimensional array are commercialized and commonly used in the medical field for imaging. These transducer elements are arranged in certain patterns for the purpose of dynamic focusing or steering ultrasonic waves, using a beam forming effect based on wave interference. The elements configuration is designed to be able to form the desired beam shape and direction of ultrasonic wave. Phased-array transducers that provide a two-dimensional or a three-dimensional images in a medium are developed for performing a reliable flaw detection. Further information on the ultrasonic instrumentation can be obtained in [9, 11, 12, 13].

A general scheme of ultrasonic based measurements and the related aspects are depicted in Fig. 10.

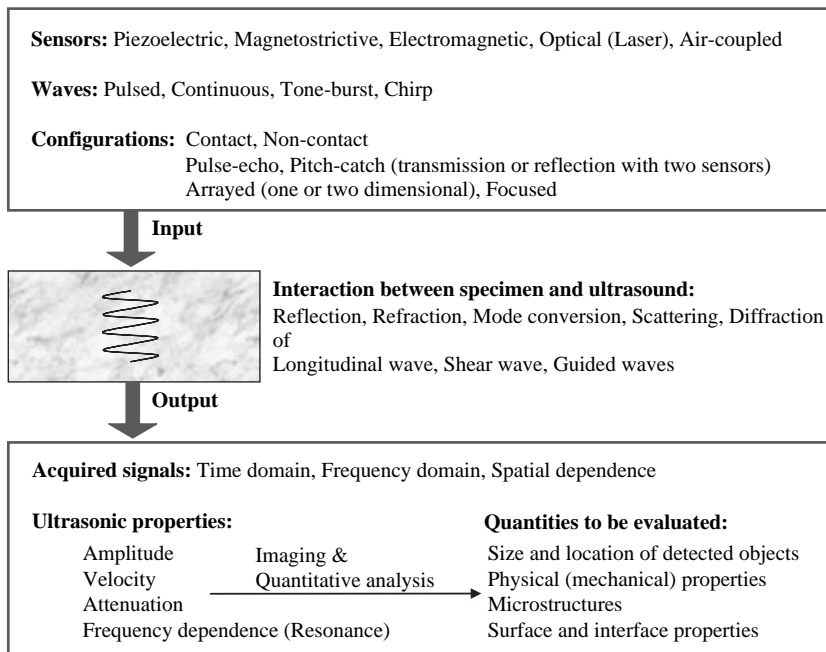


Fig. 10 General scheme of ultrasonic based measurements and evaluations

4 Applications to Nondestructive Evaluation

Ultrasonic sensors have widely been used for numerous sensing applications in the fields of engineering, physics as well as medical science. Although the ultrasonic techniques have been applied to various nondestructive evaluations such as inspections of industrial structures, quantitative characterizations of materials and structural health monitoring [12, 13, 14, 15, 16, 17, 18, 19, 20], it is still required to develop new and more effective techniques that are applicable to advanced nondestructive evaluations. One of industrial demands is to realize ultrasonic in-line monitoring in a hazardous environment such as high temperatures. In this section, recent advances showing the capabilities of using buffer rod sensors as nondestructive tool for high temperature monitoring are presented.

4.1 Buffer Rod Sensors for High Temperature Monitoring

There are several ways for ultrasonic sensing at high temperatures: laser ultrasonics, EMATs, high temperature transducers and buffer rod method (known as delay-lines or waveguides). Since each technique has advantages and disadvantages, one has to select the appropriate technique to suit the objective depending on the application. Among the techniques, buffer rod method is a classical and still an attractive approach because of its simplicity and low cost. For high temperature applications of the buffer rod method, a long buffer rod is often employed as a waveguide. A conventional piezoelectric transducer is installed to the one end of the buffer rod and the other end is in contact with the material to be measured.

The difficulty in ultrasonic measurements using a buffer rod is, in most cases, caused by spurious echoes due to interference of mode converted waves, dispersion, and diffraction within the rod of finite diameter. These spurious echoes deteriorate the signal to noise ratio (SNR) because of their possible interference with desired signals to be measured. To overcome such difficulty, tapered and clad buffer rods are developed for various applications in materials evaluations and monitoring [26, 27, 28, 29, 30, 31]. Figure 11 shows the exterior of one of the developed buffer rod sensors, consisting of a tapered clad buffer rod, a cooling pipe and a conventional ultrasonic transducer (UT). The transducer end of the buffer rod is air cooled so that conventional room temperature UTs can be used while the other end (probing end)

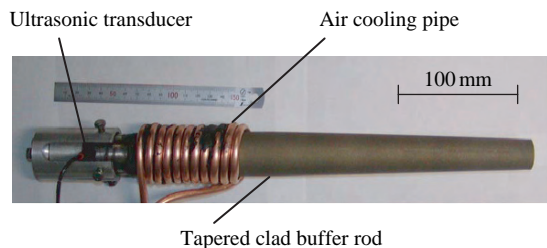


Fig. 11 External view of a buffer rod sensor for high temperature use

is in contact with a hot medium at 800°C. Because of a taper shape of the buffer rod and a cladding layer of the outer surface, the buffer rod provides high performance pulse-echo measurements with high SNR at high temperatures. The length of the rod is possible to be up to 1000 mm.

4.2 Imaging Using Focused Sensors

To provide high spatial resolution measurements, a spherical concave surface is machined at the probing end of the rod as shown in Fig. 12(a). This is expected to function as an acoustic lens for generating and receiving focused ultrasonic waves. Figure 12(b) shows a contour plot of the acoustic field in the vicinity of a focal zone in molten aluminium at 800°C, where the acoustic field is numerically examined by finite difference method [30]. We can see that the ultrasonic wave can be focused onto a small area comparable to a wavelength (460µm) so that it is expected to make high resolution measurements using the focused buffer rod sensor. It is experimentally verified that the developed focused sensor can successfully detect alumina particles of about 160µm suspended in molten aluminium [30]. Figure 13

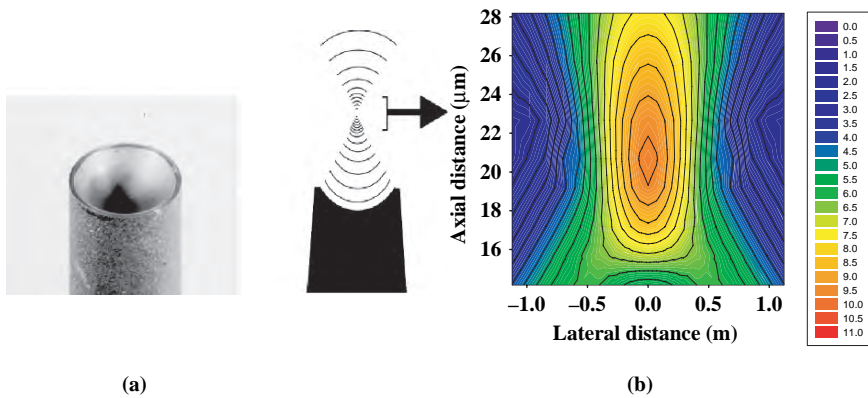


Fig. 12 (a) a concave acoustic lens fabricated at the probing end of a buffer rod sensor, (b) a simulation result of the sound field of focused ultrasonic wave at 10 MHz in molten aluminium [30]

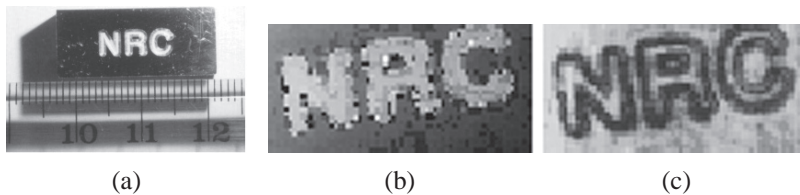
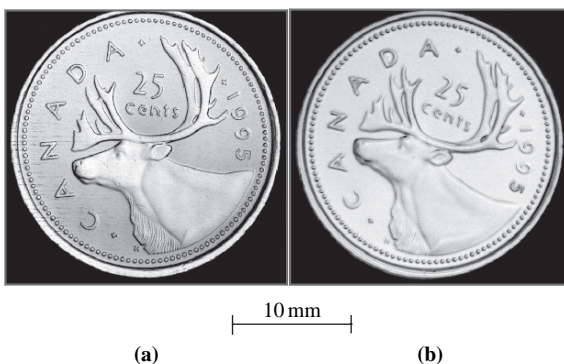


Fig. 13 Ultrasonic images in molten zinc at 800°C: (a) specimen having the three letters NRC engraved on the surface, (b) by plotting the time delay of the echo, (c) by plotting the amplitude of the echo [29]

Fig. 14 Ultrasonic images of a Canadian quarter obtained using the (a) short and (b) long buffer rods with acoustic lens in water [28]



shows ultrasonic images obtained in molten zinc at 650°C, by scanning of a focused buffer rod sensor [29]. This is probably the first ever image in a molten metal. Surprisingly, this kind of imaging is possible even using a long buffer rod of 1 m length. Figure 14 shows the images obtained in water using a short rod of 75 mm and a long rod of about 1000 mm [28]. Although the resolution of the image using the long rod deteriorates because of an attenuation of higher frequency components of the guided wave in the rod, it can be seen that the ultrasonic wave can be focused onto a small spot of about one wavelength.

4.3 In-Situ Monitoring of Solid-Liquid Interface

Using the buffer rod sensor, an attempt has been made to monitor a solid-liquid interface of aluminium alloy during unidirectional solidification at 700°C [31]. A solid-liquid interface of aluminium alloy is produced using a directional solidification furnace and then the interface behaviour is monitored during heating and cooling as shown in Fig. 15(a). Figure 15(b) shows the location of the interface determined from the transit time of ultrasonic pulse echo. The growing rate of the solidification front is estimated to be 0.12 mm/s by time-differentiating the location. The amplitude change of the interface echo is also shown in Fig. 15(b). We can observe periodical oscillations in the amplitude during heating and cooling. It is tentatively considered that these oscillations are related to the feature of solidification instabilities such as variations in cellular structure and/or mushy zone consisting of solid and liquid phases.

4.4 Monitoring of Internal Temperature Distribution

In many fields of science and engineering, there are growing demands for measuring internal temperature distribution of heated materials. Recently, an ultrasonic

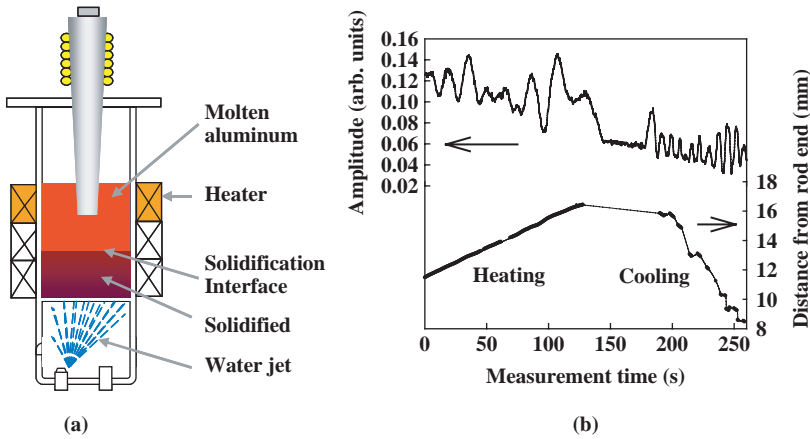


Fig. 15 (a) Schematic of experimental setup for ultrasonic monitoring of solid-liquid interface of aluminium alloy using a buffer rod sensor, (b) Monitoring result showing variations in amplitude and location of solid-liquid interface echo during heating and cooling [31]

method has been applied to internal temperature monitoring [32]. The principle of the method is based on temperature dependence of ultrasonic velocity in materials. A single side of a silicone rubber plate of 30 mm thickness is heated by contacting with a hot steel plate as shown in Fig. 16(a) and ultrasonic pulse-echo measurements are then performed during heating. A change in the transit time of ultrasonic wave in the heated rubber is monitored and used to determine the transient variation of internal temperature gradient in the rubber, where an inverse analysis is used to determine one-dimensional temperature gradient. Figure 16(b) shows the internal temperature distributions in the silicone rubber and their variations with elapsed time.

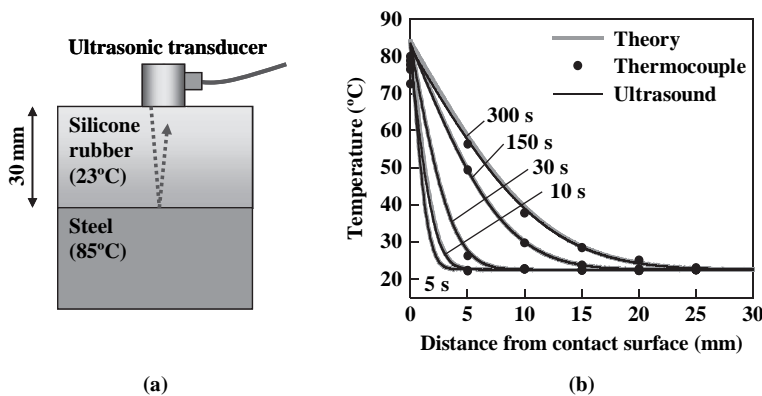


Fig. 16 (a) Schematic of ultrasonic temperature monitoring of a silicone rubber being heated, (b) Monitoring result showing internal temperature distributions in the silicone rubber and their variations with elapsed time [32]

The temperature gradient determined ultrasonically agrees well with both obtained using commercial thermocouples installed in the rubber and estimated theoretically.

Thus, recent demonstrations shown in this section reveal that even a classical method such as a pulse-echo method using a buffer rod sensor has the high potential to be applicable to a novel sensing in an unexplored field.

5 Conclusion

In this chapter a brief overview of fundamentals in ultrasonic sensing is presented. Some advanced techniques and applications to nondestructive evaluation are also introduced. The essentials of ultrasonic sensing are how to drive an ultrasonic wave into an object and how to capture the ultrasonic wave from the object. In addition, another essential is how to extract the information we want from the captured ultrasonic wave. To accomplish these and to create a useful sensing technique, it is indispensable to make an effective collaboration among researchers in different fields of engineering and science such as electrical, electronics, information, mechanical and materials. Actually, progress is being made in ultrasonic sensing technology, but, it should be noted that classical techniques and methods are still attractive and have the potential to create something new, as shown in the application of a buffer rod sensor.

References

1. H. Kolsky (1963) *Stress Waves in Solids*, Dover Publications, New York.
2. W. C. Elmore and M. A. Heald (1985) *Physics of Waves*, Dover Publications, New York.
3. D. Royer and E. Dieulesaint (2000) *Elastic Waves in Solids I & II*, Springer-Verlag, Berlin.
4. L. M. Brekhovskikh (1980). *Waves in Layered Media* 2nd Edition, Academic press, New York.
5. J. D. Achenbach (1990) *Wave Propagation in Elastic Solids*, Elsevier Science Publisher, Amsterdam.
6. B. A. Auld (1990) *Acoustic Fields and Waves in Solids* 2nd Edition Vol. 1 & 2, Krieger Publishing, Florida.
7. J. L. Rose (1999) *Ultrasonic Waves in Solid Media*, Cambridge University Press, Cambridge.
8. G. S. Kino (1987) *Acoustic Waves Devices, Imaging and Analog Signal Processing*, Prentice-Hall, New Jersey.
9. R. N. Thurston and A. D. Pierce (Editors) (1999) *Ultrasonic Instruments and Devices I & II*, Academic Press, San Diego.
10. A. Arnau (2004) *Piezoelectric Transducers and Applications*, Springer-Verlag, Berlin.
11. E. P. Papadakis (Editor) (1999) *Ultrasonic Instruments & Devices*, Academic Press, San Diego.
12. R. N. Thurston and A. D. Pierce (Editors) (1990) *Ultrasonic Measurement Methods*, Academic Press, San Diego.
13. J. Krautkramer and H. Krautkramer (1990) *Ultrasonic Testing of Materials* 4th Revised Edition, Springer-Verlag, Berlin.
14. A. Briggs (1992) *Acoustic Microscopy*, Clarendon Press, Oxford.

15. M. Levy, H. E. Bass, and R. Stern (Editors) (2001) *Modern Acoustical Techniques for the Measurement of Mechanical Properties*, Academic Press, San Diego.
16. T. Kundu (Editor) (2004) *Ultrasonic Nondestructive Evaluation*, CRC Press, Boca Raton.
17. D. R. Raichel (2006) *The Science and Applications of Acoustics 2nd Edition*, Springer Science+Business Media, New York.
18. L. W. Scherrer Jr. and S.-J. Song (2007) *Ultrasonic Nondestructive Evaluation Systems*, Springer Science+Business Media, New York.
19. B. M. Lempriere (2002) *Ultrasound and Elastic Waves: Frequently Asked Questions*, Academic Press, San Diego.
20. K. F. Graff (1991) *Wave Motion in Elastic Solid*, Dover Publications, New York.
21. J.-P. Monchalin (2007) *Laser-Ultrasonics: Principles and Industrial Applications*, in *Ultrasonic and advanced Methods for Nondestructive Testing and Materials Characterization*, chapter 4, edited by C. F. Chen, World Scientific, New Jersey, pp. 79–115.
22. H. M. Frost (1979) *Electromagnetic-Ultrasonic Transducers: Principles, Practice, and Applications: Physical Acoustics XIV*, edited by W. P. Mason and R. N. Thurston, Academic Press, New York, pp. 179–270.
23. M. Hirao and H. Ogi (2003) *EMATS for Science and Industry*, Kluwer Academic Publishers, Boston.
24. D. W. Schindel, D. A. Hutchins, L. Zou, and M. Sayer (1995) *The Design and Characterization of Micromachined Air-Coupled Capacitance Transducers*, *IEEE Trans. Ultrason. Ferroelec. Freq. Control.* UFFC-42: 42–50.
25. D. D. Sukmana and I. Ihara (2007) *Quantitative Evaluation of Two Kinds of Surface Roughness Parameters Using Air-Coupled Ultrasound*, *Jpn J. App. Phys.*, 46(5B): 4508–4513.
26. C.-K. Jen., J. G. Legoux, and L. Parent (2000) *Experimental Evaluation of Clad Metallic Buffer Rods for High Temperature Ultrasonic Measurements*, *NDT & E International* 33, pp. 145–153.
27. C.-K. Jen, D. R. França, and Z. Sun, and I. Ihara (2001) *Clad Polymer Buffer Rods for Polymer Process Monitoring*, *Ultrasonics*, 39(2): 81–89.
28. I. Ihara, C.-K. Jen, and D. R. França (1998) *Materials Evaluation Using Long Clad Buffer Rods*, *Proc. IEEE Int. Ultrasonics Symp.*, Sendai, pp. 803–809.
29. I. Ihara, C.-K. Jen, and D. R. França (2000) *Ultrasonic Imaging, Particle Detection and V(z) Measurements in Molten Zinc Using Focused Clad Buffer Rods*, *Rev. Sci. Instrum.* 71(9): 3579–3586.
30. I. Ihara, H. Aso, and D. Burhan (2004) *In-situ Observation of Alumina Particles in Molten Aluminum Using a Focused Ultrasonic Sensor*, *JSME International Journal*, 47(3): 280–286.
31. I. Ihara, D. Burhan, and Y. Seda (2005) *In situ Monitoring of Solid-Liquid Interface of Aluminum Alloy using a High Temperature Ultrasonic Sensor*, *Jpn J. App. Phys.*, Vol.44(6B): 4370–4373.
32. M. Takahashi and I. Ihara (2008) *Ultrasonic Monitoring of Internal Temperature Distribution in a Heated Material*, *Jpn J. App. Phys.*, Vol.47(5B): 3894–3898.

Part VI
Image Sensor

Multimodal Image Sensor Fusion Using Independent Component Analysis

Nedeljko Cvejic, Nishan C. Canagarajah and David R. Bull

Abstract In this chapter, we present a novel multimodal image fusion algorithm using the Independent Component Analysis (ICA). Region-based fusion of ICA coefficients is implemented, in which the mean absolute value of ICA coefficients is used as an activity indicator for the given region. The ICA coefficients from given regions are consequently weighted using the Piella fusion metric in order to maximise the quality of the fused image. The proposed method exhibits significantly higher performance than the basic ICA algorithm and improvement over the other state-of-the-art algorithms.

Keywords Data fusion · image fusion · independent component analysis · image fusion metrics

1 Introduction

A relatively lower level of interest in infrared imagery, compared to visible imagery, has been due to high cost of thermal sensors, lower image resolution, higher image noise and lack of widely available data sets. However, these drawbacks are becoming less relevant as infrared imaging advances, making the technology important for applications such as video surveillance, navigation and object tracking. Night vision cameras that produce images in multiple spectral bands, e.g. thermal and

Nedeljko Cvejic
Centre for Communications Research, University of Bristol, Merchant Venturers Building,
Woodland Road, Bristol BS8 1UB, UK, e-mail: nc332@cam.ac.uk

Nishan C. Canagarajah
Centre for Communications Research, University of Bristol, Merchant Venturers Building,
Woodland Road, Bristol BS8 1UB, UK

David R. Bull
Centre for Communications Research, University of Bristol, Merchant Venturers Building,
Woodland Road, Bristol BS8 1UB, UK

visible, have also become available. These different bands provide complementary information since they represent different characteristics of a scene or object.

Fusion of visible and infrared (IR) images and video sources is becoming increasingly important for surveillance purposes. The main reason is that a fused image, constructed by combination of features of visible and infrared inputs, enables improved detection and unambiguous localisation of a target (represented in the thermal image) with respect to its background (represented in the visible image) [1]. A human operator using a suitably fused representation of visible and IR images may therefore be able to construct a more complete and accurate mental representation of the perceived scene, resulting in a larger degree of situation awareness [2].

Image fusion is a specialisation of the more general topic of data fusion, dealing with image and video data [3]. There are a number of potential advantages of integrating the data from multiple sensors. These include [4]:

1. Redundant information provided by a group of sensors can reduce overall uncertainty and increase accuracy of the integrated image
2. Complementary information from different sensors allows features in a scene to be perceived that would not be possible from individual sensors
3. More timely information is available as a group of sensors can collect information of a scene more quickly than a single sensor.

Image fusion is defined in [5] as the process by which several images, or some of their features, are combined together to form a single image. The fusion process must satisfy the following requirements as described in [6]:

1. Preserve all relevant information from the input images in the fused image;
2. Suppress irrelevant parts of the image and noise
3. Minimise any artefacts or inconsistencies in the fused image.

Image fusion can be performed at four main levels [4]. These, sorted in ascending order of abstraction, are: signal; pixel; feature and symbolic level.

At pixel-level, images are combined by considering individual pixel values or small arbitrary regions of pixels in order to make the fusion decision. This method takes no account of what the pixels may represent, but only individual pixel value or the values of an arbitrary number of surrounding pixels to make the fusion decision. This obviously has drawbacks, as useful information of the semantic content of the image is not used in the fusion process.

Feature-level fusion has advantages over pixel-based methods as pixels are regarded as making up a feature in an image. Thus, more intelligent semantic fusion rules can be considered based on actual features in the image. For example, fusion rules can be expanded to include a number of properties of each of the features in the images, such as statistical measures, size, shape or position in a scene.

At symbolic-level, features in an image are classified as a specific type of symbol. These sets of symbols are then fused. This level requires a library of all symbol types to be fused and hence is not as general as lower levels of fusion and requires complex and time-consuming classification steps.

Nikolov et al. [7] proposed another classification of image fusion algorithms – spatial domain and transform domain techniques. The transform domain image fusion consists of performing a transform on each input image and, following specific rules, combining them into a composite transform domain representation. The composite image is obtained by applying the inverse transform on this composite transform domain representation.

Instead of using a standard bases system, such as the DFT, the mother wavelet or cosine bases of the DCT, one can train a set of bases that are suitable for a specific type of images. A training set of image patches, which are acquired randomly from images of similar content, can be used to train a set of statistically independent bases. This is known as Independent Component Analysis (ICA) [8]. Recently, several algorithms have been proposed [9, 10], in which ICA bases are used for transform domain image fusion.

In this chapter, we describe a novel multimodal image fusion algorithm in ICA domain. It uses separate training subsets for visible and IR images to determine the most important regions in the input images and consequently fuses the ICA coefficients using fusion metrics to maximise the quality of the fused image.

2 Image Analysis Using ICA

In order to obtain a set of statistically independent bases for image fusion in the ICA domain, training is performed with a predefined set of images. Training images are selected in such a way that the content and statistical properties are similar for the training images and the images to be fused. An input image $i(x,y)$ is randomly windowed using a rectangular window w of size $N \times N$. The result of windowing is an ‘image patch’ which is defined as [9]:

$$p(m,n) = w \cdot i(m_0 - N/2 + m, n_0 - N/2 + n) \quad (1)$$

where m and n take integer values from the interval $[0, N - 1]$. Each image patch $p(m,n)$ can be represented by a linear combination of a set of M basis patches $b_i(m,n)$:

$$p(m,n) = \sum_{i=1}^M v_i b_i(m,n) \quad (2)$$

where v_1, v_2, \dots, v_M stand for the projections of the original image patch on the basis patch, i.e. $v_i = \langle p(m,n), b_i(m,n) \rangle$. A 2D representation of the image patches can be simplified to a 1D representation, using lexicographic ordering. This implies that an image patch $p(m,n)$ is reshaped into a vector \underline{p} , mapping all the elements from the image patch matrix to the vector in a row-wise fashion. Decomposition of image patches into a linear combination of basis patches can then be expressed as follows:

$$\underline{p}(t) = \sum_{i=1}^M v_i(t) \underline{b}_i = [\underline{b}_1 \underline{b}_2 \dots \underline{b}_M] \cdot [v_1(t) v_2(t) \dots v_M(t)]^T \quad (3)$$

where t represents the image patch index. If we denote $B = [b_1 b_2 \dots b_M]$ and $v(t) = [v_1 v_2 \dots v_M]^T$, then (3) reduces to:

$$\underline{p}(t) = B\underline{v}(t) \quad (4)$$

$$\underline{v}(t) = B^{-1}\underline{p}(t) = A\underline{p}(t) \quad (5)$$

Thus, $B = [b_1 b_2 \dots b_M]$ represents an unknown mixing matrix (analysis kernel) and $A = [a_1 a_2 \dots a_M]$ the unmixing matrix (synthesis kernel). This transform projects the observed signal $\underline{p}(t)$ on a set of basis vectors. The aim is to estimate a finite set of $K < N^2$ basis vectors that will be capable of capturing most of the input image properties and structure. More detailed description and theoretical background of the general ICA process and its application in image processing can be found in [8, 9].

After the input image patches $\underline{p}(t)$ are transformed to their ICA domain representations $\underline{v}_k(t)$, we can perform image fusion in the ICA domain in the same manner as it is performed in e.g. the wavelet domain. The equivalent vectors $\underline{v}_k(t)$ from each image are combined in the ICA domain to obtain a new image $\underline{v}_f(t)$. The method that combines the coefficients in the ICA domain is called the 'fusion rule'. After the composite image $v_f(t)$ is constructed in the ICA domain, we can move back to the spatial domain, using the synthesis kernel A , and synthesise the image $i_f(x, y)$.

3 Overview of Fusion Metrics

As fusion metrics are an important element in the proposed algorithm's optimisation this section includes a short overview of these techniques. Objective image fusion performance evaluation is difficult due to different application requirements and the lack of a clearly defined ground-truth. Several objective performance measures for image fusion have also been proposed where the knowledge of ground-truth is not assumed.

3.1 Piella Metric

The measure used as the basis for the Piella metric is the Universal Image Quality Index (UIQI) [11]. In [11] UIQI was compared to the standard MSE objective quality measure and experimental results have shown that the new index outperforms the MSE, due to the UIQI's ability to measure structural distortions [11].

Let $X = \{x_i | i = 1, 2, \dots, N\}$ and $Y = \{y_i | i = 1, 2, \dots, N\}$ be the original and the test image signals, respectively. UIQI is defined as [11]:

$$Q = \frac{4\sigma_{xy}\bar{x} \cdot \bar{y}}{(\sigma_x^2 + \sigma_y^2) \cdot [(\bar{x})^2 + (\bar{y})^2]} \quad (6)$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad (7)$$

$$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2 \quad (8)$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \quad (9)$$

In order to apply the UIQI for image fusion evaluation, Piella and Heijmans [12] introduce salient information to the metric:

$$Q_p(X, Y, F) = \sum_{w \in W} c(w) [\lambda \cdot Q(X, F|w) + (1 - \lambda) \cdot Q(Y, F|w)] \quad (10)$$

where X and Y are the input images, F is the fused image, $c(w)$ is the overall saliency of a window and λ is defined as [12]:

$$\lambda = \frac{s(X|w)}{s(X|w) + s(Y|w)} \quad (11)$$

should reflect the relative importance of image X compared to image Y within the window w . $s(X|w)$ denotes saliency of image X in window w . It should reflect the local relevance of image X within the window w , and it may depend on e.g. contrast, sharpness, or entropy. This image fusion metric does not require a ground-truth or reference image. Finally, to take into account aspects of the human visual system (HVS), the same measure is computed with ‘edge images’ (X' , Y' and F') instead of the grey-scale images X , Y and F .

$$Q_E(X, Y, F) = Q_P(X, Y, F)^{1-\alpha} Q_P(X', Y', F')^\alpha \quad (12)$$

3.2 Petrovic Metric

The fusion metric proposed Petrovic and Xydeas [13], is obtained by evaluating the relative amount of edge information transferred from the input images to the output image. It also takes into account the relative perceptual importance of the visual information found in the input images, by assigning perceptual importance weights to more salient edges. It uses a Sobel edge operator to calculate the strength $g(n, m)$ and orientation $\alpha(n, m)$ information of each pixel in the input and output images. The relative strength and orientation ‘change’ values, $G_{AF}(n, m)$ and $A_{AF}(n, m)$, respectively, of an input image A with respect to the fused one F are defined as:

$$G^{AF}(n, m) = \begin{cases} \frac{g_F(n, m)}{g_A(n, m)} & \text{if } g_A(n, m) > g_F(n, m) \\ \frac{g_A(n, m)}{g_F(n, m)} & \text{otherwise} \end{cases} \quad (13)$$

$$A^{AF}(n, m) = \frac{|\alpha_A(n, m) - \alpha_F(n, m)| - \pi/2}{\pi/2} \quad (14)$$

These measures are then used to estimate the edge strength and orientation preservation values, $Q_g^{AF}(n, m)$ and $Q_\alpha^{AF}(n, m)$:

$$Q_g^{AF}(n, m) = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}(n, m) - \sigma_g)}} \quad (15)$$

$$Q_\alpha^{AF}(n, m) = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(A^{AF}(n, m) - \sigma_\alpha)}} \quad (16)$$

where the constants Γ_g, k_g, σ_g and $\Gamma_\alpha, k_\alpha, \sigma_\alpha$ determine the exact shape of the sigmoid nonlinearities used to form the edge strength and orientation. The overall edge information preservation values are then defined as:

$$Q^{AF}(n, m) = Q_g^{AF}(n, m) \cdot Q_\alpha^{AF}(n, m) \quad 0 \leq Q^{AF}(n, m) \leq 1 \quad (17)$$

Having $Q^{AF}(n, m)$ and $Q^{BF}(n, m)$ a normalised weighted performance metric of a given process p that fuses A and B into F is given as:

$$Q_p = \frac{\sum_{n=1}^N \sum_{m=1}^M Q^{AF}(n, m) w_A(n, m) + Q^{BF}(n, m) w_B(n, m)}{\sum_{n=1}^N \sum_{m=1}^M w_A(n, m) + w_B(n, m)} \quad (18)$$

The edge preservation values $Q^{AF}(n, m)$ and $Q^{BF}(n, m)$ are weighted by coefficients $w_a(n, m)$ and $w_b(n, m)$, which reflect the perceptual importance of the corresponding edge elements within the input images. Note that in this method, the visual information is associated with the edge information while the region information is ignored.

4 Proposed Fusion Method Using Independent Component Analysis

4.1 Separated Training Sets

In the proposed method, training images are separated in two groups prior to training process. Namely, all IR training images are grouped into a separate training subset, whereas all the visible training images constitute the second training subset.

Introduction of separate training subsets provides us with two sets of ICA bases. The first ICA bases set is used to decompose the IR input image patches $v_i(t) = A_i p_i(t)$ and the second subset to transform the visible input image patches to ICA domain $v_v(t) = A_v p_v(t)$.

Separate ICA bases sets for decomposition of input images are more specifically trained to capture statistical properties of the specific modality of the input images (IR/visual). This enables the proposed method to outperform the standard method [9], in which images of both IR and visible modality are used for training which results in an ‘average’ ICA bases set that is not able to take the full advantage of ICA decomposition. It is important to note that before the reconstruction of the fused image in pixel domain it is necessary to normalise the energy of two the ICA bases subsets. The normalisation provides the necessary amplitude balance between the ICA coefficients obtained using two different ICA bases sets.

4.2 Region-Based Fusion of ICA Coefficients

The majority of applications of a fusion scheme employ the features within the image, not in the actual pixels. Therefore, it seems reasonable to incorporate feature information into the fusion process. There are a number of perceived advantages of this, including:

1. Intelligent fusion rules: Fusion rules are based on combining groups of pixels, which form the regions of an image. Thus, more useful tests for choosing the regions of a fused image, based on various properties of a region, can be implemented.
2. Highlighting features: Regions with certain properties can be either accentuated or attenuated in the fused image depending on a number of the characteristics of the region.
3. Improved noise robustness to noise: Processing semantic regions rather than individual pixels or arbitrary regions can help overcome some of the problems with pixel-fusion methods such as sensitivity to noise, blurring effects and misregistration.

Several features can be employed in the estimation of the contribution of each input image to the fused output image. For example, the mean absolute value of each ICA coefficient can be used as an activity indicator in each input image:

$$E_i(t) = \|v_i(t)\|, \quad E_v(t) = \|v_v(t)\| \quad (19)$$

where T denotes the number of input images. As the ICA bases tend to focus on the edge information, large values for $E_k(t) (k \in \{i, v\})$ correspond to increased activity in the patch, e.g. the existence of edges or a specific texture. Based on this observation, the standard ICA image fusion method divides the ICA domain coefficients

in two groups [9]. The first group consists of the regions that contain details ($E_k(t)$ larger than a threshold) and the second group contains the region with background information ($E_k(t)$ smaller than a threshold). The threshold that determines whether a region is ‘active’ or ‘non-active’ is set heuristically. As a result, the segmentation map $s_i(t)$ is created for the IR input image:

$$s_i(t) = \begin{cases} 1 & \text{if } E_i(t) > \frac{2}{T} \sum_{k=1}^T E_i(t) \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

as well as for the visible input image $s_v(t)$. The segmentation maps of input images are combined to form a single segmentation map, using the logical OR operator:

$$s(t) = OR\{s_i(t), s_v(t)\} \quad (21)$$

After the input images are segmented into active and non-active regions, different fusion rules can be used for fusion of each group of regions. In [9] active regions are fused using the ‘max-abs’ rule, while non-active regions are fused using the ‘mean’ rule. The ‘max-abs’ rule fuses two input coefficients/vectors by selecting the one with higher absolute value. In the ‘mean’ fusion rule the fused coefficient/vector is equal to the mean value of the two input coefficients/vectors.

Because the aforementioned threshold that determines the ‘activity’ of a region is set heuristically, the regions obtained by thresholding of the ICA coefficients do not correspond always to objects in the images to be fused. Our experiments showed that important objects in the IR input images (e.g. a person or a smaller object) are often masked by textured high-energy background in the visual image. In this case the important objects from the IR image become blurred or, in extreme cases, completely masked. Thus, we propose a different rule for fusion of surveillance images. If $i_i(x, y)$ is an IR input image and $i_v(x, y)$ is a visible image, we obtain the mean absolute values of each ICA coefficient, ($E_i(t)$ and $E_v(t)$), respectively, using different ICA bases set for each image. The active and non-active regions are then determined separately, as given in (20). The active regions (energy of the region higher than the threshold $\frac{2}{T} \sum_{k=1}^T E_i(t)$) from the IR image are compared to the active regions from the visible image at the same location and the active regions from the IR image, determined by specific ICA subset for the IR images, are given higher priority than the visible image. Using this principle, we aim to transfer all the important regions (important surveillance-wise, e.g. objects representing a person walking or a source of heat) from the IR image to the fused image, not allowing them to be masked by high-energy regions from the visible image. On the other hand, in the fusion of the non-active regions the ‘max-abs’ fusion rule is used. As a consequence, most of the background details will be retained from the visible image, because of its fine high-energy texture, thus increasing the perceptual quality of the fused image.

4.3 Reconstruction of the Fused Image Using Fusion Metrics

In addition, we implement a novel method for reconstruction of the fused image, using statistical properties of the both input images. In the standard ICA method, reconstruction of the fused image is performed on the patch-per-patch base [9]:

$$p_f(t) = U_i(t) + U_v(t) + 1/2(M_i(t) + M_v(t)) \quad (22)$$

where $p_f(t)$ represents the t -th patch of the fused image $i_f(x, y)$, whereas $U_i(t)$ and $U_v(t)$ are the t -th patch obtained by inverse transform of the selected ICA coefficient from the IR image and visible image, respectively. $M_i(t)$ is the mean value of the corresponding frame from the IR input image $i_i(x, y)$ and $M_v(t)$ is the mean value of the corresponding frame from the visual input image $i_v(x, y)$. We propose a new approach for reconstruction of the fused image [10]:

$$p_f(t) = U_i(t) + U_v(t) + M_i(t) \cdot w_i + M_v(t) \cdot w_v \quad (23)$$

Weights $w_i \in [0, 1]$ and $w_v (= 1 - w_i) \in [0, 1]$ are used to balance the contributions from both visual and IR images in the synthesis of the fused image.

Weighting coefficients are set to a predefined value (e.g. $w_i = 1$ and $w_v = 0$) and then $w_v = 0$ is gradually increased. One of the fusion performance metric [12, 13] is calculated at each step. We decided to exploit the Piella metric [12] and Petrovic metric [13] because these are the most widespread tools for evaluation of image fusion algorithms. In addition, extensive experiments with multimodal images have shown that both metrics have one, global optimum, when M_i and M_v are modified. It allows us to gradually increase $w_v = 0$ by a predefined step (usually 0.1) and calculate the gradient of the given metric at each step. When the calculated gradient becomes negative, signalling that the optimum value of a fusion performance metric is reached, the process stops and reconstruction of the fused image is performed with the calculated weights. In that sense, the weighting coefficients are chosen so that the quality of the fused image is maximised.

5 Experimental Results

The proposed image fusion method was tested in different surveillance scenarios with two modalities: infrared and visible. In order to make a comparison between the proposed method and the standard ICA method, the images were fused using the approach described in [9]. We compared these results with a simple averaging method, the ratio method [14], the Laplace transform (LT) [15] and the dual-tree complex wavelet transform (DT-CWT) [16].

Before performing image fusion, the ICA bases were trained using a set of 5 images IR images and 5 visible images, with content comparable to the test set. A number of rectangular patches ($N = 4, 8, 12$) used for training were randomly selected from the training set. The lexicographic ordering was applied to the image

patches and then PCA performed. Following this, a number of the most important bases were selected, according to the eigenvalues corresponding to these bases. After that, the ICA update rule in was iterated until convergence. ICA coefficients were obtained using the principle described in Sect. 3, while reconstruction of the fused image was performed using optimisation based on the Piella fusion performance metric [12].

5.1 Comparison with the Standard ICA Image Fusion Method

In the first part, experiments were focused on the performance evaluation of the proposed algorithm and comparison with the standard ICA fusion algorithm. In order to compare how the fusion algorithm's performance depend on the training process, impact of the number of training patches taken from the training set was tested. Number of images in the training set was fixed to 10 for the standard ICA fusion method and to five images per subset of training images in the IR and visible domain, for the proposed method. The number of training patches taken from the training set (subsets) was then varied from 100 to 40000 in order to evaluate both algorithms' performance with different number of training patches. Size of training patches was 8×8 ($N = 8$) and 32 of the most significant bases obtained by training are selected using the PCA algorithm.

The results in Figs. 1–3 show that the proposed algorithm significantly outperforms the standard ICA fusion algorithm for the UN Camp and Octec surveillance image sequences, with constantly higher scores in terms of Piella and Petrovic metric. Figures 2–4. depict examples of fused images for the standard and proposed ICA fusion algorithm, where different number of training patches is used. Visual (subjective) comparison between methods indicates that our method is far superior to the basic ICA method: for example, it is clear that the fence detail from the visual image is far better transferred into the fused image in the proposed method.

In addition, the details of the tree in the visual image are visually more pleasing and the human figure is much brighter in the proposed method than in the fused image obtained by the standard ICA method. It is also noticeable that the performance of the proposed method is less dependent on the number of training patches than the standard ICA fusion method. In addition, the proposed method trained by only 200 training patches outperforms the standard ICA method trained by 40000 training patches, measured by both fusion metrics. Therefore, the proposed algorithm needs a significantly shorter training process in order to obtain fusion performance comparable to, or above, the performance of the standard ICA method.

5.2 Comparison with the State-of-the-Art Image Fusion Methods

The proposed image fusion method was tested against several state-of-the-art image fusion methods in two modalities: infrared and visible. In order to make a

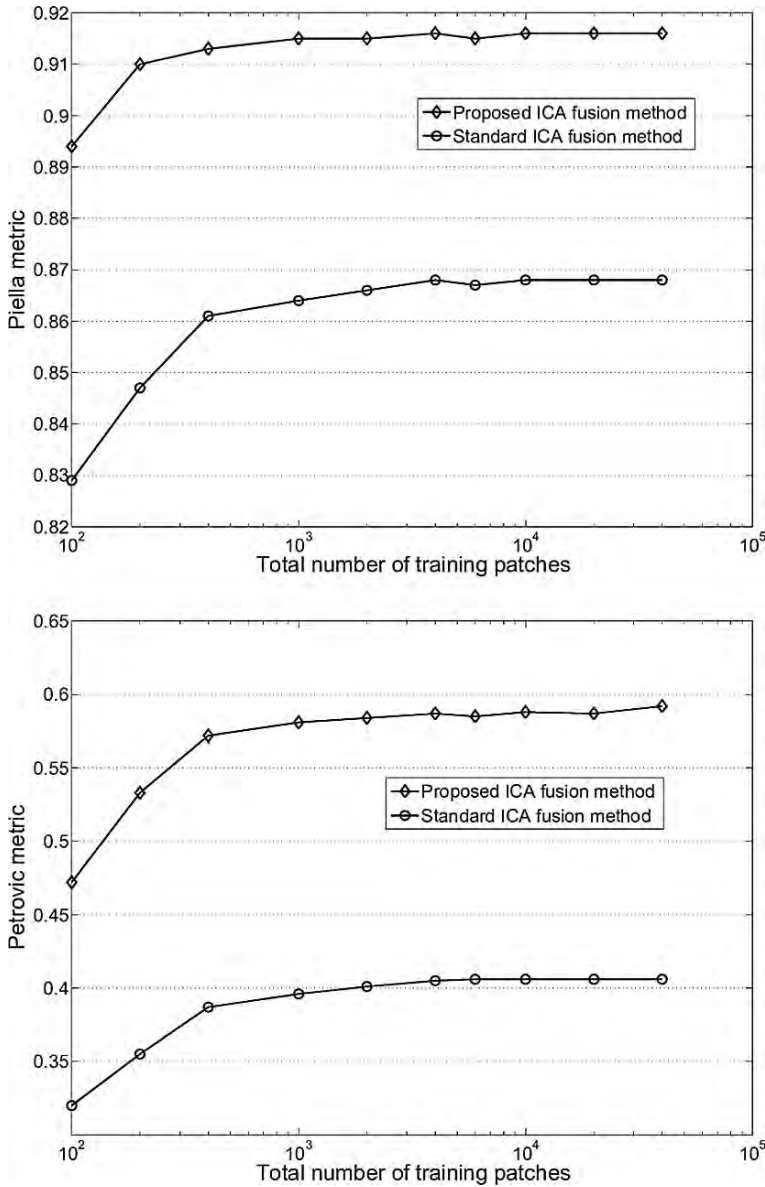


Fig. 1 Fusion performance for image 1812 from the 'UN Camp' sequence. Comparison of fusion performance vs. number of training patches for the proposed and standard ICA fusion method, Piella metric (top), Petrovic metric (bottom)

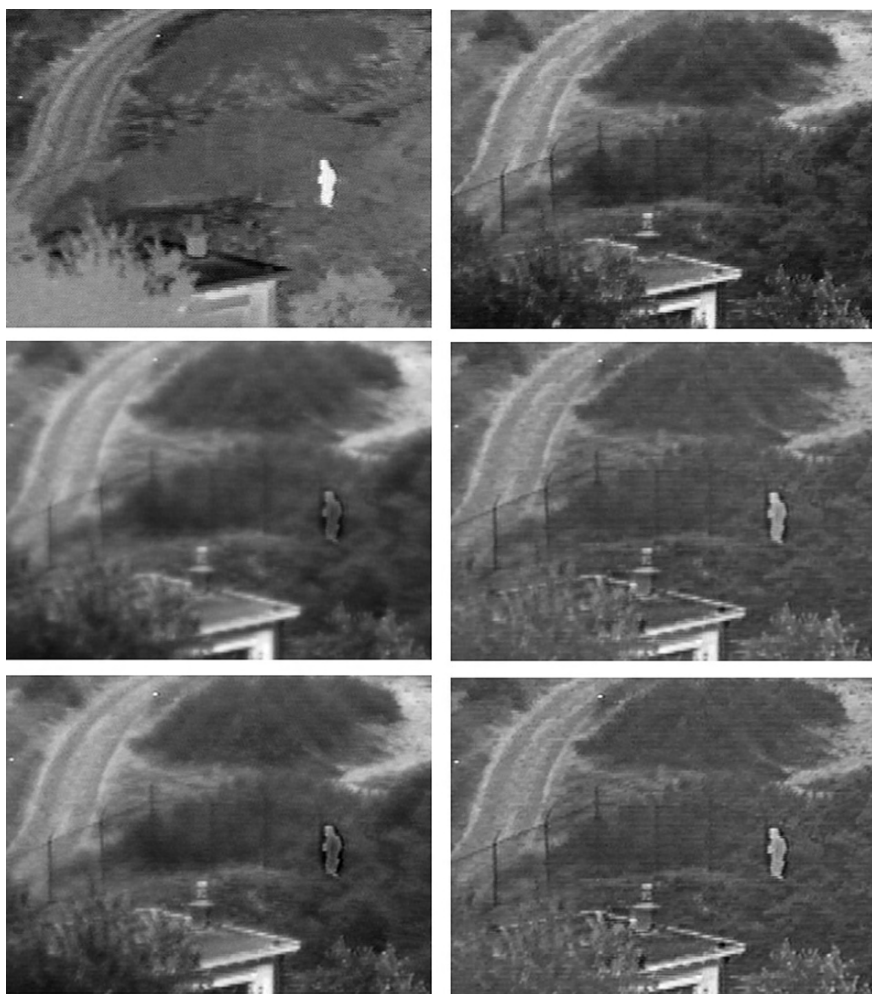


Fig. 2 Subjective fusion results. *Top*: input IR image (*left*), input visible image (*right*). *Middle*: fused image using standard ICA fusion and 100 training patches (*left*); fused image using proposed method and 100 training patches (*right*). *Bottom*: standard ICA fusion and 10000 training patches (*left*), proposed method and 10000 training patches (*right*)

comparison between the proposed method and the standard ICA method, the images were fused using the approach described in [9]. We compared these results with a simple averaging method, the ratio method [14], the Laplace transform (LT) [15] and the dual-tree complex wavelet transform DT-CWT) [16]. In the multiresolution methods (LT, DT-CWT) a 5-level decomposition is used and fusion is performed by selecting the coefficient with a maximum absolute value, except for the case of the lowest resolution subband where the mean value is used. The images fused using these algorithms are given in Figs. 5 and 6, together with IR

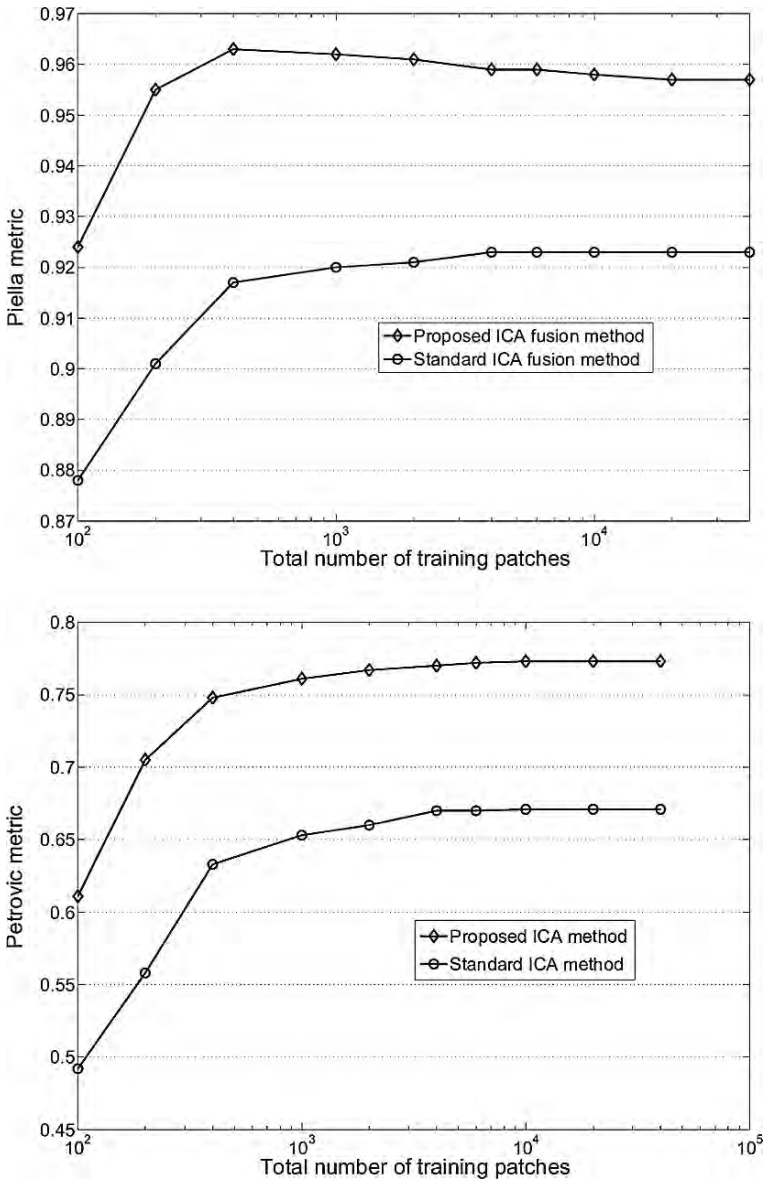


Fig. 3 Fusion performance for image 22 from the 'Octec' sequence. Comparison of fusion performance vs. number of training patches for the proposed and standard ICA fusion method, Piella metric (top), Petrovic metric (bottom)

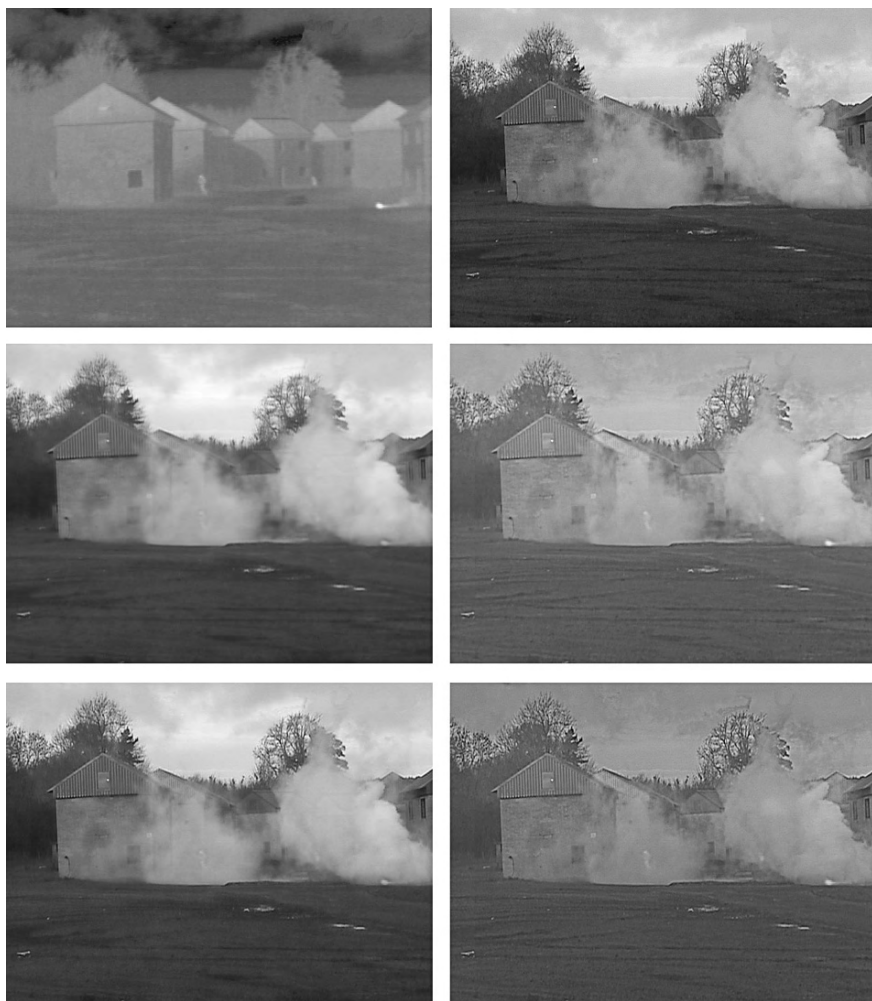


Fig. 4 Subjective fusion results. *Top*: input IR image (*left*), input visible image (*right*). *Middle*: fused image using standard ICA fusion and 100 training patches (*left*); fused image using proposed method and 100 training patches (*right*). *Bottom*: standard ICA fusion and 10000 training patches (*left*), proposed method and 10000 training patches (*right*)

and visible input images. The proposed and standard ICA method were trained using 10000 training patches taken from a set of images with similar content. Size of training patches was 8×8 ($N = 8$) and 32 of the most significant bases obtained by training are selected using the PCA algorithm. It should be noted that the adaptive fused image reconstruction adds 1–2% of computational overhead to the standard, non-adaptive ICA fusion algorithm. Visual (subjective) comparison between methods indicates that our method is far superior to the basic ICA method, but also that the proposed weighted ICA method performs slightly better than the

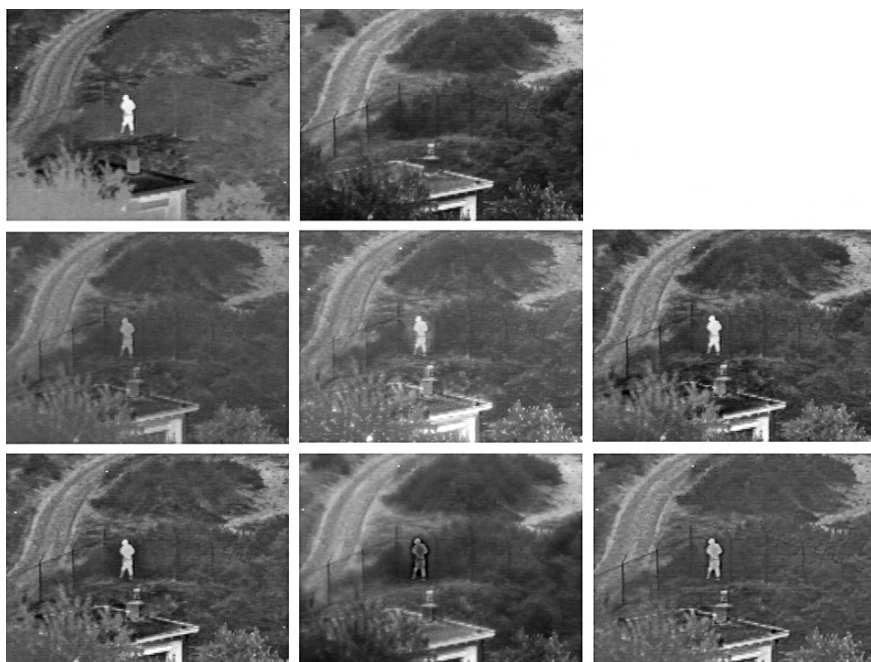


Fig. 5 Subjective fusion results. *Top*: input IR image (*left*), input visible image (*middle*). *Middle*: fused image using averaging (*left*), ratio pyramid (*middle*) and Laplace pyramid (*right*). *Bottom*: fused image using DT-CWT (*left*), standard ICA method (*middle*) and proposed ICA method (*right*)

LT and DT-CWT methods: for example, in Fig. 5 it is clear that the fence detail from the visual image is far better transferred into the fused image in the proposed method than in the standard ICA method. In addition, the details of the tree in the visual image are visually more pleasing in the proposed method than in the DT-CWT approach, although the person is brighter in the DT-CWT fused image. In Fig. 6 it is obvious that the proposed method outperforms standard ICA as the landscape structure is better represented in the fused image and the terrain information is clearer in the proposed ICA method compared to the DT-CWT and LT methods.

The results in Table 1 show that the proposed algorithm significantly outperforms the standard ICA fusion algorithm for the all tested surveillance image sequences, with constantly higher scores in terms of Piella and Petrovic metric. The proposed method generally obtains higher performance than the multiresolution methods as well, as metric values are usually higher, except for the case of the Dune sequence, where the best results are obtained by the DT-CWT fusion method. The advantage in terms of metric values is more brought up for the Petrovic metric, because it has higher dynamics (smaller differences in the fused image are discriminated with larger difference in the metric grade). The metrics' values confirm the subjective

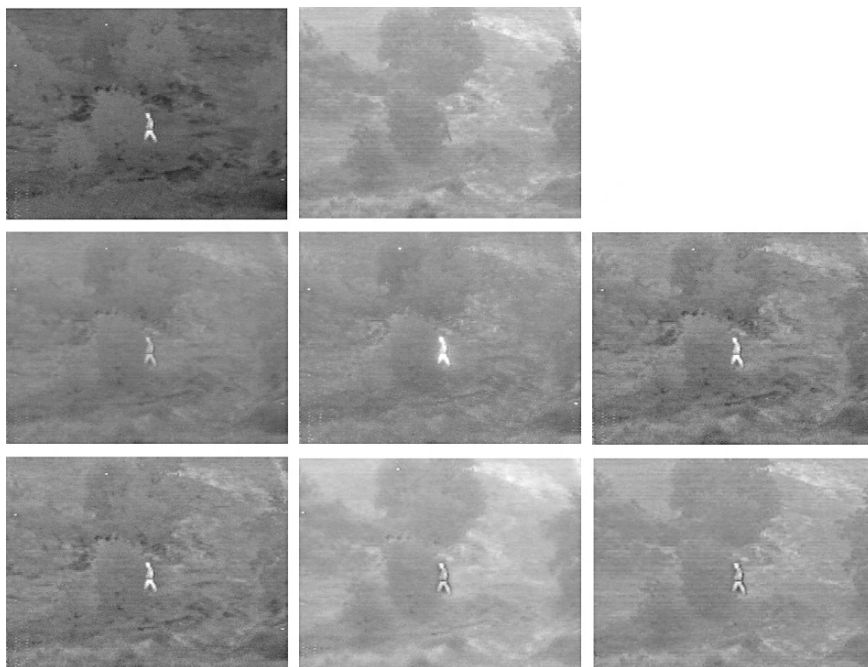


Fig. 6 Subjective fusion results. *Top:* input IR image (*left*), input visible image (*middle*). *Middle:* fused image using averaging (*left*), ratio pyramid (*middle*) and Laplace pyramid (*right*). *Bottom:* fused image using DT-CWT (*left*), standard ICA method (*middle*) and proposed ICA method (*right*)

Table 1 Performance of image fusion methods, measured by fusion metrics

Metric	Method	UN1812	Dune04	Octec22	Trees17
Piella	Average	0.86	0.96	0.87	0.91
	Ratio	0.86	0.96	0.88	0.92
	DT-CWT	0.91	0.97	0.94	0.92
	Laplace	0.91	0.96	0.94	0.92
	Standard ICA	0.87	0.95	0.92	0.91
	Proposed ICA	0.92	0.96	0.95	0.93
Petrovic	Average	0.35	0.51	0.43	0.44
	Ratio	0.41	0.53	0.50	0.47
	DT-CWT	0.46	0.60	0.77	0.55
	Laplace	0.50	0.60	0.77	0.55
	Standard ICA	0.41	0.52	0.67	0.45
	Proposed ICA	0.59	0.65	0.78	0.60

impression that the images obtained using the proposed algorithm generally incorporate more information from the visible image together with the important details from the IR image.

6 Conclusion

In this chapter, we describe an improved image fusion algorithm based on the Independent Component Analysis (ICA). In the proposed method, images used for training of ICA bases are separated in two groups prior to training process, one consisting of IR images and the second consisting of visible images. Region-based fusion of ICA coefficients is implemented, where the mean absolute value of each ICA coefficient is used as an activity indicator for the given region. Weighting of the ICA bases during reconstruction of the fused image by using the fusion metrics is used to maximize the performance of the proposed method. Experimental results confirm that the proposed method exhibits significantly better fusion than basic ICA method, as it obtains higher scores using both Piella and Petrovic metrics. The proposed method outperforms the performance of the state-of-the-art algorithms, both in terms of subjective quality and fusion metrics values.

References

1. Toet A, Ijspeert JK, Waxman AM, Aguilar M (2003) Perceptual evaluation of different image fusion schemes. *Displays*, 24:25–37
2. Toet A, Franken EM (1997) Fusion of visible and thermal imagery improves situational awareness. *Displays*, 18:85–95
3. Maitre H, Bloch I (1997) Image fusion. *Vistas in Astronomy*, 41(2):329–335
4. Abidi M, Gonzalez R (1992) *Data Fusion in Robotics and Machine Intelligence*. Academic Press, USA
5. Nikolov S (1998) Image fusion: A survey of methods, applications, systems and interfaces. Technical Report UoB-SYNERGY-TR02, University of Bristol, United Kingdom
6. Rockinger O (1996) Pixel-level fusion of image sequences using wavelet frames. In: Proc. 1996 Leeds Applied Shape Research workshop. Leeds, United Kingdom
7. Nikolov S, Bull DR, Canagarajah CN (2001) Wavelets for image fusion. In: *Wavelets in Signal and Image Analysis*. Kluwer, Dordrecht, The Netherlands
8. Hyvärinen A, Karhunen J, Oja E (2001) *Independent Component Analysis*. John Wiley and Sons, London, United Kingdom
9. Mitianoudis N, Stathaki T (2007) Pixel-based and Region-based Image Fusion schemes using ICA bases. *Information Fusion*, 8(1):131–142
10. Cvejic N, Bull DR, Canagarajah CN (2007) Region-based multimodal image fusion using ICA bases. *IEEE Sensors Journal* 7(5–6):743–751
11. Wang Z, Bovik AC (2002) A universal image quality index. *IEEE Signal Processing Letters*. 9(2):81–84
12. Piella G, Heijmans H (2003) A new quality metric for image fusion. In Proc. 2003 IEEE International Conference on Image Processing. Barcelona, Spain, 173–176
13. Xydeas C, Petrovic V (2000) Objective pixel-level image fusion performance measure. In: Proc. 2000 SPIE. Orlando, FL, 88–89
14. Toet A (1996) Image fusion by a ratio of low-pass pyramid. *Pattern Recognition Letters*, 9:245–253
15. Burt P, Adelson E (1983) Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):115–123
16. Lewis JJ, O’Callaghan RJ, Nikolov SG, Bull DR, Canagarajah CN (2007) Pixel- and region-based image fusion with complex wavelets. *Information Fusion*, 8(1):119–130

Part VII

Vision Sensing

Fast Image Capture and Vision Processing For Robotic Applications

Gourab Sen Gupta and Donald Bailey

Abstract This chapter details a technique to significantly increase the speed of image processing for robot identification in a global-vision based system, targeted at real-time applications. Of major significance are the proposed discrete and small look-up tables for Y, U and V color thresholds. A new YUV color space has been proposed which significantly improves the speed of color classification. The look-up tables can be easily updated in real-time and are thus suitable for adaptive thresholding. The experimental results confirm that the proposed algorithm greatly improves the performance of the image processing system. The results are compared with other commonly used methods such as a composite look-up table which is indexed using RGB pixel values.

Keywords Global vision · colour segmentation · YUV colour space · incremental tracking

1 Introduction

Vision systems are widely used in the industry for object tracking, intrusion detection, vehicle and mobile robot guidance, inspection automation, etc. [1]. The majority of the commodity vision systems use video signals, most often from a CCD camera, as input to the image capture and analysis subsystems. Typically, such vision systems provide frame rates of 30 Hz or field rates of 60 Hz for interlaced images. Processing these images with useful resolution of 320×240 and above in the 33.3 and 16.67 ms sample times respectively can pose a significant challenge especially when other processing tasks such as strategy and task allocation, low-level

Gourab Sen Gupta
School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand, e-mail: g.sengupta@massey.ac.nz

Donald Bailey
School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand, e-mail: d.g.bailey@massey.ac.nz

control and communication with the robots, are also to be completed. A lot of research efforts have been spent on improving the speed of image processing for robotic applications [2] and it continues to attract a lot of attention of researchers. Faster vision processing algorithms result in better motion control and hence better coordination between agents to accomplish a collaborative task.

A computationally inexpensive vision processing algorithm using Run Length Encoding (RLE) has been discussed in papers [3] and [4]. RLE is an image compression technique that preserves the topological features of an image, allowing it to be used for object identification and location [5]. Though the RLE algorithm can be implemented on commodity hardware for multi-agent collaborative systems such as the robot soccer vision system [6, 7], its significant processing speed advantage is when there are a large number of objects to track. For smaller systems with a limited number of agents, say two to five, the commonly used blob identification techniques together with the incremental tracking algorithm is adequate and equally efficient.

Interlaced images introduce the ‘image scattering’ problem for a moving object because of the time delay between the two fields of the image. To overcome this problem, the odd and even scan fields have to be processed separately. However, because of quantization errors in each field, a stationary object may appear to be in two different locations. Filtering techniques are required to minimize such a negative effect. The other sources of errors in the vision system are due to variation of light intensity and inherent sensor noise.

The cumulative effect of the errors in the vision system is very significant, especially in a highly dynamic collaborative system where the agents are moving very fast. Figure 1 shows the software hierarchy of a multi-agent robotic controller using global vision. The image processing software identifies the position and orientation of the robots and other objects of interest in the robots’ workspace. The vision data is filtered to reduce the effect of noise and passed on to the strategy/task allocation

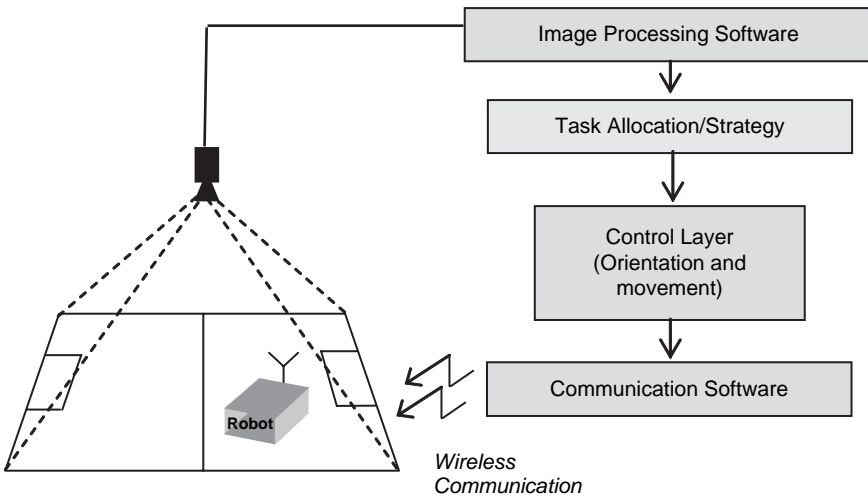


Fig. 1 Software hierarchy of a multi-agent robotic controller using global vision

layer. At this layer of the software architecture, the behavior required of an agent is determined. The errors in the vision system percolate down the hierarchy of the robotic controller and have profound effect on the robot behavior and hence the overall performance of the collaborative system. It is thus imperative that careful considerations are given to eliminate or at least minimize the vision processing errors.

2 Global Vision – Sources of Error

A global vision system uses a single or multiple cameras to detect and track several objects. The main sources of errors in the vision system are described in the following sub-sections.

2.1 *Separate Processing of Odd and Even Scan Fields of an Interlaced Bit Mapped Image*

A stationary object may be reported at different locations in each frame due to different quantization errors. This is explained using a bit mapped image of 16×16 pixel resolution as shown in Fig. 2. The object of interest, the centre position coordinates of which are required to be calculated, is of a square shape.

To calculate the position, the well known zero-order moment and centre-of-gravity (1) are used. For an $m \times n$ binary image, the coordinates are given by:

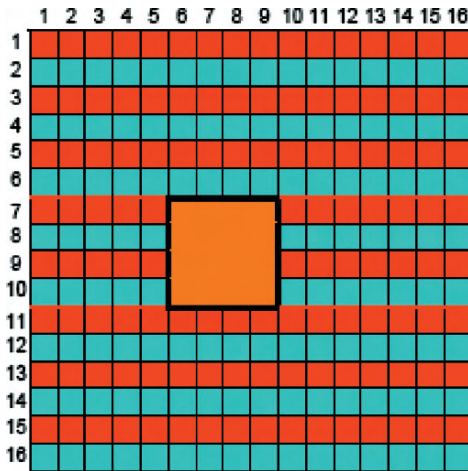


Fig. 2 Bit-map of an interlaced image

$$\begin{aligned}
 \text{Area,} \quad & A = \sum_{i=1}^n \sum_{j=1}^m B[i, j] \\
 \text{COG,} \quad & \bar{x} = \frac{\sum_{i=1}^n \sum_{j=1}^m jB[i, j]}{A} \quad \bar{y} = \frac{\sum_{i=1}^n \sum_{j=1}^m iB[i, j]}{A} \quad (1)
 \end{aligned}$$

$B[i, j]$ is the value of the bit (0 or 1) in the image at location $[i, j]$.

i is the row number (i.e. Y-coordinate)

j is the column number (i.e. X-coordinate)

For the illustrated example, the coordinates of the object in the Odd scan is (7.5, 8.0) as shown in the calculations below:

$$\begin{aligned}
 A &= 8 \\
 \bar{x} &= \frac{6 \times 2 + 7 \times 2 + 8 \times 2 + 9 \times 2}{8} = 7.5 \\
 \bar{y} &= \frac{7 \times 4 + 9 \times 4}{8} = 8
 \end{aligned}$$

The coordinates of the object in the Even scan is (7.5, 9.0) as shown in the calculations below:

$$\begin{aligned}
 \bar{x} &= \frac{6 \times 2 + 7 \times 2 + 8 \times 2 + 9 \times 2}{8} = 7.5 \\
 \bar{y} &= \frac{8 \times 4 + 10 \times 4}{8} = 9
 \end{aligned}$$

The separate processing of odd and even scan fields do not affect the X Coordinate. Only the Y Coordinate has a shift of 1 pixel. Looking at a physical area of $170\text{cm} \times 150\text{cm}$ and working with an image resolution of 320×240 pixels, 1 pixel translates into a shift of $\sim 0.62\text{cm}$ in the Y Coordinate of the object. Moreover, this offset will not be constant for a moving object since the shift can occur both in a positive or negative direction. Filtering techniques are often employed to minimize the quantization error.

2.2 Variation of Light Intensity

While errors due to separate processing of odd and even scans could be predominant under controlled light conditions, nonetheless, these errors are further compounded by variation in light intensity from one frame to another. In the real-world applications of collaborative robotics, it is often not possible to create ideal (or at least stable) light conditions. To partially overcome this problem, YUV color thresholding can be employed with the Y (intensity) range extended to the maximum limits. However, extending the color boundaries too much result in the threshold

values of two or more colors overlapping each other. Also extending the threshold values wider will make the vision system error prone as stray pixels from the background will be picked up too. This limits the color tolerance.

2.3 Inherent Sensor Noise

Certain errors are inherent in the system. These originate from the camera, the frame grabber card, connecting cables, etc.

Errors in vision-generated data have a significant impact on targeting accuracy even when intercepting or striking a stationary object. The tests carried out on moving targets, however, are more significant as interception accuracy suffers when the target is moving. This is due to the fact that actions are initiated based on predicted future positions which are different from the current position and are calculated based on velocity measures which are very noisy.

3 Experimental Hardware Setup

The experimental hardware setup consists of a Pulnix 7EX NTSC camera (www.pulnix.com) with analog composite video output and a FlashBus MV Pro image capture (frame grabber) card with PCI interface (www.integraltech.com). The image is captured at a resolution of 320×480 at a sampling rate of 30 Hz. The odd and even fields are processed separately; hence the effective image resolution is 320×240 delivered at a sampling rate of 60 Hz. The captured image is processed on a 1.8 GHz Pentium 4 PC with 512 MB RAM. The image capture card was configured for off-screen capture, as shown in Fig. 3. Off-screen capture mode facilitates fast processing of the image from the system RAM. The image is transferred to the

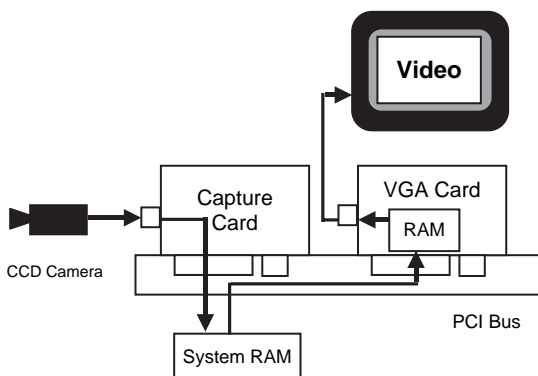


Fig. 3 Image capture card in off-screen capture mode

VGA RAM only when a live image is required to be seen on the screen, such as during the setting and testing of color thresholds. Once the color tuning is done, the transfer of image to the VGA RAM is switched off.

An important hardware feature of the FlashBus MV Pro frame grabber card is that it generates a Vertical Sync for every odd and even field. The interrupt can be detected in the software. This facilitates implementation of an interrupt based system and fixes the sample time of the controller.

4 Colour Segmentation, Area Thresholding, Blob Merging

The color image sequence is processed at three levels: pixel level, blob level, and object level. To facilitate identification and separation of individual objects, a color jacket comprising two color patches can be used on each robot, as shown in Fig. 4. In applications where several groups or teams of robots are involved, one of the color patches is used to identify the group (team color patch) and the other is used to identify which robot it is within the group (robot color patch). The centers of the two color patches, C_r and C_t , are first calculated from the image. The inclination of the line joining the two centers gives the orientation of the robot while the coordinates of the centre of the line give its position. For some target objects such as a ball in a robot soccer system, only the centre of the color patch is calculated as it does not have an orientation. The velocity of the target is used to add a direction vector to its position.

The accuracy of the angle calculation depends on the accuracy with which the centers of the color patches are detected and how far apart these centers are. If the centers are closer to each other, any small variation in the calculation of C_r and C_t will result in a large variation in angle. In order to improve the accuracy of angle calculation, experiments with different color jackets were performed. The centers of the color patches in Fig. 5 are further apart than those in Fig. 4, thus improving the accuracy of the angle calculation.

The object detection algorithm starts with color segmentation. It searches the image to determine if the pixels belong to one of the calibrated color classes. The pix-

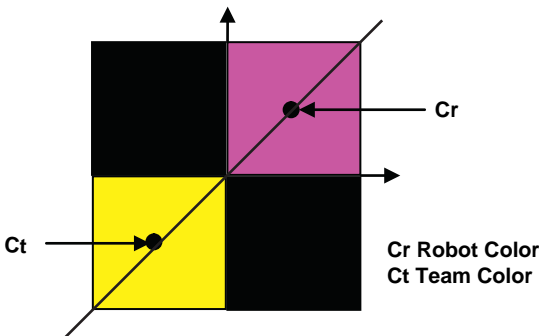


Fig. 4 Color jacket for identification of robot

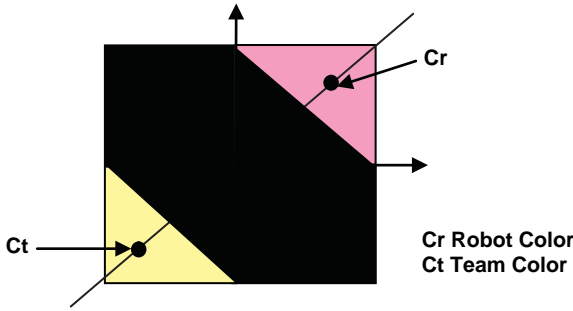


Fig. 5 Color jacket for improved accuracy of orientation

els are then grouped to create color patches using a ‘sequential component labeling algorithm’. This algorithm uses a two-pass labeling technique [8] with identifiers (labels) that increment from the value of 1. Ideally, the number of labels is equal to the number of desired color patches on the objects in the entire image.

The procedure for checking the membership and grouping of pixels consists of 5 steps split into two passes.

A. FIRST PASS (Steps 1 to 4)

1. Process the image in the tracking window from left to right, top to bottom, analyzing each pixel.
2. If the pixel in the image is within the YUV threshold values of the color of interest, then
 - (a) If only one of its upper and left neighbors has a label, copy the label.
 - (b) If both upper and left neighbors have the same label, copy that label.
 - (c) If both upper and left neighbors have different labels, copy the upper pixel’s label and enter the labels in an equivalence table as equivalent labels.
 - (d) If not (a), (b) or (c) assign a new label to this pixel and enter it in the equivalence table.
3. If there are more pixels to consider, repeat step 2 for additional pixels, otherwise proceed to step 4.
4. Find the lowest label for each equivalent set in the equivalence table and add to the equivalence table.

B. SECOND PASS (Step 5)

5. Process the picture by replacing each label with the lowest label in its equivalent set.

To illustrate the algorithm, a binary image (rather than a YUV image) is used in the following example. Figure 6 shows a representation of the binary image before the first pass of the algorithm. The 0’s represent the background of the image and the 1’s represent the objects of interest. It can be seen that there are two objects of interest in the image, one in the left and the other in the right.

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0	0	1	0
0	1	1	1	1	0	0	0	0	1	1	0	1	0
0	1	1	1	1	1	1	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 6 The binary image

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0	0	2	0
0	3	1	1	1	0	0	0	0	4	4	0	2	0
0	3	1	1	1	1	1	0	0	4	4	4	2	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 7 The image after the first pass

Figure 7 shows the image after the first pass of the algorithm. The objects now have multiple labels as the first pass of the algorithm was not able to correctly label all shapes (i.e. the object on the left has labels 3 and 1 and the object on the right has labels 2 and 4). Figure 8 shows the image after the second pass of the algorithm, which resolves the problem of multiple labels for single objects.

The algorithm described above is often called the ‘2-neighbour’ algorithm in which the upper and left neighboring pixels of the pixel under test, are considered to evaluate the membership (label) of a pixel. Another well know and very similar algorithm is called the ‘4-neighbour’ algorithm in which the four neighboring pixels – upper, left, bottom and right – are considered for evaluating the label of a pixel.

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0	0	2	0
0	1	1	1	1	0	0	0	0	2	2	0	2	0
0	1	1	1	1	1	1	0	0	2	2	2	2	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 8 Image after the second pass

This algorithm is more time consuming and gives only marginal improvement in the accuracy of color segmentation. Thus the ‘2-neighbour’ algorithm is often preferred over the ‘4-neighbour’ algorithm, especially in real-time applications where the color patches are relatively large in pixel area.

Once the colors have been segmented, two separate processing steps follow – filtering based on area threshold and blob merging. In order to reduce noise, very small color blobs are discarded if they fall below a certain area threshold. For example, it is usually safe to discard patches which are only 2 or 3 pixels in area as these would generally be noise from the background. If patches of the same color are very close to each other, then these are merged to form one patch. This blob merging technique, with a distance tolerance, is widely used in practice and reported in literature [9, 10].

Having identified the separate objects, the centre of each object is calculated using the centre of gravity calculation described earlier (1).

5 Interrupt Based Multi-Buffered Image Capture

In order to fix the sample time and delay of the vision control loop to a constant value, an interrupt driven approach is adopted. In this method the vision processing and strategy functions are placed in an interrupt service routine. The routine is serviced on each occurrence of the Vertical Sync signal on the frame grabber card which generates a Vertical Sync for every odd and even field. For an image resolution of up to 640×480 , the card can capture the image at 30 frames per second. Hence the interrupt service routine of the interlaced image is executed every 16.67 ms.

This poses a challenge – all the vision processing, strategy calculations, calculation of motion control data and transmission of RF packets to all the robots, must be completed within 16.67 ms. This is achieved by segregating the process of capturing the image and processing it. This is implemented using a four-stage ring buffer to capture and process the image. If the image processing is guaranteed to complete in the specified sample time only two buffers are required. Since this is not the case for the color segmentation and blob-identification algorithm presented in Sect. 4, a four buffer system is required. The buffer organization is shown in Fig. 9. When an

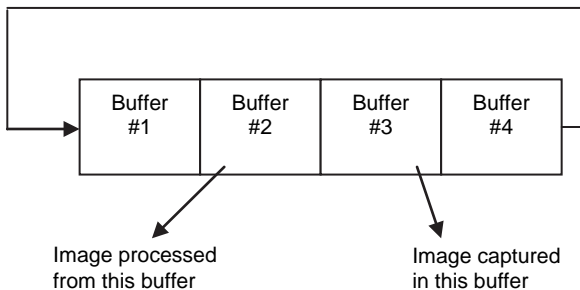


Fig. 9 Multi-buffered image captures

image is captured in a buffer, the image from the previous buffer is processed. This helps to avoid buffer contention during image capture and processing [11].

6 Full Tracking vs. Incremental Tracking

The full tracking algorithm searches through the whole image, testing each pixel whether it is a member of one of the calibrated object color classes. The full tracking process is very inefficient when the objects are small and only represent a small

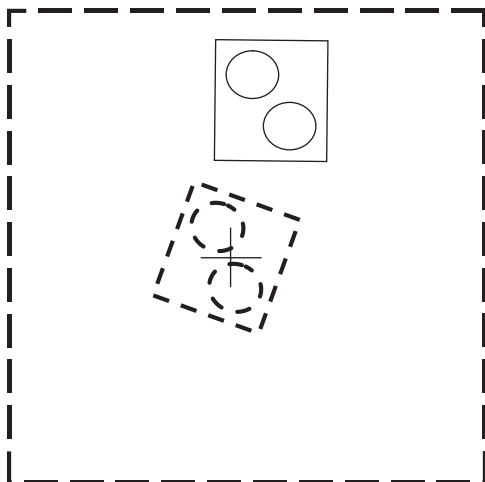


Fig. 10 Tracking window centered on last known position of object

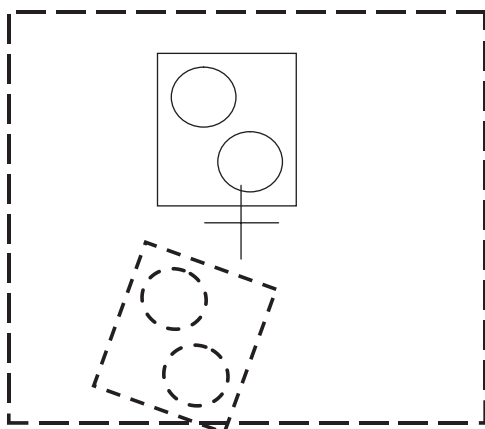


Fig. 11 Tracking window centered on predicted position of object

percentage of the whole image. For increased efficiency of image processing, incremental tracking is generally employed [12]. Incremental tracking is the approach whereby a small region around the last known position of the object (Fig. 10) or its predicted position (Fig. 11) is processed rather than the whole image, thereby decreasing the processing time or computational resources required. By limiting the objects to be tracked, it is possible to increase the incremental tracking window size and yet keep the vision processing time within the sample period of 16.67 ms. Using larger incremental tracking window sizes, the object being tracked is nearly always identified. In the event that the object is ‘lost’ (i.e. it is not inside the predicted tracking window), the fault tolerant software reverts to the full-tracking mode, whereby the whole image is analyzed to ‘recover’ the object’s position.

To locate the object positions the first time, the image must initially be scanned fully for one frame. This will identify the starting position of all the objects within the field of view. Then the incremental tracking algorithm can continue.

The number of pixels that must be processed is related to the size of the tracking window and the number of objects being tracked. This technique is significantly more efficient than full tracking if the sizes of the tracking windows are much smaller than the size of the image divided by the number of objects.

Reducing the sample period of the system can drastically reduce the required size of the tracking window, since the objects would have moved a shorter distance in the shorter sample time. In a sample case, halving the sample period would have the objects move half the distance, so the length of the tracking window can be halved. This gives an area reduction of a factor of four. Therefore doubling the frame rate (i.e. halving the sample period) would reduce the execution time of this algorithm by a factor of four. Thus, paradoxically, this algorithm is more likely to complete in the required sample time if the frame rate is higher.

In a system that has a reliable frame rate, the last known velocity of the object, rather than the last known position, can be used to centre the tracking window on the predicted position of the object. Using this method, the size of the tracking window can be reduced further as the error in the predicted position will depend only on the uncertainty in the measured velocity. The only risk associated with reducing the tracking window size relates to collisions, which can alter the velocity of the objects

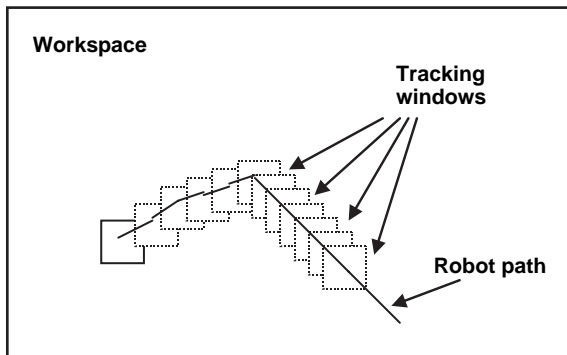


Fig. 12 Robot path and the incremental tracking window

significantly. In robot soccer, the ball is the lightest and fastest object and so is at the greatest risk of being 'lost'.

Several techniques to overcome this problem have been proposed in the literature [13] based on predicting collisions of fast moving light objects. Figure 12 shows the movement of the tracking window as the robot moves.

6.1 Adaptive Tracking Window Size

Enlarging the tracking window for the faster and lighter objects can produce reliable tracking results. A minimum tracking window size may be defined for a stationary object. As the object starts to move, the size of the tracking window will increase proportionately and adapt to the object's velocity. This will ensure that larger the velocity, bigger the tracking window size and lesser the possibility of 'losing' it. This will greatly increase the reliability of tracking. However, the downside is that even for a fixed number of objects, the number of pixels that need to be processed is not constant and can vary a lot depending on the individual object velocities. The variable number of pixels that will need to be processed creates an uncertainty in the total processing time per frame. This variability will have detrimental effect on the control of the robot motion as the sample time for error correction will no more be fixed.

7 A Fast Access Color Look-Up-Table (LUT)

7.1 Limitations of Using RGB Color Space

The blob detection algorithm can be implemented on any commodity vision system. The image digitization can be done using the FlashBus MV Pro frame grabber card

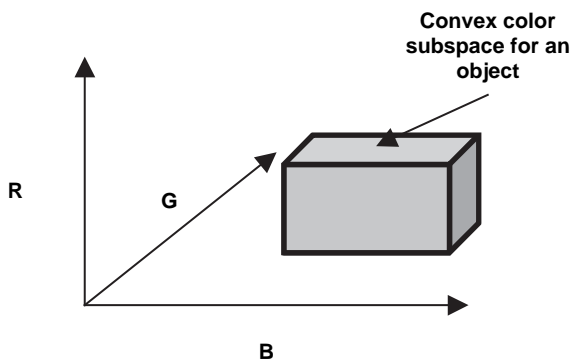


Fig. 13 Convex RGB color subspace

which provides pixel color information in RGB (Red, Green and Blue). A convex partition of the RGB color space can be created for each color identifier, as shown in Fig. 13. This convex partition (color subspace cube) is specified by a range of RGB values namely, MinR–MaxR, MinG–MaxG, and MinB–MaxB.

Blob identification based on RGB color space is not reliable in the face of varying light intensities as the luminance cannot be separated from chrominance [4]. In order to cater to a wide variation of light intensity, the volume of the color cube has to be extended. The drawback of doing this is that the color cubes of different colors will overlap and encroach into each other’s boundaries making it very difficult to segregate colors and at the same time detect them reliably. Instead a convex color subspace, defined in the YUV color space was implemented with greatly enhanced robustness (in respect of reliability of detection). The Y component independently corresponds to light intensity of the color and a wider threshold span can be set for it to cater to varying light intensities.

7.2 Defining YUV Thresholds

To define the YUV color subspace, a sample of the image is captured and the color of interest is zoomed in. In the zoomed image a rectangular region is defined, within which, each pixel is processed to calculate its YUV value using the color space transformation matrix which is shown below (2).

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

or alternatively as

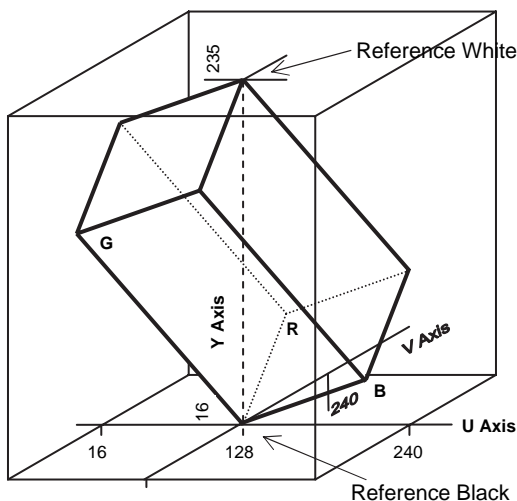


Fig. 14 Relative positioning of YUV and RGB color space

$$\begin{aligned}
 Y &= 0.299R + 0.587G + 0.114B \\
 U &= 0.565(B - Y) \\
 V &= 0.713(R - Y)
 \end{aligned}
 \tag{2}$$

The relative position and orientation of the RGB color cube in the YUV color space is shown in Fig. 14. From the computed YUV values, the MinY, MaxY, MinU, MaxU, MinV and MaxV are set. A user may manually fine-tune these thresholds using the application's GUI. Usually a wider range of Y values are desirable giving it more bandwidth to account for varying light intensity.

7.3 Membership Testing

For the two-pass sequential algorithm for blob-detection, each RGB color pixel of an image needs to be tested to determine its color sub-space membership. Since the color boundaries are defined in YUV color space, each pixel value would have to be converted from RGB to YUV, using (2), before testing membership using (3). This is a computationally intensive process and the overall performance could be quite low. Hence the mechanism used for thresholding warrants close scrutiny and requires careful efficiency consideration.

$$\begin{aligned}
 &\text{IF}((Y \geq \text{MinY}) \text{ AND } (Y \leq \text{MaxY}) \text{ AND} \\
 &\quad (U \geq \text{MinU}) \text{ AND } (U \leq \text{MaxU}) \text{ AND} \\
 &\quad (V \geq \text{MinV}) \text{ AND } (V \leq \text{MaxV})) \\
 &\text{THEN pixel_of_interest} = \text{TRUE}
 \end{aligned}
 \tag{3}$$

Equation (2) requires to perform 5 multiplications and the whole implementation of equation (3) may require up to 6 logical ANDing operations to determine whether a pixel belongs to a color subspace and is thus of interest. To improve the computational efficiency, implementations using Boolean valued decomposition of the multidimensional threshold have been tested [4] on static images resulting in substantial reduction in processing time. This, however, still requires the color space to be transformed from RGB to YUV. This method had not been tested previously by any researcher on live images in real time.

7.4 A One-Dimensional Color Look-Up-Table

An initial implementation in this research work used a large one-dimensional color look-up-table (LUT) and an indexing technique based on the RGB value of the pixel. The index is created, using the (4) below, which is used to access the LUT.

$$\text{index} = R * 65536 + G * 256 + B
 \tag{4}$$

For a 24-bit RGB color output from the frame grabber card, the maximum value of R, G or B is 255. Thus the size of the LUT is $256 \times 256 \times 256$ bytes (16 MB). For each RGB value, a unique index is created by the (4).

7.5 Posting the Look-Up-Table

Once the YUV thresholds have been defined for each color, the LUT is posted with color identities (IDs) for the entire RGB color space as shown in the code segment in Fig.15.

```

for (r=0; r<256; r++)
  for (g=0; g<256; g++)
    for (b=0; b<256; b++)
      {
        y=(299*r+587*g+114*b+500)/1000;
        u=(565*(b-y) + 128000)/1000;
        v=(713*(r-y) + 128000)/1000;
        index = r*65536 + g*256 + b;

        //-- initialise on update --
        LUT[index] = NoCOL;

        //-- Reference Colour range --
        if ( (MinY<=y && y<=MaxY) &&
            (MinU<=u && u<=MaxU) &&
            (MinV<=v && v<=MaxV))
          {
            LUT[index] = RefCOL;
          }
      }

```

Fig. 15 Posting the LUT with color ID

The time it takes to update the LUT is not of any consequence as the update is done during the color tuning phase. It is important that during the inspection time, the processing should not take unduly long and hence repeated multiplications and logical ANDing must be avoided.

7.6 Inspecting the Look-Up-Table

To test whether a pixel is in the YUV sub-space, given its RGB value, the index is calculated using (4) and the LUT content at that indexed location is tested as shown in the code segment in Fig.16.

```

index = r<<16 + g<<8 + b;
if ( LUT[index] == RefCol )
  //-- it is a desired pixel
  {
    //-- process the pixel
  }

```

Fig. 16 Inspecting the LUT

To further improve the processing speed, the multiplications in (4) were replaced by shift-left operations as in (5) below.

$$\text{index} = R \ll 16 + G \ll 8 + B \quad (5)$$

To classify each pixel in an image into one of a discrete number of color classes, the index is created from the pixel's RGB values and the LUT is queried. The returned value indicates color class membership.

The advantage of this method is that it does not require the RGB value of each pixel to be converted to YUV, which otherwise would take considerable processing time. However, this method has the following drawbacks-

- It takes very long to update the LUT. Using the experimental hardware setup detailed in Sect. 3, it took 909 ms to update the LUT. This eliminates the possibility of updating the LUT in a real-time processing environment and hence the thresholds defined for each discrete color class cannot be adapted to variations in light intensity.
- Because of the huge size of the LUT (16 Mbytes), the algorithm runs slower on a computer with a small cache memory, as there is frequent memory swapping. Nonetheless, in an image where the majority of the pixels belong to the background color class, this is not a significant problem as the same part of the LUT will be accessed most of the time.

8 Discrete YUV Look-Up-Table

Recent work has focused on efficiency issues so that effective classification can be provided in real-time. For simplicity of tuning, each color is classified with a pair of thresholds on each of the Y, U, and V axes as illustrated in Fig. 17.

Since each axis is independent, the large LUT may be decomposed into separate Y, U, and V LUTs of 256 elements each. The total size of all the discrete LUTs put together is only 768 bytes, greatly reducing the memory requirements compared to the LUT described in Sect. 7.4. For each array element, one bit is used to represent each color class as shown in Fig. 18.

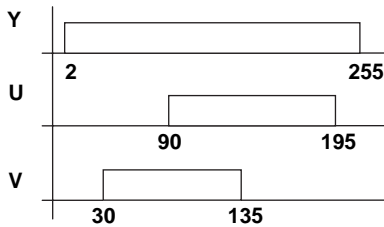


Fig. 17 YUV thresholds for color 3

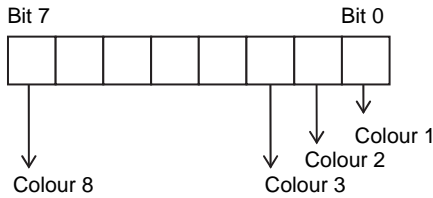


Fig. 18 Color representations in the LUT element

In programming terms, the color IDs are defined as-

```
#define colour1_ID 0x01
#define colour2_ID 0x02
#define colour3_ID 0x04
#define colour4_ID 0x08
#define colour5_ID 0x10
```

Using arrays of bytes, 9 different color classes can be represented (including 8 of interest and the background). To cater to more colors, the system can easily be scaled up to implement arrays of 16-bit or 32-bit integers.

Once the color thresholds have been defined, the YUV LUTs are then posted with the color IDs as shown in Fig. 20. The shaded cells store a value of 1.

8.1 Populating the Discrete YUV Look-Up-Table

After defining the YUV thresholds for every color class, each LUT is posted individually with color IDs. The code segment shown in Fig. 19 updates the Y-LUT. The U- and V-LUTs are similarly posted.

To update the three LUTs it takes only 8.2 μ s using the same experimental hardware setup. This enables the LUTs to be updated in real-time and hence may be used in implementing adaptive color thresholding.

```

for (y=0; y<=255; y++) {
    if ((y >= Col1_MinY) &&
        (y <= Col1_MaxY))
        Y_LUT[y] |= Colour1_ID;
    //-- repeat for other colours
}
    
```

Fig. 19 Posting the LUT with color ID

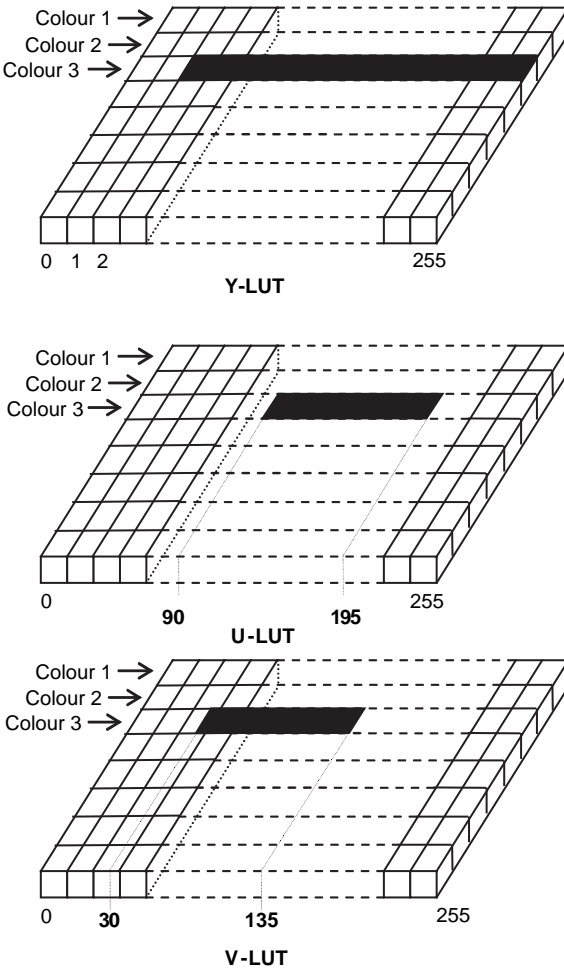


Fig. 20 YUV LUT populated for color 3 (not to exact scale)

8.2 Testing Color Class Membership

A pixel belongs to a color class only if it is within all three Y, U, and V ranges. The color class membership can therefore be computed as a bitwise AND of the elements of each component array. This is shown in the code segment in Fig. 21.

```

if (Y_LUT[Y] & U_LUT[U] & V_LUT[V] &
    Colour1_ID)
{
    //the pixel belongs to Color1 class
}
    
```

Fig. 21 Testing color class membership

The membership testing is very fast as the bitwise AND operation is computationally inexpensive. This method, however, requires the YUV values of each pixel to be available, which is often not the case for commodity hardware. Thus the advantage gained by using a smaller LUT is partly offset by the additional computation time required to map a pixel from RGB color space to YUV color space. In Sect. 8.3 two methods of speeding up this mapping are presented and the performance evaluation is detailed in Sect. 8.5.

8.3 Color Space Transformation

The standard transformation matrix of (2) for mapping RGB to YUV involves several floating point multiplications, which are potentially very time-consuming. The floating point arithmetic may however be replaced by integer operations as in (6) below.

$$\begin{aligned}
 Y &= (299R + 587G + 114B)/1000 \\
 U &= 565(B - Y)/1000 + 128 \\
 V &= 713(R - Y)/1000 + 128
 \end{aligned}
 \tag{6}$$

The U and V components are offset by 128 to bring them into positive range to facilitate array indexing. Equations (6) still use division operations. The division operations may be eliminated by scaling the coefficients by powers of 2 rather than powers of 10, allowing the use of a computationally less expensive shift-right operation as shown in (7) below.

$$\begin{aligned}
 Y &= (9798R + 19235G + 3736B) \gg 15 \\
 U &= 18514(B - Y) \gg 15 + 128 \\
 V &= 23364(R - Y) \gg 15 + 128
 \end{aligned}
 \tag{7}$$

8.4 A New Color Space

This research makes a significant contribution by proposing a novel Y'U'V' color space. This new color space not only retains all the colors of the RGB color cube, it actually increases the volume of the YUV color cube, thereby enhancing the resolution of spatial color separation. The proposed transformation is governed by (8):

$$\begin{bmatrix} Y' \\ U' \\ V' \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (8)$$

The new Y'U'V' color space has several advantages over the standard YUV space represented by (2).

- Computationally it is very inexpensive. Only integer additions and subtractions are involved; all floating point operations and logical shift operations have been completely eliminated.
- The white point corresponds to equal quantities of R, G, and B.
- The new Y'U'V' color space provides better resolution. Without scaling, the Y' range is 0 to 765, U' range is from -510 to 510 and V' range is from -255 to 255. This enables colors that are closer together to be detected reliably.
- The Y', U' and V' axes are orthogonal, making the transformation back to RGB similarly simple. It also gives better decorrelation of the color space for many images.

This larger LUTs (2298 elements as compared to 768) increase the LUT update time to 19.2 μ s. This is still extremely fast and causes no concern for real-time update of the LUT.

8.5 Experimental Results and Discussion

The proposed improvements to the methods of transforming from RGB to YUV color space and the new Y'U'V' color space were evaluated using the robot soccer system, which offered the possibility of testing the algorithms for real-time processing with differing number of objects. A very high precision counter was implemented in the software which enabled measurement of time with an accuracy of a hundredth of a milli-second. Tests were done with 4 objects (3 home robots and ball), 7 objects (3 home robots, 3 opponent robots and ball) and 11 objects (5 home robots, 5 opponent robots and ball) for incremental and full tracking. To evaluate the robustness, the tests were done for incremental as well as for full tracking modes. The test results are summarized in Table 1, and compared in Figs. 22 and 23. The tracking times were measured for 1000 frames and averaged.

Table 1 Summary of test results

System	#Objects	Average tracking time (ms)		LUT update time
		Incremental	Full	
#1	4	5.27	15.54	909 ms
	7	5.62	20.97	
	11	5.95	28.34	
#2	4	5.38	21.89	8.2 μ s
	7	5.64	30.37	
	11	5.99	41.72	
#3	4	5.34	17.22	8.2 μ s
	7	5.56	23.41	
	11	5.85	31.68	
#4	4	5.15	14.34	19.2 μ s
	7	5.38	19.22	
	11	5.68	25.76	

With respect to the data presented in Table 1, the various systems are as follows:

- System #1: Large composite LUT indexed using RGB – (5)
- System #2: Separate YUV LUTs, with integer division used to map RGB to YUV – (6)
- System #3: Separate YUV LUTs using the >> operation to map RGB to YUV – (7)
- System #4: New color space (Y’U’V’) and separate Y’U’V’ LUTs – (8)

The radar charts in Figs. 24 and 25 show the vision processing time taken by the different systems for incremental and full tracking respectively for 4, 7 and 11 objects. The system using Y’U’V’ color space and discrete look-up-tables takes the least amount of time.

Figures 26 and 27 highlight the relative improvements achieved in the vision processing time by using the new Y’U’V’ color space for different number of objects.

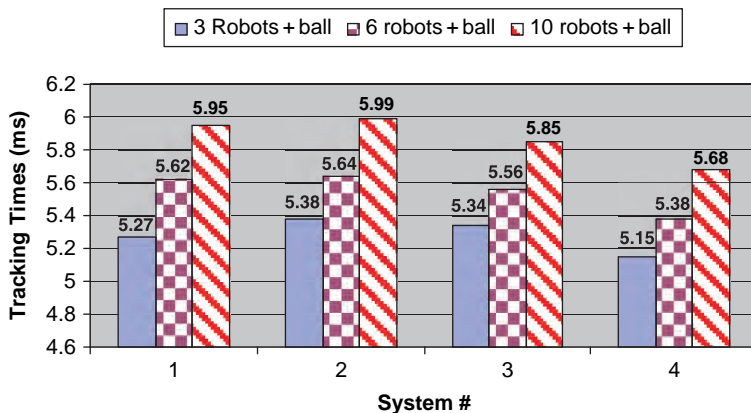


Fig. 22 Comparison of incremental tracking time

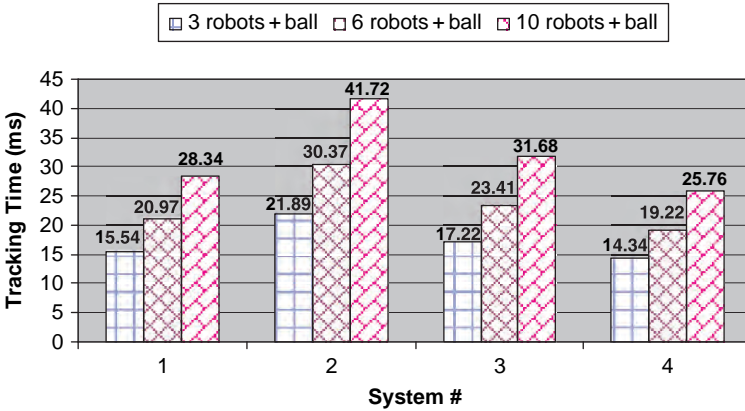


Fig. 23 Comparison of full tracking time

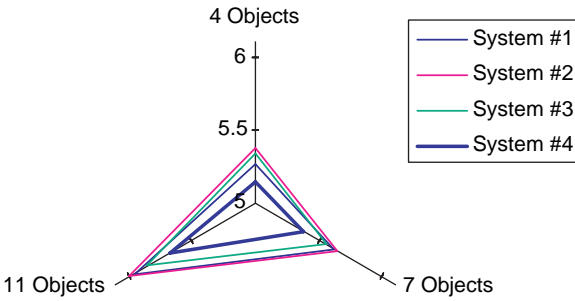


Fig. 24 Chart showing the performance of the four systems for incremental tracking

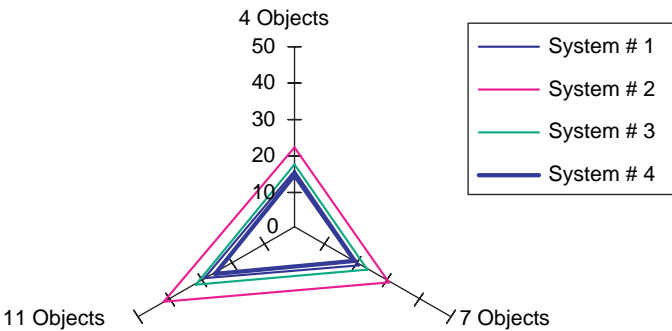


Fig. 25 Chart showing the performance of the four systems for full tracking

The improvements reported are with respect to the other three systems. As can be seen, the improvements are substantial.

To summarize the experimental results, it can be said that an efficient arrangement of discrete Y, U and V LUTs for fast color segmentation has been proposed and tested for real-time vision processing applications. A significant reduction in

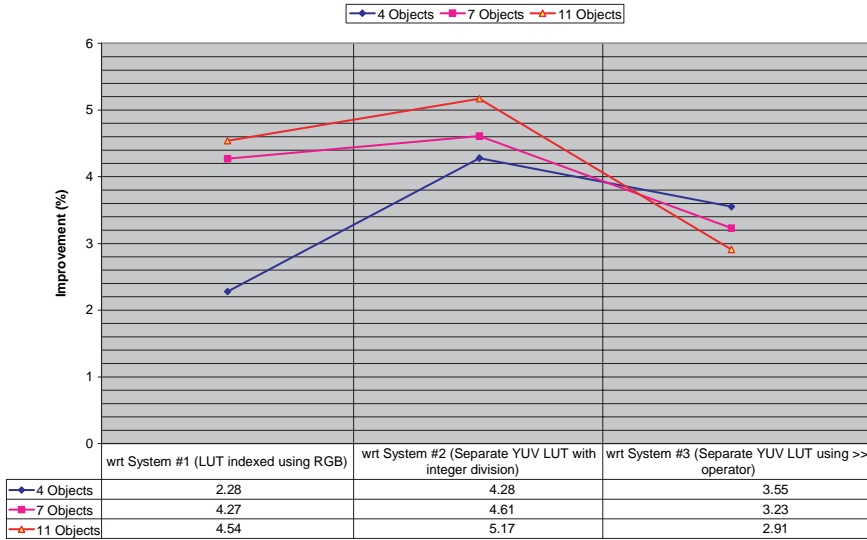


Fig. 26 Summary of improvements in processing speed (in %) achieved by Y'U'V' for incremental tracking

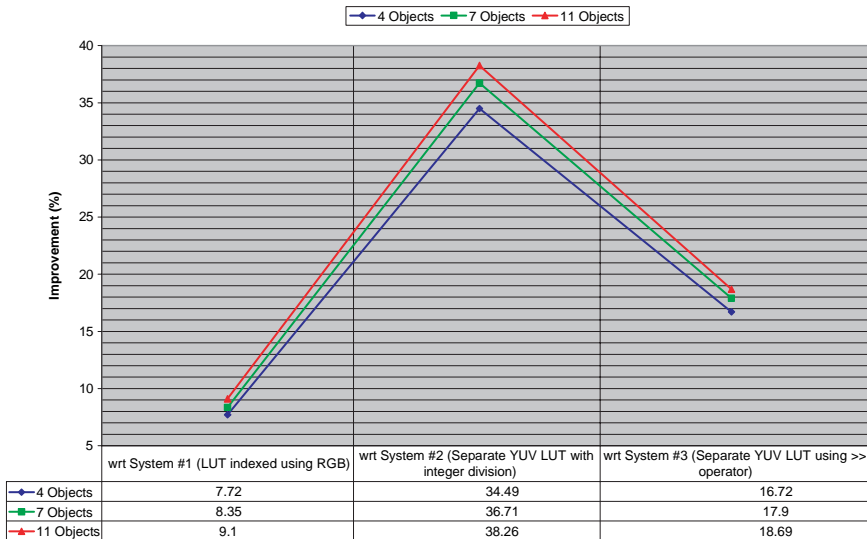


Fig. 27 Summary of improvements in processing speed (in %) achieved by Y'U'V' for full tracking

LUT size (and hence the memory requirement) has been achieved. This translates into large increase in program execution speed for incremental and full tracking of multiple objects, especially on processors with small cache. The time to update the discrete LUT is negligible (8.2 μ s for YUV and 19.2 μ s for Y'U'V') and hence is very suitable for real-time update. This is a vast improvement over the method that

employs a large LUT with indexing using RGB and takes 909 ms to update the LUT. This has laid the foundation which will enable further work to be done to make the LUT ‘adaptive’ in order to cater to variation of light intensities and color distortions, possibly due to reflections from nearby objects.

To enhance the gains derived from the discrete YUV LUTs, the proposed new color space, $Y'U'V'$, further simplifies the transformation from RGB to a YUV-like color space. The time to fully track 7 objects reduces from 30.37 ms (using YUV LUT) to 19.22 ms (using $Y'U'V'$ LUT), which is an improvement of 36.71%.

Furthermore, the $Y'U'V'$ color space allows better color resolution, thereby increasing the robustness of color classification. The results compare favorably to the color threshold based approaches discussed in [7].

References

1. Berthold K.P. Horn, *Robot Vision* (MIT Electrical Engineering and Computer Science), MIT Press, 1986.
2. M. Jamzad, B.S. Sadjad, V.S. Mirrokni, M. Kazemi, H. Chitsaz, A. Heydarnoori, M.T. Hajiaghahi, and E. Chiniforoosh, “A Fast Vision System for Middle Size Robots in RoboCup”, A. Birk, S. Coradeschi, S. Tadokoro (Eds.): *RoboCup 2001: Robot Soccer World Cup V*, Springer Verlag 2002, pp. 71–80.
3. C.H. Messom, S. Demidenko, K. Subramaniam, and G. Sen Gupta, “Size/Position Identification in Real-Time Image processing using Run Length Encoding”, IMTC, Alaska, USA, 2002, pp. 1055–1059.
4. J. Bruce, T. Balch, and M. Veloso, “Fast and Inexpensive Color Image Segmentation for Interactive Robots”, IROS 2000, San Francisco, 2000, pp. 2061–2066.
5. F. Ercal, M. Allen, and F. Hao, “A Systolic Image Difference Algorithm for RLE-Compressed Images”, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 11, No. 5, May 2000, pp. 433–443.
6. H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara, “RoboCup: A Challenge Problem for AI”, *RoboCup-97: Robot Soccer World Cup I*, Springer Verlag, London, 1998, pp. 1–19.
7. J. Baltes. “Practical camera and colour calibration for large rooms”, Manuela Veloso, Enrico Pagello, and Hiroaki Kitano (Eds.): *RoboCup-99: Robot Soccer World Cup III*, pages 148–161, New York, 2000. Springer, pp. 148–161.
8. M. Ramesh Jain, R. Kasturi, and B.G. Schunck, *Machine Vision*, McGraw-Hill International Editions, Computer Science Series, International Edition, 1995.
9. Sangho Park and J.K. Aggarwal, “Segmentation and tracking of interacting human body parts under occlusion and shadowing”, *Proceedings of the workshop on Motion and Video Computing*, 5–6 Decemcer 2002, pp. 105–111, ISBN: 0-7695-1860-5.
10. C. Garcia, and X. Apostolidis, “Text detection and segmentation in complex color images”, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00*, 06/05/2000 – 06/09/2000, Istanbul, Turkey, pp. 2326–2329, ISBN: 0-7803-6293-4.
11. M. Shand, “Flexible image acquisition using reconfigurable hardware”, *IEEE Symposium on FPGA's for Custom Computing Machines (FCCM '95)*, 1995, pp. 0125.
12. C.H. Messom, G. Sen Gupta, and H.L. Sng, “Distributed Real-time Image Processing for a Dual Camera System”, *CIRAS 2001*, Singapore, 2001, pp. 53–59.
13. A. Chakravarthy and D. Ghose “Obstacle avoidance in a dynamic environment: a collision cone approach”, *IEEE Transactions on Systems, Man and Cybernetics, Part A*, September 1998, Vol. 28, No. 5, pp. 562–574, ISSN: 1083–4427.

Part VIII
Sensors Based on Human Parameter

Affection Based Multi-robot Team Work

Sajal Chandra Banik, Keigo Watanabe, Maki K. Habib and Kiyotaka Izumi

Abstract Multi-robot task allocation, cooperation and interaction among the members of a team are very complex topics that need to be explored more. A task can be accomplished by a multi-robot team with required performance and reliability being operated with a proper cooperative plan. A proper cooperative plan includes an intelligent task allocation method in a productive and efficient manner such that the assigned task to the team is performed with a level of performance satisfaction. The robots need to be intelligent enough to dynamically adjust with changing workload either by changing actions or by making new cooperative plan. In this chapter, we describe proposed approaches to multi-robot task allocation and cooperation in a chronological way such that they can be studied and compared for future development with affection based augmentation. In respect of some drawbacks (like high communication overhead, dead lock, etc) with the existing approaches, we present the affection based task allocation and cooperation that has been used for a very few cases. We also present the complexity of the affective method and give some hints to compensate the complexity problems. Later on, we present also a stochastic approach for affection based task allocation, cooperation and interaction for a multi-robot team.

Sajal Chandra Banik

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan,
e-mail: baniksajal@yahoo.com

Keigo Watanabe

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan,
e-mail: watanabe@me.saga-u.ac.jp

Maki K. Habib

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan. Currently with the American University in Cairo, Egypt, e-mail: maki@ieee.org

Kiyotaka Izumi

Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University, 1-Honjomachi, Saga 840-8502, Japan,
e-mail: izumi@me.saga-u.ac.jp

1 Introduction

Multi-robot system is one of the main topics in research area to application field having a variety in structure, team size, goals and application domains. In many cases, a significant benefit like: reliability, performance and economic value can be had by engaging multi-robot system instead of a single robot. In addition to that, a team of robot gives a good level of robustness, fault tolerance and flexibility because the failure of one robot does not break the common goal of the team due to task sharing by other robots. Emotions have important roles in intelligence, planning, learning, interactions, perception, creativity and more [1]. In an organization or in a team, a lot of importance is given on the emotional state of members of a team and in this way each one behaves to others and can understand the emotional state of others. With the ongoing research on emotion and its application for artificial agents, some researchers have already agreed that a multiagent (pure or mixed agent) system with emotional facts can have the same advantages as emotion brings to human team [2].

In some cases, we use multi-robot system to distribute the activities and intelligence among the members depending on the complexity of problems. Sometimes it is needed to divide a complex task into small tasks and distribute the small tasks to members of team when problems are widely distributed and heterogeneous in functional terms. With the limited ability and knowledge, a robot can have a satisfactory role by performing the assigned small task with high performance. When working in a group, a robot needs to develop intelligent behavior with the artificial intelligence. On the contrary, artificial intelligence is facing some critical problems in projecting the common senses into knowledge with some rules. In general, a truly autonomous robot should develop its rules that govern its behavior. If one expands the concept of autonomy by including self motivation, then emotions might play a role because these are considered to be essential for human team with reasoning.

Cooperative multi-robot systems need to have adaptability to a changing environment, flexibility of responses to various tasks, better performance and easier interaction among the robots. With affection based intelligence a robot can make a cooperative plan in a more advanced way to decide its goal specific actions to be performed and also makes a request to others with task allocated information. A robot having emotional intelligence will be able to work in a mixed agent system (a system of multi-robot and human) to cooperate and to interact like a human with its full extent.

The concept of artificial emotion is expanding and increasingly used to design autonomous robots with the augmented capability like emotion based experience of environment, emotional interaction, etc. [3]. Recently, researchers are very interested in developing robot with emotional intelligence even for multi-robot system to make it more autonomous and efficient. Although, it is a very complex task to realize the emotion creating, expressing and understanding process; with some simplicity and basic emotional rationality, it is possible to introduce emotional intelligence even for multi-robot system that can work like human team somehow. The emotional system and the action-behavior selection process will depend on the taxonomy of the multi-robot system. Before going to discuss about affection based

multi-robot team (MRT), it will be helpful to give a brief idea about the taxonomy of MRT prevailing in the robotics literature.

2 Taxonomy of MRT

A cooperative MRT has the ability to perform tasks independently or by cooperation. The cooperation among the members of an MRT can be of two types: implicit cooperation and explicit cooperation [4]. For the implicit one, each team has a common goal and each member performs its individual task independently and the collection of individual task is targeted to satisfy the common goal. For example, we can say if an MRT is engaged in cleaning a big area and the area is divided among the members to be cleaned, then each member is performing the task of individual area cleaning which collectively satisfies the common goal of cleaning the whole area. This type of cooperation is also called asynchronous cooperation, because it need not synchronize in time and/or space while performing task. For the explicit case, each robot is in synchronized state while performing a task. For example, in the case of weight lifting each robot takes a correct position simultaneously with the helper one and then need to hold, lift and release the object with synchronization.

To perform a task completely, it needs a proper task allocation (task assignment) process for a multi-robot system. Task allocation process solves many task related problems (like which robot will perform a task? Which task will be done by a robot? When the task will be done? Where should be the task performed? Does it need implicit/explicit cooperation?). The task allocation procedure depends on team composition, team size, communication style, cooperation level, etc. Multi robot task allocation problem includes mainly four types of basic strategies [5]:

Auction: In this method, task is announced among the members of the team and each robot returns a bid specifying the fitness to perform the task. A best-fit selection algorithm is used to select the best robot for the task. Gerkey and Mataric presented an auction-based task allocation system called as MURDOCH where auction protocol follows some sequence of steps like (for more details see [6, 7, 8]) broadcasting of task, metric evaluation, submission of bid, close of auction and progress monitoring or renewal of contract (if necessary). In case of affection based MRT, the best fit selection function algorithm can be based on emotional state of each robot. Because, emotional state reflects the internal condition as well as the environmental changes.

Motivation based: In this approach, a suitable action is selected through some internal motivation mechanism. Parker's ALLIANCE is one of the available architectures that has used a fault tolerant behavior-based architecture where robots choose tasks by two motivational mechanisms named as impatience and acquiescence (for more details see [9, 10]).

Team agreement: Each member of the team works under a general agreement targeting to a goal. Chaimowicz et al. used this approach for the coordination among

the members of RoboCup team [11]. For the case of affection based robots, a common agreement can be done to make each one ‘happy’ by assisting each other if necessary.

Broadcasting of local eligibility: In this approach, each robot in the team determines its own utility to accomplish the available task and the task allocation process depends on the local efficiency of a robot. The most efficient robot directly inhibits the other robots surrounding it, takes the liability of the task and performs the task. Werger and Mataric have used such kind of task allocation method as described in [12, 13]. In affection based MRT, the members of the team may be in different emotional state. A robot with good emotional state can take the responsibility of a task and inhibits those who have no emotional fitness (for example, one robot with sad or distress may have poor performance in working and if it is allowed to perform a task then outcome will be poor, so it is better to inhibit the sad robot from doing task for a while).

There are also some other task allocation approaches that have been applied for some specific purposes. For example, an emergency handling approach has been used in [14] where robots respond to an audible alarm and follow the sound gradient to its source. Another famous approach is plan-merging protocol, where each robot individually makes plan and then these plans are merged into a directed acyclic graph (DAG) to resolve temporal constrains (for more details see [15, 16]).

A taxonomy of MRT can be developed based on cooperative technique and system properties that affect on team development. Farinelli *et al.* has proposed a taxonomy of MRS (multi-robot system) based on coordination dimensions and system dimension [17]. The coordination dimensions include four dimensions like cooperation level, knowledge level, coordination level and organization level (see Fig. 1). A cooperative system is composed of compliant and/or benevolent agents that willingly perform tasks to satisfy a global or common goal. It is better for an agent to have knowledge about its team members (although unaware robots have no

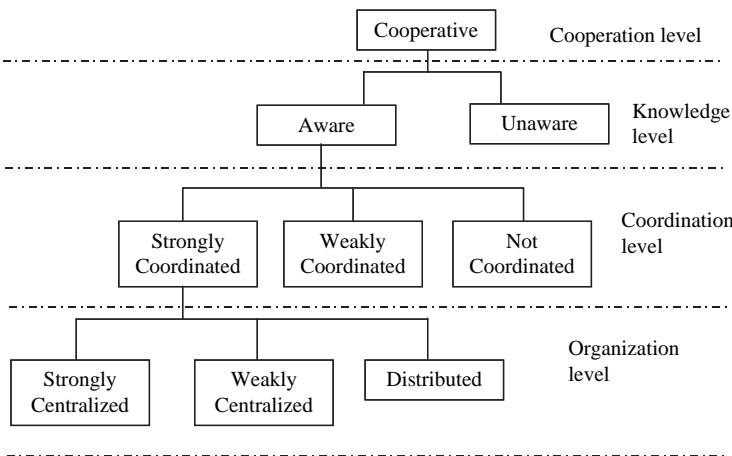


Fig. 1 Taxonomy of MRS according to [17]

knowledge of others in the cooperating team [17]). To achieve an effective cooperation, it needs a proper cooperative procedure or coordination protocol that is followed by each robot during cooperation. A cooperative system may have different types of organization levels for decision making such as centralized (strongly/weakly) and distributed ways.

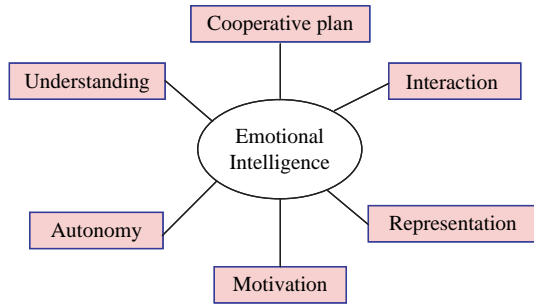
Farinelli *et al.* also grouped system features for including in system dimension which includes communication, team composition, system architecture and team size. To make an interaction and cooperation among the team members, it is necessary to have an easier communication technique that depends on various system features such as degree of cooperation, team composition which may be heterogeneous or homogeneous (for more details see [18]), team size, team architecture, etc. To know about the impact of communication capabilities on system and details of communication, the work of Dudek *et al.* [19, 20] can be referred.

Affection based system relates with emotion, feelings or mood of an entity of the system. To develop an affection based MRT, many questions arise that it has to be solved for a particular system. What is the scientific framework for approaching ‘emotion’ for a MRT? How many and what emotions are to be applied for a system? How to integrate the emotion system to other system, such as learning, sensory, action, communication, etc.? What are the computational mechanisms that reflect the complexity existing in emotional process? What kind of emotional model can be used suitably for robot’s/agent’s performance? To what extent these models can replicate the biological phenomena behind emotion generation? Although these types of questions are not so easy to reply from any theoretical or computational concepts, the emotional concept can be applied to MRT system with some assumption and simplicity. This simplicity will depend on the taxonomy of MRT. In the next section, we try to define ‘emotion’ and discuss its perspective to robotic system. We also describe the emotional intelligence that may be applicable for MRT to develop a more autonomous and dexterous system.

3 Emotion and Emotional Intelligence

Until now, there is no concrete and universally accepted definition of emotion that can be considered as the ideal one to be used in robotic system. Researchers are using these phenomena with their inventive ideas for different cases. Emotion is a complex biological process within the brain and body that can be also created artificially and then can be applied for robotic system. Nowadays, ‘being emotional is not good or irrational’ –such kind of talking is useless. If we think about the beneficial aspects of emotion that prevail in biological entities and if we can create these emotional aspects artificially, then this artificial emotional system will augment the robots with similar benefits like emotional intelligence, thinking, planning, creativity, robust decision-making, etc. M. Minsky has written in [21], “The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions”. From the perspective of MRT, we find some most important beneficial features from emotional intelligence as shown in Fig. 2.

Fig. 2 Beneficial effects of emotional intelligence



S. H. Kenyon has defined emotional robot as “An emotional robot would be a cognitively and physiologically biomimetic machine. The ‘body’ of the robot, including all sensors and actuators must be included in the design of the emotion system” [22]. He also stated that emotions are very closely related to an organism’s internal state, i.e. what it feels inside.

Some researchers also differentiated reactions, emotions and moods based on time scale as shown in Fig. 3. Reactions are created for very short period, whereas emotions stay for longer period, but moods being longer period than emotions [23]. Normally, emotions are elicited from stimuli and psychological phenomenon of short-period, whereas a mood is disseminated and long-lasting phenomena. Mood can also be described as “an emotional state, perhaps low intensity, capable of lasting for many minutes or several hours” [24].

The complexities in MRT coordination are of many such as: to find a better communication technique by removing/reducing communication traffic, to find an easier interaction among the agents whether it is pure agent system or mixed agent system that consists of human and robot, a better representation of each agent to others and also easier understanding among the members, removing deadlock and unexpected delays and thus increasing performance, etc. To resolve some of the problems, some researchers have already applied affection based control mechanism for multi-robot system and have got some advantages (for more information see [5, 25]).

There are two main areas of research where emotion is being used [26]: human-robot interaction [1, 27, 28] and affection based internal architecture [29, 30, 31, 32]. First one deals with the interaction method among human and robots/machines and bringing improvement by introducing emotion. The second one deals with modeling

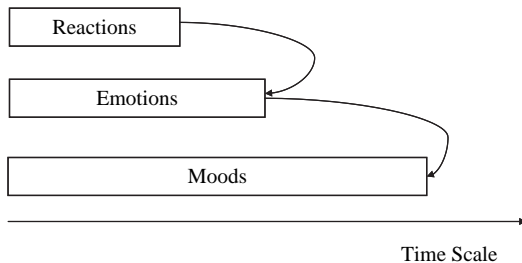


Fig. 3 Moods, emotions and reactions with sustaining time

and computing of emotional architecture to be applied in robots/machines control process. In [26], emotional research project described as “in general, emotion-based projects expect that including an emotion model into computational system they can improve machine performance in terms of decision-making competence, action selection, behavior control and autonomous and trustworthy system response”. Before going to application of emotion, we need to have knowledge of emotional states and their varieties. The following section will briefly discuss the various emotional states.

4 Emotional States

To describe emotional states, there are many theories and opinions given by theorists and researchers. Some believe on discrete state of emotions whereas some believe on continuous dimension of emotion. Both concepts have individual advantages and different applications. R. W. Picard has also given her consent about this in *Affective Computing* book [1] as “The question of whether to try to represent emotions with discrete categories or continuous dimensions can be considered a choice, as each representation has advantages in different applications. The choice of discrete or continuous states is, in one sense, like the choice of particles or waves in describing light: the best choice depends on what you are trying to explain”.

To characterize emotion in continuous dimensional space, there are two mostly used dimensions like *arousal* (calm/excited) and *valence* (positive/negative). Basic emotions can be placed in the two dimensional space defined by these two dimensions, though there are some limitations. For example, ‘fear’ and ‘anger’ both emotions have negative valence and high arousal. Sometimes another third dimension is included named as *potency* (powerfulness/powerlessness) to minimize this coherency problem.

Modern concept says that emotional states are the brain states which can rapidly assign value or valence to the result of consequences and can provide a plan for systematic action. R. Plutchik has classified emotions into eight basic emotions such as anger, fear, sad, joy, disgust, surprise, curiosity and acceptance [33]. Some researchers also say that there is similarity in concept with ‘prime colors’ and ‘basic emotions’. With the blending of basic emotions the full spectrum of emotions can be obtained. Each basic emotion is related with some specific behaviors with different survival value, for example, fear motivates to flight away and anger motivates to fight against for survival. According to Ortony, Clore and Collins, the most common basic emotions are fear, anger, sadness and joy and the next most common are disgust and surprise [34]. In case of emotional application to MRT/multiagent system (MAS), it is unwise to include many emotions unnecessarily which increase complexity in computations. Selection of emotions depends on the system structure, usability, task type, environment, etc. Even, sometimes one emotion is sufficient to develop an affection based system as used in [5] by A. Gage in his task allocation problem for multi-robot.

5 Emotional Roles in MAS/MRT

There are many psychological evidences supporting the emotional concept to be needed for getting automation in agents. Now it is a matter of thinking whether emotions could have the same functional roles for artificial system (like multirobot system) as those prevailing in natural system. Scheutz [3] has found such 12 roles of emotions that can be used for artificial agents (may be for single agent or multi-agents system) to develop emotional control mechanism. Some of them which are important for MAS/MAT are described bellow:

Adaptation: Emotion can play a role in short or long term behavior changes to adapt with dynamic environment under several constraints (like time limitation, resource limitation, etc.), especially for the environment which can not be predicted perfectly [24, 33]. A multiagent system is usually engaged in performing complex task where some uncertainty exists in the working environment. For this a complete premature plan does not work well, whereas it needs a step by step plan and then responses to new environment. The adaptive behavior can emerge through the current emotional state of each agent which can make the agent more autonomous.

Action selection: A behavior is a set of some actions. An agent can select an action (what to do next) based on the present emotional state to show its adaptive behavior to adjust with the contingencies of the world.

Managing social regulation: In multiagent system, it is required to develop an easier method for interaction and communication which can be achieved with some emotion based concepts. Although some emotional expressions and their understanding have already been developed, but it is for only limited cases and most of them for exposing and interacting agents. In [4], Plutchik stated that emotions are functional adaptations to corroborate a kind of social interaction. Emotion also plays a great role to communicate links in mutual plans among the individuals in a social group [35].

Sensory integration: Sometimes, it is also possible to filter or to reorganize data preceding based on emotional situation. Some strong emotional state will prohibit some minor feeling/senses or will consider it for next step giving priority to the present situation which causes the present emotional state. For example, if someone is seriously wounded in leg with high pain (and being very sad) and at that time if any ant or insect bites him at any place in the body, he will not bother even not feel it at that moment due to high concentration in his leg. But if he is sitting with normal condition, he will definitely response to the bite of ant or insect. This is one kind of sense-filtration process by emotional state.

Motivation and learning: Several motives can be created as an integral part of emotional coping mechanism [3]. The current emotional state as well as past history affects on the motive-creating and also on learning mechanism. In reinforcement learning mechanism, the environment or situation can be evaluated by emotion based score which can be used as Q-values.

By studying the various theories of emotion, the roles of emotion are summarized as follows:

- Emotions are the reason of various actions.
- Goal oriented action is also affected by emotion.
- Emotion enables adaptation to any harmful condition prevailing in any process or in any environment without having to reason about the cause [25].
- Sometimes emotion can also be goal.
- Emotion can be triggered by cognition process. Behavior is affected by emotion and also vice versa [36].

With taking the advantages from emotional roles, an MRT can increase overall performance by managing a proper task assignment process, by using a proper cooperative plan and traffic free communication system. In Sect. 2, we have described some of the emotional based issues for task allocation problems. The application of emotion for MRT is not yet well established by having some conceptual problems (for more details see Sect. 4 of [37]), though some researchers have already applied emotion for MRT/MAS as in [2, 5, 25, 38] with different computational models and mechanisms. Some researchers have also studied with the communication system for MRT in respect to performance evaluation in a variety of tasks and have found that the communication provides some advantages and for some cases, even a small amount of communication can render a great benefit [39, 40]. But for some cases, such as interactive agents (for example, robots performing theater or drama), human-robot interaction (for example, personal robots, dental robots, nursing robots, robot prototyping or resembling a patient like human that used by internee doctors for operation), mixed agent system (composed of human and robot) that works as a team, etc. it would be better and easier if the communication system is developed on some emotional phenomena.

6 Development of Affection Based MRT

In recent years, numerous affection based robotic systems have been developed arising many questions and giving hints about the bottlenecks of the emotion based system. Are there any common definitions of emotions for modeling? What is the fundamental mechanism of emotions playing in cognition, action and their integration? What kind of emotion model should be used for an application? What is the best way for sensing, recognizing and expressing emotion during interaction among agents? These types of inquiries are commonly arising and some of them are demolished during the procedure of modeling and implementation.

To develop emotion based architecture for robotic system, a number of approaches have been used to increase the autonomy and adaptation in the working environment. For adaptation, usually two types of adaptation are considered: short term adaptation and long term adaptation (for more details see [41]). For the first type of adaptation, robots need to suit with short term changes in environment with rapid decision and action selection. To suit with the long lasting changes of environment, robots require the ability to update behavior over time which needs learning and memory control mechanism. Cañamero [37] has discussed some computational

models that have been developed to design emotion for action-behavior control, emergent emotional behavior, and learning and memory control. Now, we will discuss some of the well-known computational models of emotions that have been developed to be used in various artificial intelligent fields.

Cathexis Model: Cathexis model is proposed by Velásquez [42] considering six basic emotions: anger, fear, distress/sadness, enjoyment/happiness, disgust and surprise. This model considered several aspects of emotion process like neurophysiology, sensory motor aspect (such as body movement, facial expression, gestures and postures, etc.), motivational states and even appraisals. This model did not introduce adaptation in emotion modeling, therefore, it is not a flexible model and there is no integration of personality in the model.

This model consists of a network of nodes where nodes are the ‘proto-specialists’ representing the different emotion types. Each ‘proto-specialist’ can detect internal and external stimuli having different sensors. The stimuli detected by the sensors can activate the emotion of ‘proto-specialist’ or change the emotional intensity. A schematic overview of the system is shown in Fig. 4.

Elliott’s Affective Reasoner: Elliot’s affective reasoner [43] is a multi-agent model of emotions through the adaptation of OCC model [34] which is capable of producing 24 emotions including some directive emotions like happy-for, sorry-for, gratitude, etc. This model can simulate in a nice way describing emotion generation, emotional expression and social interactions, but this model still faces some difficulties. The *Affective reasoner* is not a quantitative model: i.e., does not consider the intensities of emotions. This model is also limited by the use of domain-specific rules for appraising events.

FLAME: FLAME model [44] is a computational model of emotions based on fuzzy logic, considering an even-appraisal method which is composed of three major components: an emotional component, a learning component and a decision making component (as shown in Fig. 5). The learning component increases the adaptation in modeling emotions. It also has an emotion filtering component which can resolve conflicting emotion by considering motivational states. This is also an

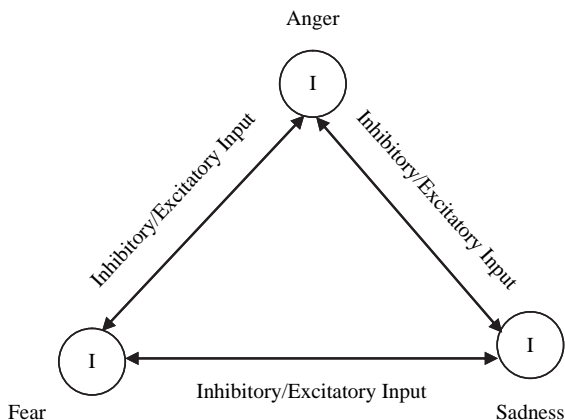
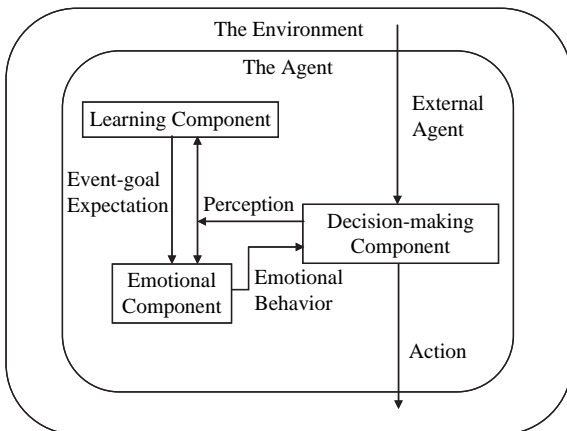


Fig. 4 An overview of Cathexis model for anger, fear and sadness ‘proto-specialists’ where ‘I’ represents the intensity of the respective emotion

Fig. 5 Agent architecture of FLAME-An overview (adopted from [44])



inflexible model (because it uses a predefined reward value for the impact value of user’s action on an agent’s goal) and personality is not considered.

Bates’ OZ Project: J. Bates in his OZ project [45] built believable agents following an even-appraisal model in Ortony *et al.*’s work [34] to provide users the exciting experience of living in a dramatically interesting micro-world composed of emotional agents. The emotional state is governed by the rules of OCC model (the emotion and their causes are shown in Table 1). This project has addressed many important emotional aspects; but still has some limitation - especially with the emotional synthesis processes that use expectation values statically derived from some predefined rules. However, in reality expectation values are not fixed: the expectation is changed with experience of each agent.

Gratch’s Émile: In the Émile [46] model, Gratch has used a classical method for detecting and resolving threats to appraise the emotional significance of events. He

Table 1 Emotions and their causes used in OZ project [45]

Emotion	Cause
Joy	Goal success.
Distress	Goal failure.
Hope	Prospect of goal success.
Fear	Prospect of goal failure.
Pride	Action of self approved according to standards.
Shame	Action of self disapproved according to standards.
Admiration	Action of other approved according to standards.
Reproach	Action of other disapproved according to standards.
Love	Attention to liked object.
Hate	Attention to disliked object.
Gratification	Action of self causes joy and pride.
Gratitude	Action of other causes joy and admiration.
Remorse	Action of self causes distress and shame.
Anger	Action of other causes distress and reproach.

proposed a simplified way to calculate a probability of goal achievement: the probability of threats to a goal, and how much it has importance of emotional effect. There are some problems with the threat detecting approach because it would mistreat an event which is both establisher and threat to the agent's goal [46, 47]. It also does not consider the value of event unexpectedness during the calculation of even-based emotion and emotional intensity. Émile also does not consider about the way of motivational states and personality influence on emotion.

ParleE: ParleE is a quantitative, flexible and adaptive model of emotions suitable for an embodied agent to be believable, placed in a multi-agent environment [47]. It generates emotions based on OCC model. This model has given attention to the integration of personality and motivational states and their roles in emotion generation. By the level of appraisal, this model can distinguish between moods and emotions adopting the concept of two thresholds associated with each emotion for activation of emotion type and saturation of emotion. Motivation states influence emotions by operating the threshold values for each emotion. An overview of ParleE system is shown in Fig. 6. Planner produces a plan to achieve a goal and calculates the probability of success that is supplied to the emotional appraisal component to generate an emotion impulse vector (EIV).

Markovian Emotion Model: The Markovian emotion model is a stochastic model of emotion where nodes represent some pre-defined states and the arcs show the probabilities of transition between the emotional states as shown in Fig. 7. Discrete-state homogeneous Markov model is very suitable to model human emotion as well as to clone human emotion in believable agents [48]. The emotional state is derived from the present state and this model has memoryless property. Markov model is very suitable for modeling emotion because behaviors highly depend on emotional present state rather than the past history. K. Kühnlenz and B. Martin have proposed an emotion core for autonomous robots based on hidden Markov model where emotional states are represented by hidden states [49]. State transitions are possible due

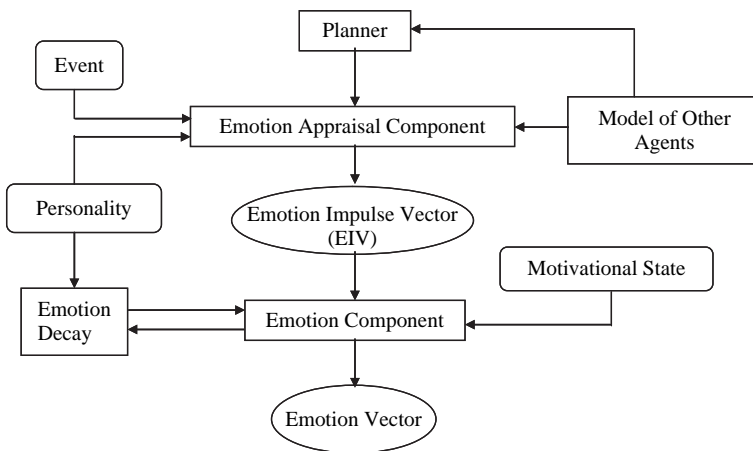
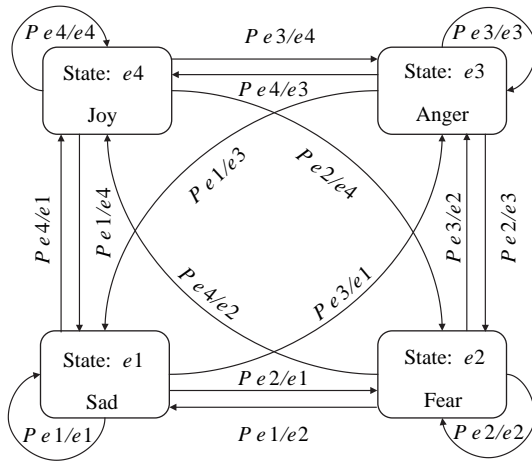


Fig. 6 An overview of ParleE System (adopted from [47])

Fig. 7 Markovian emotion model



to influence by itself or from outside stimuli and also particular character can be developed in a robot by carefully tuning the parameters of the emotion core.

We have also developed an emotion model to be implemented for MRT system through the adaptation of some features of emotion model from [48, 49]. This model is used to develop some intelligent behavior of job distributed mobile robots in a simulated environment (for more details see [50]) and also used to develop an affection-based task allocation method for cooperative multi-robot system (see [51]). The emotion generation system is shown in Fig. 8. Some input stimuli like workload (w), barrier-level (bl), energy level (e), etc. give emotion changing impact on Markov model through updating of *emotion-inducing factors* such as α, β, γ and δ which are for joy, anger, fear and sad respectively. For example, the *joy inducing*

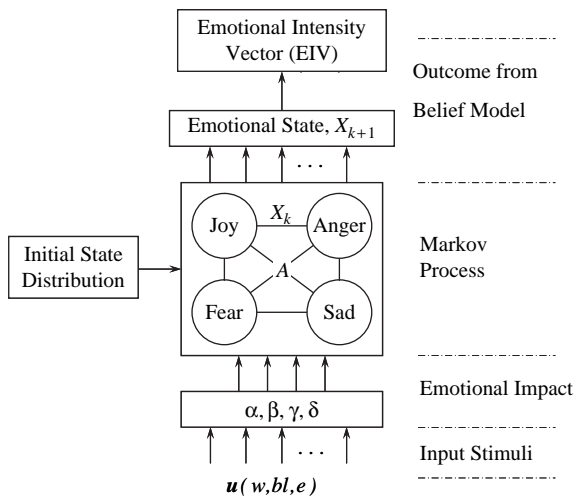


Fig. 8 The emotion generation system which can generate emotional state through the perception of input stimuli from environment

factor (α) will have a positive impact on the transition probability from any state to ‘happy’ state and on the contrary, it will oppose the transition from ‘happy’ state to other states. The elements of state transition matrix (A) are changed in each step with the updated value of *emotion-inducing factors*. In this model, we have considered four basic emotions for simplicity. But, a complex model can be developed including more emotions with multi-layered Markov model considering hierarchical stage of emotions.

The Markovian emotion model with four states can be expressed as follows:

$$X_{k+1} = AX_k \tag{1}$$

with emotional state points

$$\Omega = \{\text{Joy, Anger, Fear, Sad}\} \tag{2}$$

where X_k represents the current emotional state and A is the emotional state transition matrix (so called stochastic matrix) which can be expressed as follows:

$$A = \begin{bmatrix} P_{e4/e4} & P_{e4/e3} & P_{e4/e2} & P_{e4/e1} \\ P_{e3/e4} & P_{e3/e3} & P_{e3/e2} & P_{e3/e1} \\ P_{e2/e4} & P_{e2/e3} & P_{e2/e2} & P_{e2/e1} \\ P_{e1/e4} & P_{e1/e3} & P_{e1/e2} & P_{e1/e1} \end{bmatrix} \tag{3}$$

We found the model works well with adjustability and simplicity in structure. We have applied the model in a room cleaning task for simulation as shown in Fig. 9. Figure 10 shows the emotional changes of one robot with the progress of work where, emotional intensity is derived through a belief model from the probability of each emotion. For the task allocation problem [51], we have used three robots (A, B and C) in simulated environment of which the emotional states are shown in Figs. 11, 12 and 13. We found the emotional states are very logical considering the internal and external situations that prevail during the task performance. In [49], Kühnlenz and Martin also found satisfactory results in simulation with emotion core

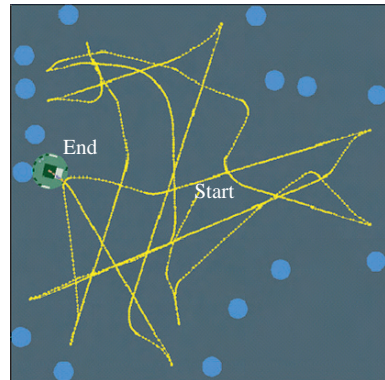


Fig. 9 The room cleaning task for a simulated robot as in [50]

Fig. 10 Emotion intensity changes with time step as in [50]

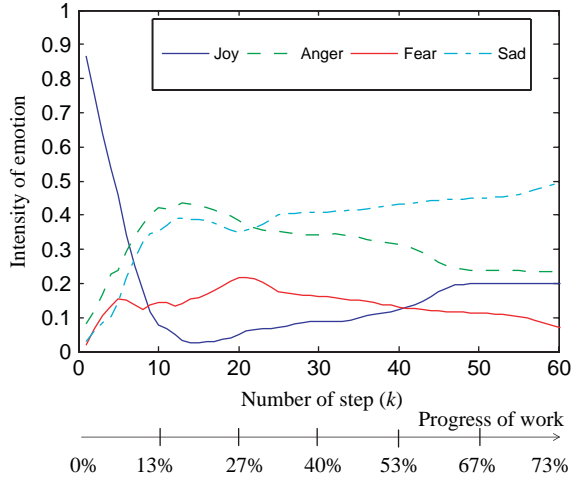


Fig. 11 Dominating emotional state of Robot A [51]

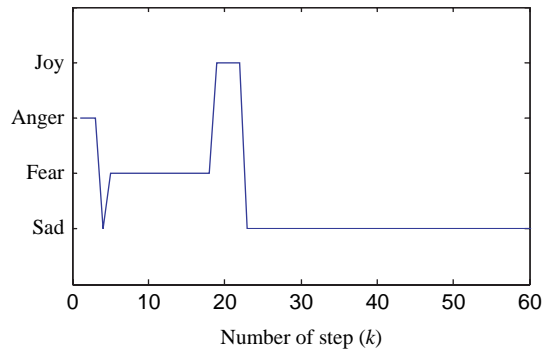
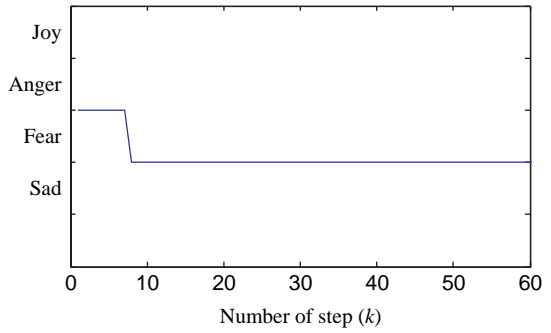
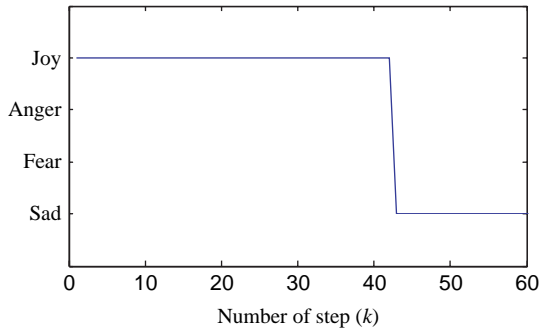


Fig. 12 Dominating emotional state of Robot B [51]

Fig. 13 Dominating emotional state of Robot C [51]



based on Markov model. In Fig. 14, we can see the emotional change from the joy to the fear state after receiving the perception of a frightening stimulus and it is remaining until the stimulus is weak enough. One of the drawbacks of this model is that the solution time increases exponentially with the size of emotional state. To reduce the time, C. Arun has proposed a hardware approach for complex emotional states [48].

In this section, we have discussed in brief some important models of emotions for AI agents with their advantages and limitations. There are still many open issues about universal vs. agent-specific emotions, pure vs. mixed emotion, emotion vs. moods, etc. There are no universally accepted concepts for these types of comparative studies.

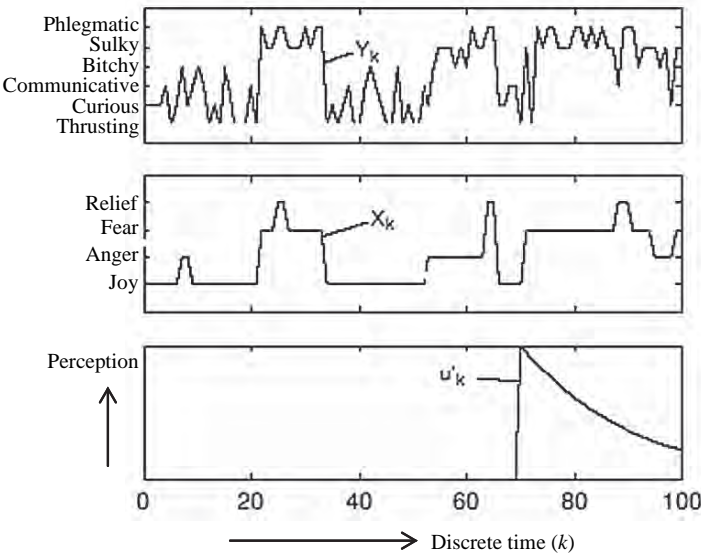


Fig. 14 The Observations Y_k and emotional state X_k of HMM core over discrete time k under perception of frightening stimulus u' as in [49]

7 Affective Computing, Sensing, Expressing and Synthesis

Affective computing is in infancy state, but results have already been proved that it is no longer an oxymoron to think about its application for robotics. A robotic system with affective abilities and logical reasoning abilities will bring balance and reason to their logical skills. When a system faces many problems, where possible solutions can not be evaluated in the available time, affective decision making gives a good solution [1].

There are so many technical issues to develop MRT system with the ability to recognize, express and have emotions. There should be an easier way for representing input and internal signals, understanding the patterns of signals, synthesizing the expressions, generating states and analyzing situation and so on. R. W. Picard in her *Affective computing* book [1], has given some guidelines for all of these points to develop affective systems.

In the simulation, we have used the model of a miniature robot like Khepera which has eight light sensors. These sensors are also functioning as proximity sensors and each sensor is having a field-of-view of about 120°. Each sensor has been simulated with a neural network model by which it can detect obstacles or light source. We have also used a model of the linear vision extension module that used in Khepera.

The input stimuli from the environment are the sensor’s output to the emotion generation system. There is a fuzzy rule based perception system to map these inputs to the *emotion inducing factors* and with these updated values the new emotional state is generated. The basic fuzzy rule based perception system is shown in Fig. 15 and some of the fuzzy rule surfaces are shown in Figs. 16, 17 and 18.

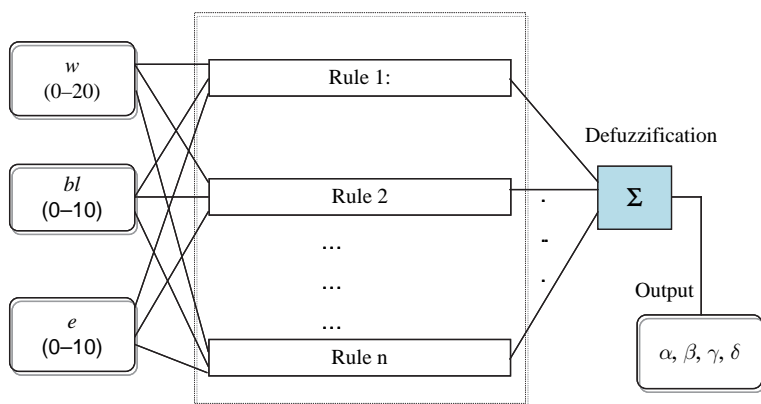


Fig. 15 Fuzzy rule based perception system

Fig. 16 Surface view of joy factor (α) at $bl = 10$

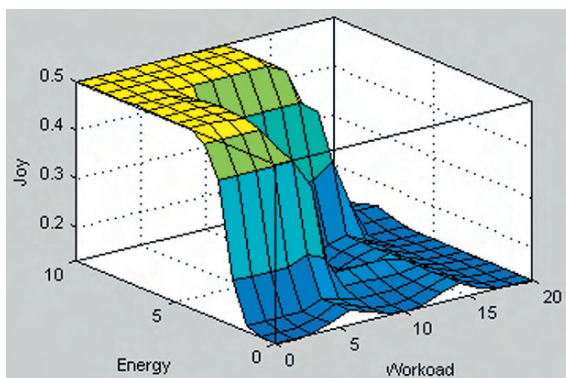


Fig. 17 Surface view of anger factor (β) at $bl = 10$

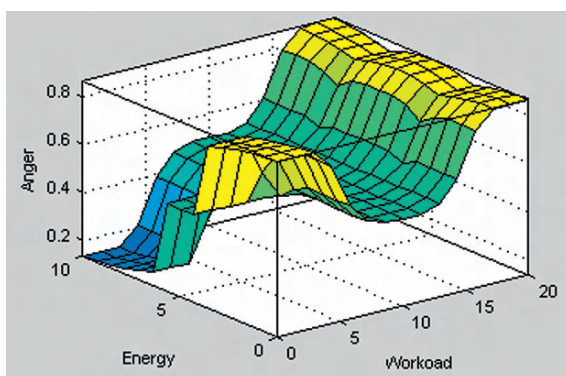
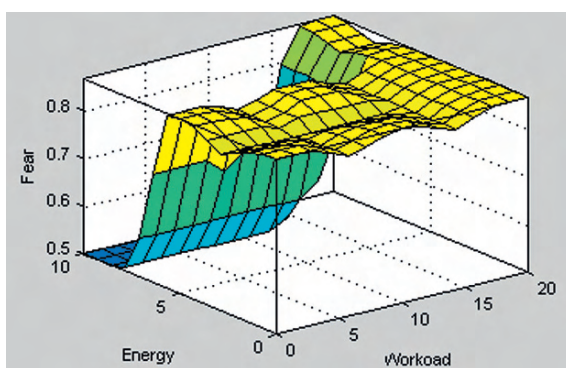


Fig. 18 Surface view of fear factor (γ) at $bl = 10$



8 Conclusion

In this chapter, we have discussed some of the important features of emotions and their modeling from the perspective of MRT/MAS. The research works on this field already proved that, it needs ideas and interests from multi disciplines related to affective sciences like as neuroscience, philosophy, psychology, etc. A common effort from specialists of these fields will play a great role to develop affection based robotics. With the emotionally capabilities, robots will be more life-like and acceptance of these robots will be increased to general people. A common platform can be developed where human and robots will work together in a team with such a feeling that robots are colleagues when having lively features with emotional capabilities. But before that, we need more research to remove the complexities and limitations that face with emotion modeling, interacting, expressing, sensing, application, etc.

References

1. Picard, R. W., "Affective computing", Cambridge: MIT Press, 2000, ISBN 0-262-16170-2.
2. Nair, R., Tambe, M., Marsella, S., "The role of emotions in multi-agent teamwork: A preliminary investigation", in Fellous, J-M., Arbib, M. (Eds), Who needs emotions: the brain meets the robots. Oxford: Oxford University Press, 2005.
3. Scheutz, M., "Useful roles of emotions in artificial agents: A case study from artificial life", Proceedings of AAAI conference, pp. 42–48, July 2004.
4. Baghaei, K. R., Agah, A., "Task allocation methodologies for multi-robot systems", Technical Report ITTC-FY2003-TR-20272-01, Information and Telecommunication Technology Center, University of Kansas, 2002.
5. Gage, A., "Multi-robot task allocation using affect", PhD Thesis, 2004, University of South Florida.
6. Gerkey, B. P., Mataric, M. J., "Pusher-watcher: an approach to fault-tolerant tightly-coupled robot coordination", Proceedings of International Conference on Robotics and Automation. 2002, IEEE Catalog Number 02CH37292, ISBN 0-7803-7272-7, 464–469.
7. Gerkey, B. P., Mataric, M. J., "Sold!: Auction methods for multirobot coordination", IEEE Transactions on Robotics and Automation 2002, 18, 758–768.
8. Gerkey, B. P., Mataric, M. J., "Multirobot task allocation: analyzing the complexity and optimality of key architectures", Proceedings of International Conference on Robotics and Automation. 2003, IEEE Catalog Number 03CH37422, ISBN 0-7803-7736-2, 3862–3868.
9. Parker, L. E., "Alliance: architecture for fault tolerant and multirobot cooperation", IEEE Transactions on Robotics and Automation 1998, 14, 220–240.
10. Parker, L. E., "Alliance: An architecture for fault tolerant, cooperative control of heterogeneous mobile robots", Proceedings of International Conference on Intelligent Robots and Systems (IROS'04). 1994, IEEE Catalog Number 94CH3447-0, ISBN 0-7803-1933-8, 776–783.
11. Chaimowicz, L., Campos, M. F. M., Kumar, V., "Dynamic role assignment for cooperative robots", Proceedings of International Conference on Robotics and Automation. 2002, IEEE Catalog Number 02CH37292, ISBN 0-7803-7272-7, 293–298.
12. Wergler, B. B., Mataric, M. J., "Broadcast of local eligibility for multi-target observation", Proceedings of DARS 4, pp. 347–356, Oct. 2000.

13. Werger, B. B., Mataric, M. J., "From insect to internet: Situated control for networked robot teams", *Annals of Mathematics and Artificial Intelligence* 2001, 31, 173–198.
14. Ostergaard, E. H., Mataric, M. J., Sukhatme, G. S., "Distributed multi-robot task allocation for emergency handling", *Proceedings of International Conference on Robotics and Systems*. 2001, IEEE Catalog Number 01CH37180, ISBN 0-7803-6612-3, 821–826.
15. Alami, R., Fleury, S., Herrb, M., Ingrand, F., Robert, F., "Multi-robot cooperation in the MARTHA project", *IEEE Robotics and Automation Magazine* 1998, 5, 36–47.
16. Simmons, R., Singh, S., Hershberger, D., Ramos, J., Smith, T., "First results in the coordination of heterogeneous robots for large-scale assembly", *Lecture Notes in Control and Information Sciences* 2000, 271, 323–332.
17. Farinelli, A., Iocchi, L., Nardi, D., "Multi-robot system: A Classification focused on coordination", *IEEE Transactions on System Man and Cybernetics* 2004, part B, 2015–2028.
18. Stone, P., "Layered learning in multiagent system: A winning approach to robotic soccer", MIT Press, 2000, ISBN 0-262-19438.
19. Dudek, G., Jenkin, M., Milius, E., Wilkes, D., "A taxonomy for multi-agent robotics", *Autonomous Robots* 1996, 3(4), 375–397.
20. Dudek, G., Jenkin, M., Milius, E., "A taxonomy of multirobot systems", in Tucker, B., Lynne, E. P. (Eds), *Robot Teams: From Diversity to Polymorphism*, Natick, MA, Canada: A K Peters, Ltd., 2002.
21. Minsky, M., "The society of mind", New York: Simon & Schuster, Inc., 1988, ISBN 0-671-60740-5.
22. Kenyon, S. H., "The need for emotional architectures in practical robots", [online] http://stardec.ascc.neu.edu/~kenyon/personal/papers/emotional_arch_robot_v1.pdf.
23. Levenson, R. W., "Emotion and the automatic nervous system: A prospectus for research on autonomic specificity", in Wagner, H. L. (Ed.), *Social Psychophysiology and Emotion*, Chichester: John Wiley & Son, 1988.
24. Oatley, K., Johnson-Laird, P. N., "Towards a cognitive theory of emotions", *Cognitive and Emotion* 1987, 1(1), 29–50.
25. Murphy, R. R., Lisetti, C. L., Tardif, R., Irish, L., Gage, A., "Emotion-based control of cooperating heterogeneous mobile robots", *IEEE Transactions on Robotics and Automation* 2002, 18(5), 744–757.
26. de Freitas, J. S., Gudwin, R. R., Queiroz, J., "Emotion in artificial intelligence and artificial life research: Facing Problems", [online] <http://www.dca.fee.unicamp.br/projects/artcog/files/freitas-iva05-extended.pdf>.
27. Brave, S., Nass, C., "Emotion in human-computer interaction", in Sears, A., Jacko, J. A. (Eds), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, NJ, USA: Lawrence Erlbaum Associates, Inc., 2002.
28. Breazeal, C., "Emotion and sociable humanoid robots", *International Journal of Human-Computer Studies* 2003, 59, 119–155.
29. Custódio, L., Ventura, R., Pinto-Ferreira, C., "Artificial emotions and emotion-based control systems", *Proceedings of 7th International Conference on Emerging Technologies and Factory Automation*. 1999, IEEE Catalog Number 99TH8467, ISBN 0-7803-5670-3, 1415–1420.
30. McCauley, T. L., Franklin, S., "An architecture for emotion", *Proceedings of AAAI Fall Symposium*, pp. 122–127, 1998.
31. Nehaniv, C., "The first, second and third person emotions: Grounding adaptation in a biological and social world", 5th International Conference of the society for adaptive behavior (SAB), August 1998. <http://www.ofai.at/~paolo.petta/conf/sab98/final/nehaviv.ps.gz>.
32. Gadanho, S., Hallam, J., "Emotion-triggered learning in autonomous robot control", *Cybernetics and Systems* 2001, 32, 531–559.
33. Plutchik, R., "A general psychoevolutionary theory of emotion", in Plutchik, R. and Kellerman, H. (Eds), *Emotion: Theory, Research and Experience*, Vol. 1, New York: Academic Press, 1980.

34. Ortony, A., Clore, G. L., Collins, A., "The Cognitive Structure of Emotions", Cambridge, MA: Cambridge University Press, 1988, ISBN 0-521-38664-0.
35. Michaud, F., Robichaud, E., Audet, J., "Using motives and artificial emotions for prolonged activity of a group of autonomous robots", Emotional and Intelligent II: The Tangled Knot of Social Cognition-AAAI Fall Symposium, Technical Report FS-01-02, pp. 85–90, Nov. 2001.
36. Mook, D. G., "Motivation: The organization of action", New York: W. W. Norton & Co.
37. Cañamero, L., "Emotion understanding from the perspective of autonomous robots research", Neural Networks 2005, 18, 445–455.
38. Schneider-Fontan, M., Mataric, M., "Territorial multi-robot task division", IEEE Transaction on Robotics and Automation 1998, 14, 815–822.
39. Balch, T. R., Arkin, R. C., "Communication in reactive multi-agent robotic systems", Autonomous Robots 1994, 1, 1–25.
40. MacLennan, B., "Synthetic ethology: An approach to the study of communication", in Langton, C. G., Taylor, C., Farmer, J. D., Rasmussen, S. (Ed.), Artificial Life II: The Second Workshop on Synthesis and Simulation of Living Systems (pp. 631–658). Redwood City, CA: Addison-Wesley, 1991.
41. Maes, P., "Modeling adaptive autonomous agents", in Langton, C. G. (Ed.), Artificial Life: An overview (pp. 176–181). Cambridge, MA: The MIT Press, 1995.
42. Velásquez, J. D., "Modeling emotions and other motivations in synthetic agents", Proceedings of AAAI Conference, pp. 10–15, July 1997.
43. Elliott, C., "The Affective Reasoner: A process model of emotions in a multi-agent system", PhD thesis, Institute for the Learning Sciences, Evanston, IL: Northwestern University, 1992.
44. El-Nasr, M. S., Ioerger, T., Yen, J., "FLAME: Fuzzy logic adaptive model of emotions", Autonomous Agents and Multi-Agent Systems 2000, 3(3), 219–257.
45. Reilly, W. S., Bates, J., "Building emotional agents", Technical Rep. CMU-CS-92-143, Pittsburgh, PA: Carnegie Mellon University, 1992.
46. Gratch, J., "Émile: Marshalling passions in training and education", Proceedings of the 4th International Conference on Autonomous Agents, pp. 325–332, June 2000.
47. Bui, T. D., Heylen, D., Poel, M., Nijholt, A., "ParleE: An adaptive plan based event appraisal model of emotions", Lecture Notes on Artificial Intelligence 2002, 2479, 129–143.
48. Arun, C., "A computational architecture to model human emotions", Proceedings of International Conference on Intelligent Information System, pp. 86–89, Dec. 1997.
49. Kühnlenz, K., Buss, M., "Towards an emotion core based on a hidden Markov model", Proceedings of the 13th International Workshop on robot and human interactive communication (ROMAN'04), pp. 119–124, Sept. 2004.
50. Banik, S. C., Watanabe, K., Izumi, K., "Intelligent behavior generation of job distributed mobile robots through emotional interaction", Proceedings of 13th International Conference on Advanced Robotics, pp. 1215–1219, Aug. 2007.
51. Banik, S. C., Watanabe, K., Izumi, K., "Task allocation with a cooperative plan for an emotionally intelligent system of multi-robots", SICE Annual Conference, pp. 1004–1010, Sept. 2007.

Part IX
Displacement Sensor

Displacement Sensor Using Magnetostrictive Wire and Decrease of its Hysteresis Error

Hiroyuki Wakiwaka

Abstract A displacement sensor using magnetostrictive wire is a sensor which estimates the displacement from propagation time of an elastic wave that is caused and detected by using the magnetostrictive effect and the inverse-magnetostrictive effect. This sensor can be used for measurement up to 60 meters in simple structure, therefore it is appropriate for industry applications. This chapter describes the followings; In case of unipolar pulse current used until now, 1.3 millimeters hysteresis error is produced. Then, various methods for reducing the hysteresis error were proposed. Here, a method for reducing the error is introduced by the form of the pulse current. The bipolar pulse current is used instead of the unipolar pulse current. As the result, it is able to decrease the hysteresis error to 0.3 millimeters or less.

1 Introduction

A displacement sensor using a magnetostrictive wire is an absolute type displacement sensor using the magnetostrictive effect and the inverse-magnetostrictive effect. With its advantages such as simple structure, high precision, wide measurement range, good reliability and so on, this sensor has shown its practical value. It can be used up to 60 meters [1] in simple structure, therefore it is appropriate for industry applications. The magnetostrictive wire used in this sensor is made of Ni-Span-C which is a nickel based alloy. Ni-Span-C is a ferromagnetic material. Therefore, the displacement error occurs with magnetic hysteresis of the magnetostrictive wire. Various methods for compensating for hysteresis error in the magnetostrictive linear position sensor were already presented [2]. However, we examine a simpler compensation method for the hysteresis error for practical uses.

Hiroyuki Wakiwaka

*Shinshu University, 4-17-1, Wakasato, Nagano 380-8553, JAPAN,
e-mail: wakiwak@shinshu-u.ac.jp

In this chapter, the following contents are described:

- (1) Operation principle of the magnetostrictive wire type displacement sensor.
- (2) The cause of hysteresis error generation.
- (3) The hysteresis error comparison by a bipolar pulse current and a unipolar pulse current.

2 Operation Principle of Magnetostrictive Wire Type Displacement Sensor

Figure 1 shows the basic structure and operation principle of the sensor [3]. An elastic wave is generated near the magnet beside the wire, and it propagates to the detecting element. The absolute position sensor can be realized, because the position of the magnet can be estimated from propagation time of the elastic wave.

In Fig. 1, when pulse current I_p is flowing through the magnetostrictive wire, magnetic field Φ_i will be induced around the wire. Meanwhile, the magnet near the wire produces magnetic field Φ_m that is parallel with the wire. The two fields Φ_i and Φ_m combine to make in twisting field Φ .

Because of the magnetostrictive effect, an instantaneous distortion of the magnetostrictive wire will appear in the resultant field Φ . According to the Wiedenman effect, a twisting elastic wave will be generated in the wire and spreads to the two ends at a certain velocity of v . When the wave arrives at the detecting coil, because the change of the mechanical stress causes a change of magnetic field in the magnetostrictive wire, the magnetic flux density B changes at the same time. According to

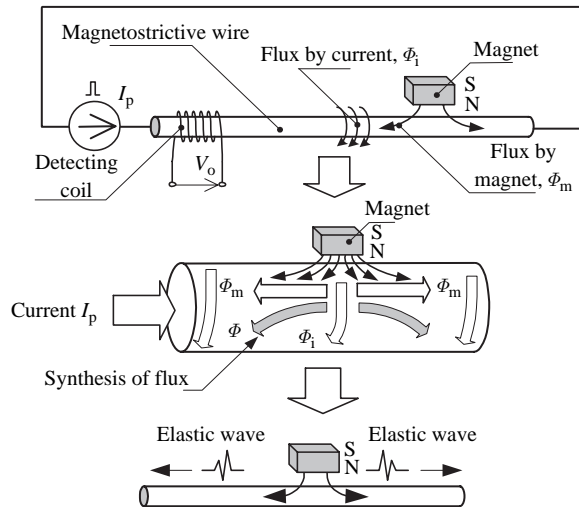


Fig. 1 Basic structure and operation principle of the sensor

Faraday’s Law, induced voltage will be produced at the detecting coil. The formula is given by (1).

$$V_o = -NS \frac{dB}{dt} \tag{1}$$

where, V_o : detected voltage[V]; N : number of detecting coil turns; B : magnetic flux density[T]; S : cross section area of the wire[m²].

3 Measurement of the Displacement Error

3.1 Measuring Method

The basic configuration for measuring the displacement error is shown in Fig. 2. The position of the permanent magnet is made to move in 10 millimeters interval. It is moved 50 millimeters in both directions, both toward and away from the detection coil, and the displacement error is measured.

Propagation distance of an elastic wave is shown by the following equation;

$$x_d = v \cdot t_d \tag{2}$$

where, v : velocity of the elastic wave [m/s], t_d : propagation time of the elastic wave [s], x : displacement of the permanent magnet [m].

The displacement error ε is obtained from an indicated value x_s of a linear encoder in equation (3);

$$\varepsilon = x_d - x_s \tag{3}$$

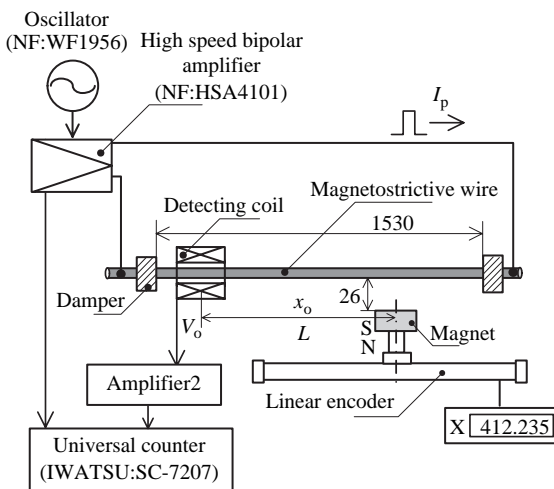


Fig. 2 Measurement of the displacement error

3.2 Measurement Result

Experimental conditions for the displacement sensor using magnetostrictive wire are listed in Table 1. The unipolar and bipolar current waveforms are shown in Fig. 3. The displacement error curve is shown in Fig. 4. When the unipolar pulse current is flowing though the magnetostrictive wire, a large displacement error width of 1.3 millimeters is observed.

When the bipolar pulse current is flowing though the magnetostrictive wire, the width of the displacement error became about 0.3 millimeters or less.

Table 1 Experimental conditions for the displacement sensor using magnetostrictive wire

Item		Symbol [Unit]	Material, Value
Magnetostrictive wire	Material		Ni-Span-C
	Length	L [m]	1.53
	Diameter of wire	d [mm]	0.5
	Resistance	R [Ω]	9.5
Pulse current	Amplitude	I_p [A]	0.7
	Period	T [ms]	5
	Width of pulse	T_w [μ s]	22
Magnet	Material		Nd-Fe-B
	Size	[mm]	$20 \times 15 \times 10$
Amp.	Amplification factor	A_v dB	36.1
Detecting coil	Number of turns	N [turn]	1500
	Length	l [mm]	5
	Wire diameter	d [mm]	0.06

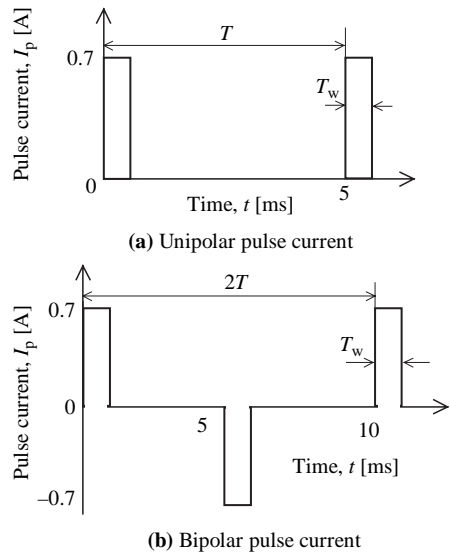
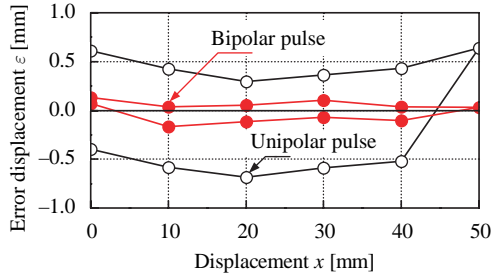


Fig. 3 Pulse current wave form

Fig. 4 Displacement error



Then, the displacement error became about 1/4 or less than when the unipolar pulsed current is flowing through.

4 Consideration on the Reduction of Displacement Error

A hysteresis loop of magnetization is generated in the magnetostrictive wire. When a unipolar pulse current is flowing, and a magnet reciprocates, the minor loop shift model is shown in Fig. 5. The magnetization characteristic of the magnetostrictive wire is changed, and the generation point of elastic wave is different. As a result the displacement error occurs.

For this reason, by the change of the direction of the pulsed current, when the bipolar pulse current is flowed, the action of demagnetization worked in the magnetostrictive wire. Therefore it seemed to hold the change of the elastic wave generation position constant. The minor loop shifts, when only the positive pulse

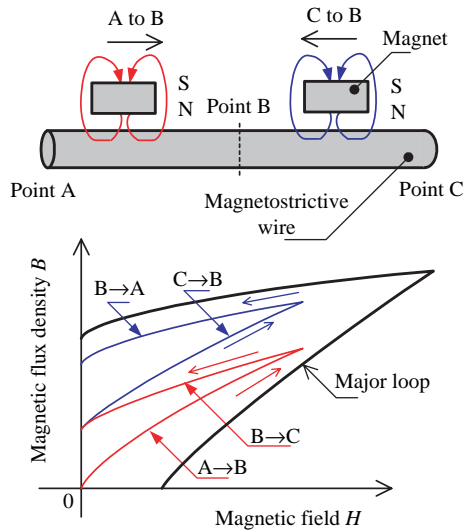


Fig. 5 Minor loop shift model.

was applied. The negative pulse functions as a reset pulse of the minor loop. As a result, the hysteresis error is reduced.

5 Conclusions

The conclusions are described as follows;

- (1) The basic structure of the displacement sensor using the magnetostrictive wire was shown. The structure of this displacement sensor is simple. It is possible to measure the displacement more accurately than previously possible.
- (2) By applying the bipolar pulse current, the magnetostrictive wire is demagnetized at every pulse current. Then, it reduces the measurement displacement error from 1.3 millimeters to 0.3 millimeters or less. This means that the error is reduced 1/4 or less.

References

1. H. Wakiwaka, M. Shimada, S. Hattori, J. Nishiyama, K. Murata, K. Ito, "Elastic wave signal processing on a long displacement sensor using magnetostrictive wire", The 1st International Symposium on Linear Drives in Industry Applications 1995, FA-20 p. 243 (1995).
2. F. Seco, J.M. Martin, J.L. Pons, A.R. Jimenez, "Hysteresis compensation in a magnetostrictive linear position sensor", *Sensor and Actuator A* 110, pp. 247–253 (2004).
3. X. Chang, H. Wakiwaka, H. Yamada, "Comparison of two detection methods using a long-scale displacement sensor with magnetostrictive wire", *Journal of the Magnetics Society of Japan*, Vol. 22, No. 4–2 pp. 681–684 (1998).

Part X
THZ Sensor

Submillimeter-Wave Coherent and Incoherent Sensors for Space Applications

Goutam Chattopadhyay

Abstract Most of the radiation in the Universe is emitted at wavelengths longer than 10 microns (30 THz), and this peaks at about 100 microns (3 THz), if we exclude contributions from the cosmic microwave background (CMB). Radiation in these wavelengths highlights warm phenomena, processes of change such as star formation, formation of planetary systems, and galaxy evolution; atmospheric constituents and dynamics of the planets and comets and tracers for global monitoring and the ultimate health of the earth. Sensors at far-infrared and submillimeter wavelengths provide unprecedented sensitivity for astrophysical, planetary, earth observing, and ground-based imaging instruments. Very often, for spaced based platforms where the instruments are not limited by atmospheric losses and absorption, the overall instrument sensitivity is dictated by the sensitivity of the sensors themselves. Moreover, some of the cryogenic sensors at submillimeter wavelengths provide almost quantum-limited sensitivity. Frequency sources at submillimeter wavelengths with adequate output power for transmitters and local oscillators are not easily available, and pose the greatest challenge for advancement of this field. This article provides an overview of the state-of-the-art of submillimeter-wave sensors for a variety of space-borne applications and their performance and capabilities.

1 Introduction

Submillimeter wavelength is loosely defined as $1\text{ mm} > \lambda > 100\mu\text{m}$ (100 GHz to 3 THz frequency range). Figure 1 shows a schematic representation of the electromagnetic spectrum from radio to gamma rays highlighting the submillimeter wave region which is sandwiched between the microwave and infrared. Sensors at these frequencies are used for a variety of applications, from unravelling the mystery of star formation in far-away galaxies to detecting the presence life on another planet; from detecting presence of environment altering chemical species in the

Goutam Chattopadhyay

Jet Propulsion Laboratory, California Institute of Technology Pasadena, California, United States,
e-mail: goutam@jpl.nasa.gov

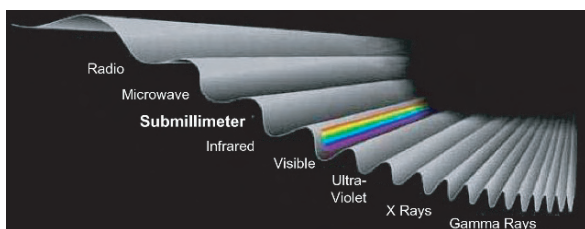


Fig. 1 Submillimeter wavelength is in between microwave and infrared, loosely defined as 100 GHz to 3 THz

Earth's atmosphere and protect the health and wellbeing of our planet to detecting contrabands at stand-off distances to protect the global citizenry from evil-doers; to answering the most important question in every human being's mind – where did we all come from.

Over this frequency range, the Earth's atmosphere is opaque due to the presence of a very large number of atomic and molecular spectral lines. On the other hand, it is precisely this wealth of spectral lines that makes this region of the electromagnetic spectrum so valuable for the astrophysics, planetary, and Earth observing studies. Therefore, in spite of submillimeter-wave technologies getting a lot of attention in recent years, technology development at these frequencies are driven largely by applications in astronomy, Earth, and planetary sciences. Very high atmospheric opacity from pressure broadened vibrational and rotational absorption signatures of water and oxygen prevents propagation of submillimeter-wave signals to large distances under standard temperature and pressure, as shown in the representative plot in Fig. 2. Even observing relatively strong emissions at submillimeter wavelengths from nearby star forming regions going above most of the atmosphere, either to high and dry mountain top observatories, stratospheric aircraft and balloons, or orbital platforms. However, once away from the atmospheric absorptions, the wealth of information about our own atmospheric processes, the constituents and conditions in the atmospheres of solar system bodies, and the vast amount of cosmic chemistry that occurs both around and within stellar systems, all provide a strong motivating force for the development and deployment of submillimeter-wave sensors from space-borne platforms [1].

Radiation at these wavelengths highlights warm phenomena, processes of change such as star formation, formation of planetary systems, and galaxy evolution; atmospheric constituents and dynamics of the planets and comets and tracers for global monitoring and the ultimate health of the earth. Then there is the cosmic microwave background (CMB) – the cooled radiation (2.7 Kelvin) that permeated the Universe for 15 billion years – which is largely confined to wavelengths between one and five millimeters with peak intensity at two millimetres, is the earliest electromagnetic relic of our Universe [2]. Determining the minute anisotropy in its temperature distributing and its polarization will lead to the knowledge of inflationary gravitational waves and the nature of the Universe at the very first fraction of a second after the Big Bang.

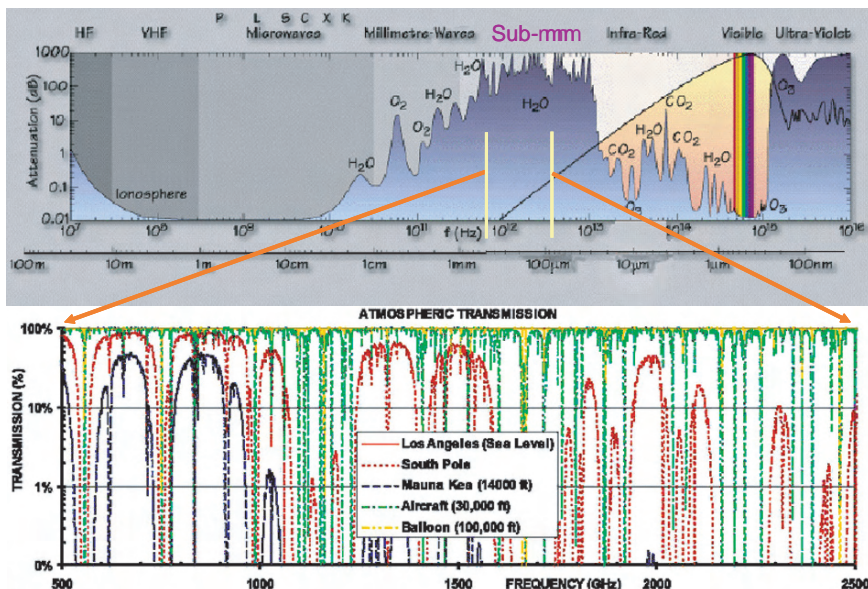


Fig. 2 Atmospheric attenuation (*top*) and transmission (*bottom*) as a function of frequency at different locations and altitudes. The bottom figure shows the blow-up of the submillimeter region

There is a lot of interest to know more about the planets in our solar system, as some of the planets and their moons have characteristics similar to our Earth. Instruments in the submillimeter wavelengths have the capability to unearth a lot of information by studying their atmospheres, surfaces, and subsurface water and ice contents [3, 4]. An immediate objective of the many of the planetary programs is the discovery of past or present habitable environments on Mars, Jupiter, Saturn, and their associated moons that could have supported microbial life and provided conditions favorable for the capture and preservation of biosignatures. In-situ and orbital instruments provide an important platform for the development and testing of the emerging and established technologies that will be used to assess past and present habitability and explore for biosignatures during future missions. Microscopic imaging, combined with spectroscopic methods at submillimeter wavelengths, provides a powerful capability for the identification of both minerals and organic matter in rocks, soils and ices on Mars, Jupiter and its moon Europa, and Saturn and its moon Titan. This approach provides observations at a spatial scale relevant to microbial life and is required for realistic assessments of habitability.

In spite of being the most fascinating part of the electromagnetic spectrum, submillimeter-wave frequency range remains one of the least explored bands. The reason is believed to be the non availability of commercially available sensor components, sub-systems, and instruments. However, with the new emerging applications such as imaging from space platforms [5, 6, 7], stand-off contraband detections and reconnaissance [8, 9], medical imaging [10], and even in the art world – for painting analysis [11]; the far-Infrared and submillimeter-wave band is increasingly

playing an important role in pushing the technology frontiers in this electromagnetic band. In this article we review the current status of submillimeter-wave sensors, instruments, and their applications.

2 Applications

Although the development of submillimeter-wave sensors and sources has primarily been driven by the astronomers to have instruments and detectors with almost quantum-limited sensitivities [12], in recent years planetary, Earth observing, medical imaging, and homeland security applications are playing a major role as well. In astronomy, the challenge lies in unearthing signals from the distant stars and galaxies embedded in the noise of the interstellar medium and the noise generated by the sensor themselves. They operate in a regime where signal to noise is less than one. That led to the development of the highly sensitive sensors and data/image processing techniques which found their way to other fields [13]. In the following sections, we list a couple of important areas which has been the driving force for the development of far-infrared and submillimeter-wave sensors and sources.

2.1 Astronomy and Astrophysics

Measuring primordial abundance of a variety of atoms and molecules such as deuterium strongly constrains the physical conditions during the first few minutes after the Big Bang and represents the most sensitive probe of the baryonic density Ω_b in the universe. Constraining the baryonic density is vital to predicting the density of the universe, and submillimeter-wave observations of the star forming regions is believed to be the key in getting this answered. The scientific importance of velocity-resolved high-resolution spectroscopic observations at submillimeter wavelengths is widely recognized by the international astronomy and astrophysics community. This importance is also underscored by the key role of heterodyne spectrometers in the ESA cornerstone Herschel Space Observatory [14] as well as the ground-based Atacama Large Millimeter Array (ALMA) [15] and airborne Stratospheric Observatory for Infrared Astronomy (SOFIA) [16]. Star formation and key phases of galaxy evolution occur in region enshrouded by dust that obscures them at infrared and optical wavelengths, while the temperature range of the interstellar medium of ten to a few thousand Kelvin in these regions excites a wealth of submillimeter-wave spectral lines. Figure 3 shows a schematic representation of the spectrum of a typical star-forming region. With high-resolution spectroscopy, resolved line profiles reveal the dynamics of star formation, directly revealing details of turbulence, outflows, and core collapse. Observations of emission from ionized species such as N^+ at 1461 GHz allow one to effectively count ionizing ultraviolet photons from newborn stars still enshrouded in their stellar nurseries. These allow one to understand the

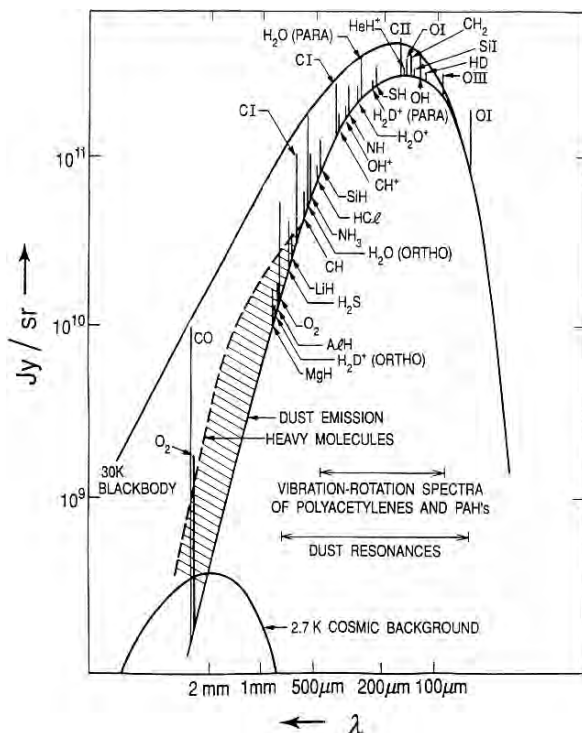


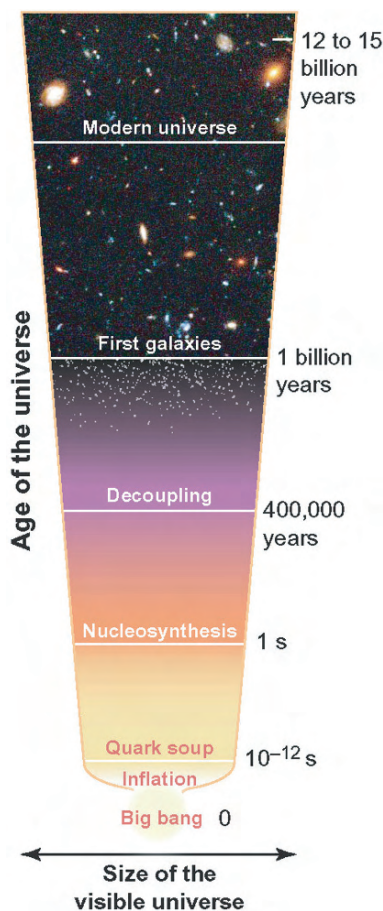
Fig. 3 Schematic representation of the spectrum of a star forming region (*bottom*). Spectral line emission is superimposed on the dust continuum [17, 18]

energy balance and dynamics of star forming regions and galaxy interactions, and are essential components to understanding the origins of both stars and galaxies.

Submillimeter-wave signals emitted from the stars and galaxies are mostly obscured from most Earth-based observations because of the atmospheric absorptions. This provides strong motivation for a number of existing or upcoming space astrophysics instruments. Most notable are the Submillimeter Wave Astronomy Satellite (SWAS) [19], launched in December 1998, and currently sending back signals. The future missions such as Herschel Space Observatory [14] and Exploratory Submillimeter Space Radio-Interferometric Telescope (ESPRIT) [20] will advance our knowledge about this universe by many folds.

Cosmology is another area of astrophysics which has been studied in detail at submillimeter wavelengths to answer key questions about the Big Bang and the life of the universe at its infancy. Figure 4 shows the pictorial depiction of our universe as we know it. Study of the CMB, which was emitted approximately 500,000 years after the big bang, when electrons and protons in the primordial plasma – the hot, dense soup of subatomic particles that filled the early universe – first combined to form hydrogen atoms [2], and its temperature anisotropy and polarization will answer many of the key features of the universe such as the presence of the inflationary gravitational waves.

Fig. 4 Pictorial representation of the evolution of the universe



2.2 Planetary Sciences

Planetary and small-body (asteroids, moons, and comets) observations have been a major space application for submillimeter-wave sensors. Orbital remote sensing or lander-based robotic in-situ observations of the atmospheres of Venus, Mars, Jupiter, and Saturn, as well as their moons such as Europa, Ganymede, and Titan have been either proposed or currently in orbit [21]. Submillimeter-wave sensors in the spectroscopic instruments on planetary missions allow a large number of chemical species in the atmospheres of Mars and Titan to be detected at concentrations below a part per billion, and their location to be precisely pinpointed in latitude, longitude, and in altitude. Specific species of interest include water, NO_2 , N_2O , NH_3 , SO_2 , H_2S , CH_4 , and HCN , among others. Moreover, the radiometer instruments at submillimeter wavelengths allow determining the nature and composition of cometary and planetary surfaces such as the Mars, Europa, and Titan by measuring the polarization-sensitive thermal emission from the dielectric surfaces.

In March 2004, MIRO (Microwave Instrument for Rosetta) was launched on the Rosetta spacecraft that will rendezvous with Comet 67 P/Churyumov-Gerasimenko [22]. MIRO will measure the near surface temperatures of the comet nucleus and possibly of an asteroid, thereby allowing scientists to estimate the thermal and electrical properties of these surfaces. In addition, the spectrometer portion of MIRO will provide measurements of water, carbon monoxide, ammonia, and methanol in the gaseous coma of the comet [23]. These measurements will provide insight as to how the comet nucleus material sublimates (changes from its frozen state, ice, to a gas) in time and distance from the sun. In addition the Doppler shifts of the spectral lines will characterize the velocities of gas outflow from the nucleus. Figure 5 shows the chemistry of the middle atmosphere, cloud layers, lower atmosphere, and surface on the planet Venus which can be detected by submillimeter-wave sensors, and its submillimeter-wave spectrum as a function of altitude and brightness temperature.

An ongoing challenge in the exploration of the outer solar system is how to determine the prospects for life and how life might have evolved on Earth. The larger moons of Jupiter and Saturn are among the most interesting bodies in the solar system because their composition includes a large amount of water ice. In the case of Europa there is strong evidence that there is at least some liquid water under the surface and the possibility of a biosphere, as illustrated in Fig. 6. A submillimeter-wave radar spectrometer, as shown in Fig. 7, will be ideal to explore the biotic and non-biotic materials of the Europa-surface. The products of high-resolution submillimeter-wave remote sensing, such as composition, temperature, pressure, and gas velocity (winds) offer the planetologist a wealth of information on a global scale for planetary surfaces and atmospheres such as the Titan. Saturn’s largest moon Titan has dense atmosphere, as shown in Fig. 8, whose characteristics are in many ways similar to our home planet. Fig. 8 (right) shows the schematic of a future mission being considered for Titan. Given the potential of submillimeter-wave technologies, it is not unreasonable to suppose that the first detection of planets

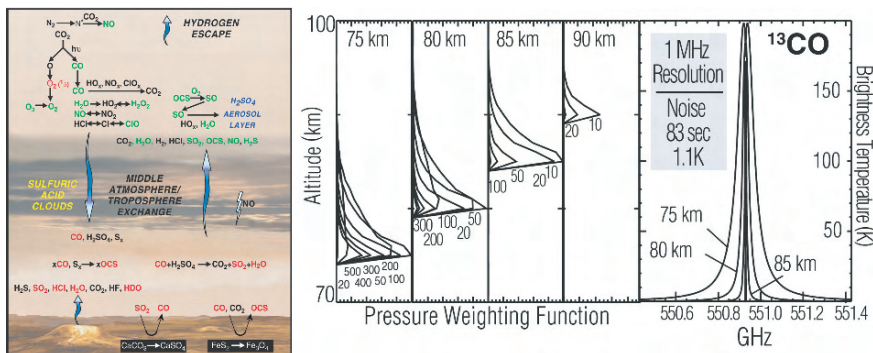


Fig. 5 The chemistry of the middle and lower atmosphere and surface on the planet Venus and its submillimeter-wave spectrum

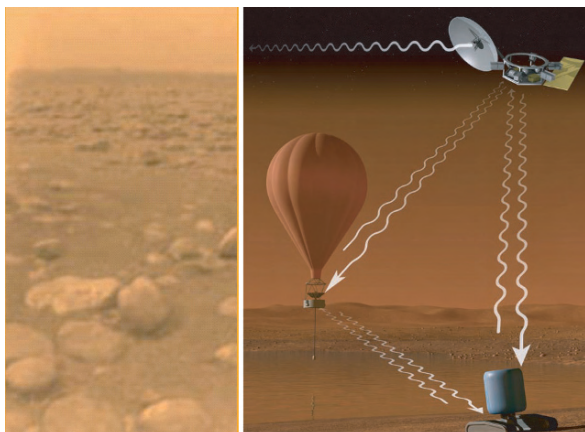


Fig. 8 Titan’s atmosphere (*left*), and schematic of a conceptual mission to Titan

are thermal emission lines from gases that appear in the Earth’s stratosphere and upper troposphere that serve as pointers to the abundances, distributions, and reaction rates of species involved in ozone destruction, global warming, total radiation balance, and pollution monitoring. Many key species either have thermal emission line peaks or their first rotational or vibrational line emissions in the submillimeter, especially between 300–2500 GHz. Again, these emission lines are best observed from platforms above the Earth’s atmosphere. Figure 9 shows Ozone distribution (and the so called Ozone hole) on Earth observed from an orbital platform.

It is now well established that human activity has begun to affect the health of our planet [25]. A prime example is anthropogenic effects on stratospheric chemistry that lead to global depletion of the protective ozone layer and the Antarctic ozone hole [26]. Tropospheric ozone and related trace gases have also been perturbed significantly and are likely to have modified the atmospheric oxidizing capacity and contributed to climate change [27, 28]. Remote sensing of Earth’s atmosphere at submillimeter wavelengths is an important method of obtaining global observations needed for atmospheric chemistry and climate [29]. Submillimeter-wave

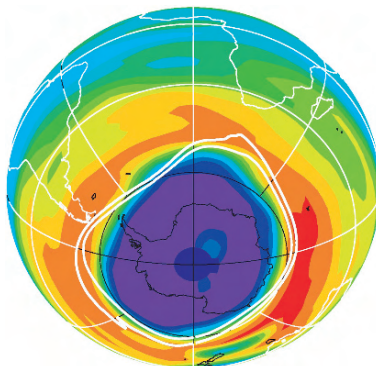


Fig. 9 Ozone distribution on Earth observed at submillimeter wavelengths

measurements are obtained from observations of atmospheric spectral line thermal emission, allowing daily global coverage from a satellite-based instrument. Additional important features include the ability to (a) make chemical measurements in the presence of dense volcanic aerosol, smoke, and ice cloud, and (b) measure signals from weak spectral lines in the presence of nearby very strong ones. These features are due to the relatively long wavelengths – compared to infrared, visible, and ultraviolet spectral regions; and the excellent spectral resolution available with heterodyne techniques at submillimeter-wave frequencies [30].

Microwave limb sounding is a proven remote-sensing technique that resolves the spectra of microwave thermal emission along a limb view of the earth's atmosphere with a cold space background. The temperature and composition of the atmosphere as a function of altitude is retrieved by analyzing the spectra returned from a vertical scan of the limb. These techniques have already been developed and applied to stratospheric chemistry measurements from space.

The Microwave Limb Sounder (MLS) instrument [31] on the NASA Upper Atmosphere Research Satellite (UARS), launched in 2004, was the first experiment to study the microwave limb from space, and was followed by the current EOS MLS instrument on the Aura spacecraft. High-resolution heterodyne receivers at 118, 190, 240, 640, and 2520 GHz were designed for this mission to take advantage of the information content available through high-resolution spectroscopic measurements of these gases at submillimeter-wave frequencies. Unlike the astrophysical sources, even modest diameter collecting surfaces are fully filled by the signal beam in atmospheric observations. Resolution requirements are set by the orbital path and speed or by the atmospheric processes themselves. In both limb sounding (scanning through the atmospheric limb) or nadir sounding (looking straight down through the atmosphere), precise spectral line-shape information is required to separate out the effects of pressure and Doppler broadening at each altitude along the emission path. Spectral resolution of better than one part in a million is typically needed for line widths that range from tens of kilohertz in the upper stratosphere to 10 MHz or more lower down. Figure 10 shows the capability and coverage of submillimeter-wave sensors for tropospheric chemistry measurements.

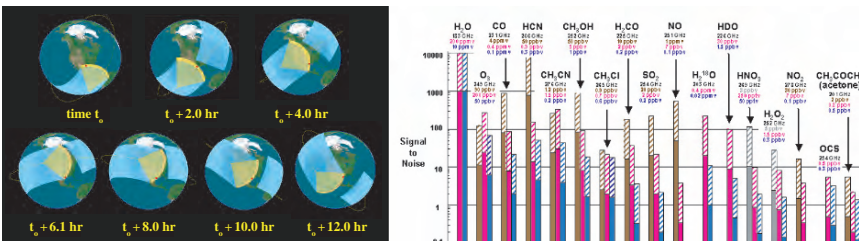


Fig. 10 Earth observing Microwave Limb Sounder (MLS) coverage of the troposphere in each pass (*left*), and mid-upper tropospheric chemistry measurements achievable by a future satellite-based sensor with sensitive ($T_{\text{sys}} = 100$ K) submillimeter-wave sensors (*right*). Many molecules have strong spectral emissions at frequencies between 180 and 280 GHz, detectable with short integration times

2.4 Other Applications

Recent progress in submillimeter-wave and terahertz (THz) technology, as well as the demand for new surveillance capabilities, has led to the development of prototype submillimeter imagers capable of detecting weapons concealed within clothing or packages [5, 6, 7, 8, 9, 32, 33, 34, 35, 36]. Imaging in the submillimeter wavelengths is attractive because wavelengths in the range $100 \mu\text{m} < \lambda < 1 \text{ mm}$ are short enough to provide high resolution with modest apertures and long enough to penetrate materials such as cloth or cardboard, as shown in Fig. 11. However, current approaches to submillimeter-wave imaging do not yet meet all of the real-world and often conflicting requirements of standoff range, portability, high speed, penetrability, and target identification amongst clutter.

For example, while active submillimeter-wave imaging systems using high-power coherent illumination and ultra-low-noise heterodyne detection show great promise, they often face operational drawbacks such as requiring cryogenic detectors or bulky laser sources. A more fundamental difficulty with coherent active imaging is that by relying on a single frequency, target recognition is reliant on an object's contrast and brightness which, in turn, are highly sensitive to incidence angle of radiation, clutter signal from the foreground or background, and interference and speckle effects.

Imaging radar at these frequencies is believed to solve many of the problems listed above. A room temperature active submillimeter imager can be used in the swept-frequency frequency modulated continuous wave (FMCW) radar mode to map a target in three dimensions. Figure 12 shows images using such a radar which can distinguish targets with centimeter-scale resolution in both range and cross-range. The images clearly indicate that radar capability may emerge as a key component of active submillimeter-wave imagers.

Submillimeter-wave imaging techniques are also showing a lot of promise in biological systems, from imaging of tooth cavity to cancerous cells. A detailed review of this subject is available in [10]. Another application of submillimeter-wave sensors has been in the areas of plasma fusion diagnostics and gas spectroscopy. An excellent review of submillimeter-wave techniques in the fusion field can be found in [37].



Fig. 11 Application of submillimeter-wave imaging techniques for reconnaissance from unmanned aerial vehicles (*left*), suicide bomber detection (*center*), and contraband detection at standoff distance (*right*)

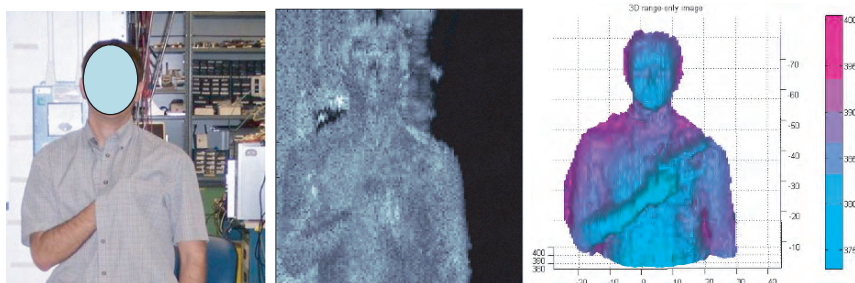


Fig. 12 Photograph of target scenario, where a concealed gun replica is held his shirt at 4 m stand-off range (*left*), total-power image of the received intensity at 600 GHz, on a logarithmic scale, covering the entire final IF bandwidth of 96 kHz (*center*) - which results in very poor scene contrast since no range information is available, and three-dimensional target reconstructions of the front-surfaces encountered by the 600 GHz radar beam during the scan, effectively removing the shirt, leaving the concealed gun exposed on his torso

3 Submillimeter-Wave Sensors

Submillimeter-wave sensors can be broadly categorized in two distinct regimes depending on their application. In the first case, they are used to make images by detecting the thermal emission from the objects. In the second case, spectroscopic studies are carried out either in absorption or in emission. In spectroscopy, there are two generic approaches. Measurements with low spectrum resolution $\lambda/\Delta\lambda$ of 3–10 are called photometry. They are used to characterize broad spectrum sources such as electron synchrotron and thermal bremsstrahlung emission, and thermal emission from interstellar dust grains. Photometry at millimeter and submillimeter wavelengths plays a key role in astrophysics. For example, photometry is used to study the cosmic microwave background (CMB), which lights up the sky at millimeter wavelengths and makes up most of the electromagnetic energy in the universe. CMB measurements provide us a snapshot of the early universe 500,000 years after the Big Bang. Measurements of the anisotropy of this radiation have already shown that the geometry of the universe is flat [38, 39, 40], have provided the most accurate values for several important cosmological parameters [41], and have provided very strong support for an inflation-motivated cosmological model dominated by dark energy and containing substantial dark matter. Another example is the study of interstellar dust emission from galaxies. It has recently been discovered that for some of the earliest, very distant galaxies, most of this starlight does not reach us [42, 43, 44, 45, 46]. The starlight is instead absorbed by dust and reradiated at longer wavelengths. In fact, the total amount of energy in submillimeter light in the universe is about equal to that in the UV/visible/near-IR band [47, 48, 49]. Furthermore, the submillimeter band is unique in that galaxies may be seen out to very large distances, because the dust emission spectrum has a very steep long-wavelength slope ($\sim \nu^{-3.5}$), so that the cosmological frequency redshift essentially compensates for the dimming due to the increasing distance. Measurements of this dust emission by

submillimeter photometry give us the total luminosity of the galaxy, which is related to the rate at which stars are forming, and also provides information about the behaviour of the dust.

The detector technologies used for photometry depend on the wavelength. Cooled high electron mobility transistor (HEMT) amplifiers, followed by diode detectors, are often used for $\lambda > 3$ mm [50]. Thermal direct detectors, especially semiconducting bolometers [51, 52, 53] are in current use for $3 > \lambda > 0.2$ mm. Semiconducting photon detectors are generally used for wavelengths $\lambda < 0.2$ mm [54].

Measurements with higher spectral resolution ($\lambda/\Delta\lambda \sim 10^6$ or higher) are generally different from photometry, and are typically used to characterize molecular and atomic spectral lines [55]. Molecular rotation lines dominate at millimeter and submillimeter wavelengths and atomic fine structure lines at far-IR wavelengths [18]. Resolutions as large as $\lambda/\Delta\lambda \sim 10^6$ are often needed to measure Doppler shifts and spectral line profiles.

The sensors at submillimeter wavelengths can be broadly categorized into two distinct sets: *coherent detectors* and *incoherent (direct) detectors* [56]. At submillimeter wavelengths, coherent detection is mostly done using heterodyne techniques. Heterodyne receivers are quite familiar to most electrical engineers: an RF signal picked up by an antenna at frequency f_{signal} and the sine-wave output of a local oscillator (LO) at frequency f_{LO} are combined in a nonlinear device known as a mixer, which generates the beat frequency f_{IF} , also known as the intermediate frequency (IF). The IF signal may then be further downconverted or demodulated. For radio astronomical spectroscopy, the IF signal is processed by a multichannel spectrum analyzer, such as a filter bank, acoustooptical spectrometer (AOS) [57], or a digital correlator [58, 59, 60, 61, 62]. The IF spectrum of a heterodyne receiver is an exact replica of the original RF spectrum with both the phase and amplitude information. That allows heterodyne systems to have very high spectral resolution, since $f_{\text{IF}} \ll f_{\text{signal}}$. Figure 13 shows the typical schematic block diagram of a submillimeter-wave heterodyne receiver.

Direct detection instruments such as bolometers behind grating or Fabry–Pérot spectrometers [63, 64, 65, 66, 67] are used at moderate resolution, while heterodyne receiver systems are used for higher spectral resolution ($\lambda/\Delta\lambda > 10^3$) [12], [68, 69, 70, 71]. These include the superconductor–insulator–superconductor (SIS) tunnel junction mixer for $3 > \lambda > 0.25$ mm and hot-electron bolometer (HEB) mixers at shorter wavelengths and Schottky diode mixers are used for the entire

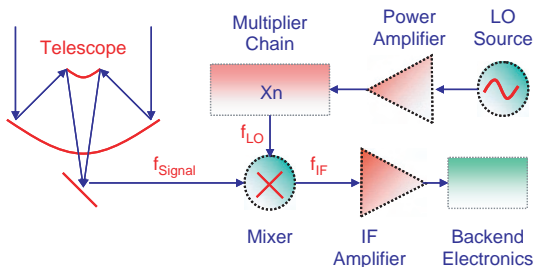
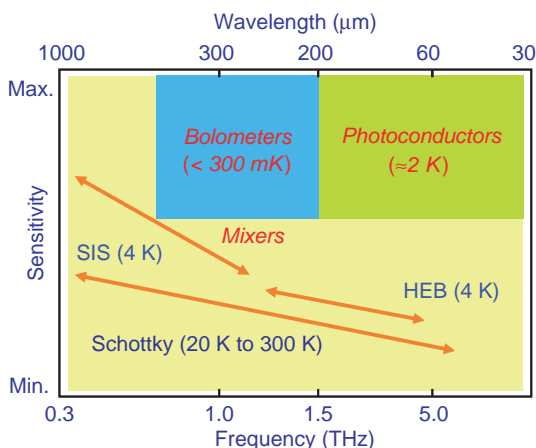


Fig. 13 Typical schematic block diagram of a submillimeter-wave heterodyne receiver

Fig. 14 Generic representation of different submillimeter-wave sensors and their respective sensitivities



submillimeter-wave spectral range. Figure 14 shows a generic representation of different submillimeter-wave sensors – both coherent and incoherent – and their respective sensitivities.

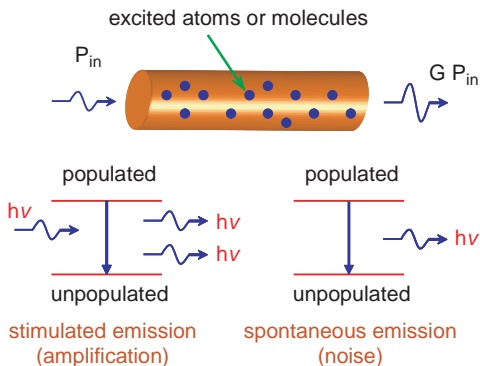
The primary distinction between coherent and incoherent (or direct) detection is the presence or absence of quantum noise. Coherent receivers preserve information about both the amplitude and phase of the electromagnetic field while providing large photon number gain. As a result, coherent receivers are subject to quantum noise, which can be expressed as a minimum noise temperature of $T_n = h\nu/k_B$, or 48 K/THz. Quantum noise is equivalent to the shot noise produced by a background radiation flux of one photon per second per Hertz of detection bandwidth. At radio wavelengths, the background is significantly larger than this value and in any case never falls below the 2.7 K cosmic microwave background (CMB), and so the use of coherent receivers at radio wavelengths need not lead to a loss of sensitivity. In contrast, at optical or infrared wavelengths the quantum noise of coherent receivers is intolerably large, far larger than the typical backgrounds, and so direct detection is strongly preferred. In the following section we compare the sensitivity of coherent detection with incoherent detection in detail and discuss which detection method is preferable under a given circumstances.

3.1 Sensitivity of Coherent (Heterodyne) and Incoherent (Direct) Detection

3.1.1 Quantum Limit for Coherent Detection

Coherent detection is fundamentally different than direct detection. In a coherent system, the incoming signal photons collected by the antenna are first amplified before they are detected. This photon amplification process, which in some cases

Fig. 15 An illustration of quantum noise in a maser amplifier. This fictitious maser amplifier consists of a tube filled with a gas of molecules or atoms. An input signal P_{in} is amplified by stimulated emission and the output is GP_{in} , where G is the power gain. However due to spontaneous emission, noise photons emerge from the amplifier output even when $P_{in} = 0$ [74]



may be simultaneously accompanied by frequency conversion (for heterodyne detection), imposes a limitation on the sensitivity known as the ‘quantum limit’, which is usually expressed as an equivalent noise temperature $T_{QL} = h\nu/k$. Quantum noise was discovered in 1957 [72] in connection with the development of the maser. As shown in Fig. 15, the maser amplifier provides a simple intuitive picture which explains quantum noise as the result of spontaneous emission by the inverted states which produce the amplification. It is easy to see that noise photons would emerge from the amplifier output even when no signal is present at the input. Subsequent work [73] has shown that quantum noise is a general limitation on all *phase insensitive* linear amplifiers and not a peculiar feature of maser amplifiers.

In contrast, there is no fundamental limit to the sensitivity of direct detection. It is possible in theory to make a direct detector which does not produce any signal at its output unless photons are being absorbed. The difference between coherent and incoherent detection can also be understood in terms of the excitation temperature of the quantum states involved in the amplification or detection processes: an ideal direct detector has a small *positive* excitation temperature, $T_x = +\epsilon$, while an ideal amplifier has a small but *negative* excitation temperature, $T_x = -\epsilon$, where in both cases $\epsilon \ll h\nu$.

3.1.2 Direct Detection and Background Noise

The limitations to the sensitivity of direct detection are largely practical. Detectors always produce some noise, although in many cases the detector noise is small compared to the fluctuations in the thermal background radiation from the telescope or the atmosphere. However, detector noise can be an important issue for spectroscopy, where background fluctuations are reduced due to the narrow bandwidth received by the detector. In fact, background-limited high resolution spectroscopy using a cooled space telescope in the submillimeter wavelengths is still largely out of reach with current detector technology.

Whether or not the quantum noise limit for coherent detection is actually a significant factor depends largely on the level of the background radiation. For instance, a lower limit to the thermal background is provided by the cosmic microwave background (CMB) radiation. This means that the quantum noise is actually irrelevant at radio frequencies $h\nu/k < 2.7\text{ K}$ or $\nu < 56\text{ GHz}$., regardless of the temperature and location. In the submillimeter and far-infrared ($\nu < 3\text{ THz}$), the quantum limit is important for cooled telescopes but is not a limiting factor for ground-based or airborne receiver systems.

3.1.3 Comparison of Detection Methods

The background-limited sensitivity of a direct detector receiving a single mode (i.e., diffraction-limited beam and a single polarization) is expressed by the following equation:

$$\sigma_P^d = \frac{h\nu}{\sqrt{\Delta\nu T}} \sqrt{\frac{n_0(1 + \eta^d n_0)}{\eta^d}} \Delta\nu$$

where σ_P^d is the uncertainty in the incident power in a detection bandwidth $\Delta\nu$ after an integration time T using a detector with effective quantum efficiency η^d (including optical losses) [74]. Here n_0 is the mean photon occupation number associated with the thermal background radiation entering the system. For example, in the case that the background arises from objects at a single temperature T_{bg} with total emissivity ε , n_0 will be given by:

$$n_0 = \varepsilon \left[e^{h\nu/kT_{bg}} - 1 \right]$$

For a coherent receiver, the corresponding expression is given by

$$\sigma_P^c = \frac{h\nu}{\sqrt{\Delta\nu T}} \frac{1}{\eta^c} (1 + \eta^c n_0) \Delta\nu$$

The two terms in the sum $(1 + \eta^c n_0)$ correspond to quantum and background noise, respectively. Thermal emission from the optics inside the instrument has been neglected for both types of detection systems. The expression for a coherent receiver assumes that an amplifier or a single-sideband mixer is being used; a factor of 2 must be inserted on the background term for a double-sideband mixer.

We must make assumptions about the quantum efficiencies in order to make a comparison. In general, coherent systems have extremely simple optical paths. In contrast, direct detection systems, particularly high resolution spectrometers, have rather complicated filtering systems to restrict the bandwidth of the radiation landing on the detector, which inevitably reduces the optical transmission and the effective quantum efficiency. For this reason, we assume $\eta^c = 0.5$, which is readily achievable, while for the direct detector spectrometers we assume $\eta^d = 0.1$, which is fairly optimistic for resolutions of 10^4 or higher.

3.2 Coherent (Heterodyne) Sensors

Since low-noise amplifiers are not available in the submillimeter band, the first operation that a coherent receiver must perform is frequency down-conversion, or heterodyning, from submillimeter-wave frequencies to GHz frequencies. This is accomplished in the usual way, using a mixer and a local oscillator; however the mixer noise sets the system sensitivity and should be as low as possible. Figure 13 shows the typical schematic block diagram of a submillimeter-wave heterodyne receiver.

There are at least three different mixer technologies for submillimeter-wave heterodyne detection. The particular choice is mostly dictated by the available local oscillator power, receiver sensitivity criteria, and whether the mixer will operate at room temperature or cryogenic temperatures. In the following section we evaluate the Superconductor insulator superconductor (SIS) mixers, Hot Electron Bolometer (HEB) mixers, and Schottky mixers for their usage at submillimeter wavelengths.

SIS mixers are the most sensitive mixers available today in the 100–1200 GHz frequency range [75]. These mixers are less sensitive at frequencies beyond the superconductor bandgap (2Δ for NbTiN \approx 1500 GHz) where reverse tunnelling becomes a factor and mixer performance is dominated by circuit losses. SIS mixers typically operate at temperatures below 5 K (well below the superconductor critical temperature T_c). Typical state-of-the-art double sideband (DSB) noise temperature for SIS mixers are about 85 K at 500 GHz, have approximately 1 dB of mixer conversion loss, and require approximately 40–100 μ W of local oscillator pump power [76]. For optimal performance, SIS mixers require magnets to suppress Josephson currents. Figure 16 shows the operating principle, energy band diagram, and photo of a SIS mixer device along with the mixer in a receiver configuration.

Hot electron bolometer (HEB) mixers [77] have excellent noise performance from 500 GHz to 5 THz. Both the diffusion cooled and phonon cooled variety use a short superconducting bridge connecting two normal metal pads as the mixing element, and they generally operate at temperatures below 4 K. Typical DSB noise temperatures of HEB mixers are around 600 K at 500 GHz with approximately 10–15 dB of mixer conversion loss. They require approximately 1–2 μ W of LO pump power, which is substantially less than SIS mixers. HEB mixers tend to have only a few gigahertz of IF bandwidth, which is a problem for heterodyne imagers where higher IF bandwidth is advantageous. High temperature HEBs, operating in the 100 K range, have shown promise for use as mixers; however, no published results are available. Figure 17 shows the schematic drawing and close-up photo of a HEB mixer device along with the photo of a 1.6 THz quasi-optical HEB mixer.

Schottky diode mixers operate at frequencies up to well beyond 5 THz [78]. One of the major advantages of Schottky mixers compared to SIS and HEB mixers is that they operate at room temperature, although optimum performance is achieved at or below 20 K. Schottky mixers require high local oscillator pump power, approximately in the 1 mW range. Typical DSB noise temperatures for room temperature Schottky mixers are about 1800 K at 500 GHz with approximately 8 dB of conversion loss. However, their noise temperature improves when cooled, e.g., reaching approximately 1200 K (DSB) at 77 K. It has also been shown that the Schottky mixers

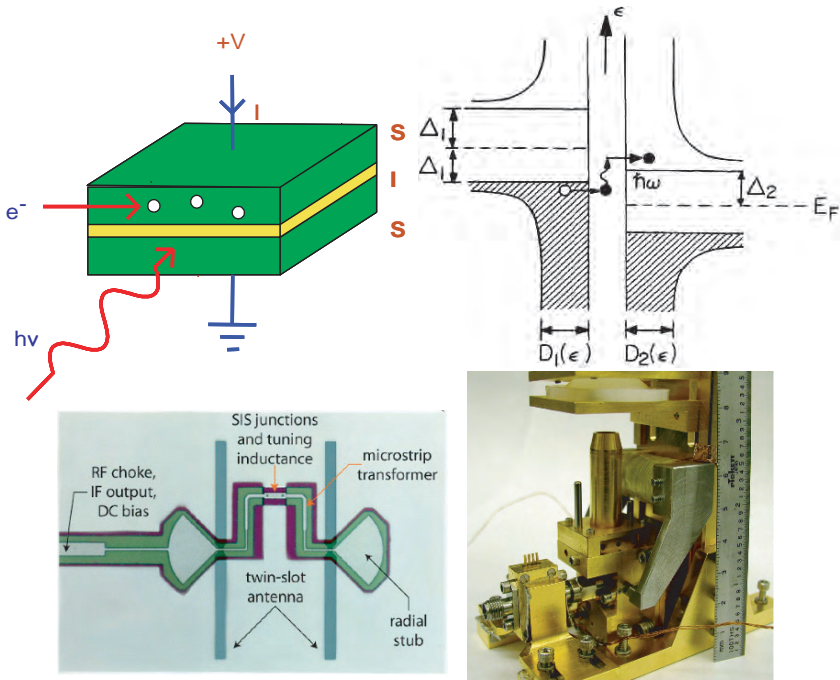


Fig. 16 Operating principle and energy band diagram for SIS mixers(top). Bottom: Quasi-optical SIS mixer device (left) and receiver system (right)

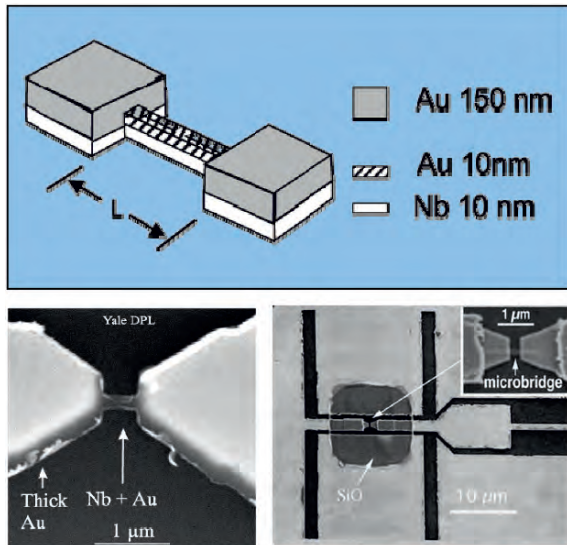


Fig. 17 Schematic drawing and device photo of HEB mixers. Bottom left shows the photo of a 1.6 THz quasi-optical HEB mixer

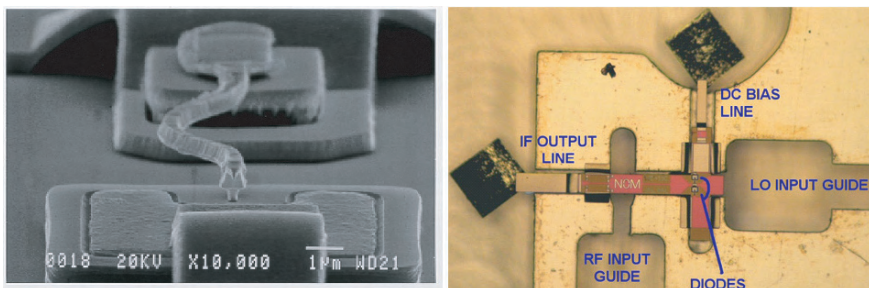


Fig. 18 SEM photo of a 2.5 THz GaAs Schottky diode mixer (*left*) and a photo of a 560 GHz fundamental balanced mixer inside a waveguide block

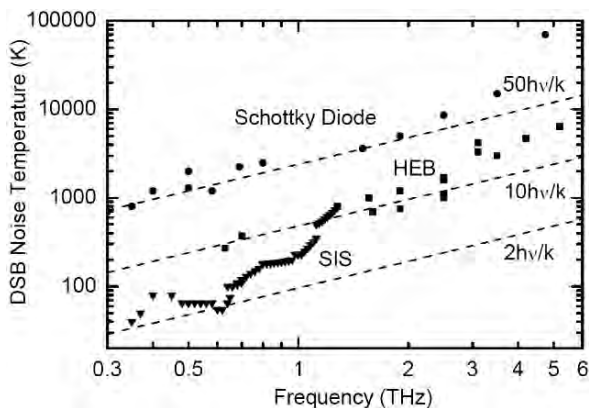


Fig. 19 Double sideband (DSB) noise temperature performance of SIS (*triangles*), HEB (*square*), and Schottky diode (*circle*) mixers. Also shown are the 2-, 10-, and 50-times quantum noise limit lines for comparison

can be operated with reduced LO power at the expense of marginally higher mixer noise temperature [79]. Figure 18 shows photos of GaAs Schottky diode mixers operating at submillimeter wavelengths. Figure 19 shows the performance of different coherent sensors (heterodyne mixers) at submillimeter wavelengths.

3.3 Incoherent (Direct) Detectors

Direct detectors function by absorbing photons in a material and sensing the resulting change in a physical property of that material. In semiconductor photo-detectors, the mobile charge carriers created by photons are sensed by measuring the current flow in response to an applied electric field. The minimum photon energy required to create an excitation sets a long-wavelength cut-off.

Superconducting bolometers are the most sensitive incoherent detectors, their sensitivity increases when the operating temperature of the detector is reduced.

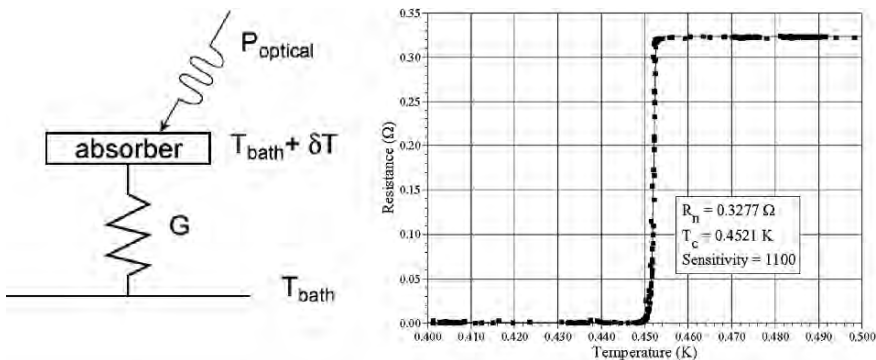


Fig. 20 Operating principle of a transition edge sensor (TES). On the right is a typical resistance vs. temperature plot for such sensors showing the transition temperature between normal metal to superconductor [55]

Superconducting transition edge sensors (TES) thermistors are the most popular direct detectors which promise to have unprecedented sensitivity [80]. A TES bolometer for millimeter and submillimeter wavelengths consists of a radiation absorbing element attached to a thin superconducting film with a transition temperature T_c , which is weakly coupled to a heat sink at temperature $T_0 \sim T_c/2$, as shown in Fig. 20. Also shown in Fig. 21 is the resistance of the TES as a function of temperature, showing the sharp transition at T_c . Figure 21 shows two direct detector array instruments currently in operation.

An alternative approach to photon detection using superconductivity is to operate far below the transition temperature T_c . In this situation, most of the electrons are bound together into Cooper pairs [81]. Photons absorbed in the superconductor may break Cooper pairs to produce single electron quasi-particles, similar to electron-hole pair creation in semiconductors. However, it is difficult to separate the quasi-particles from Cooper pairs in this method.

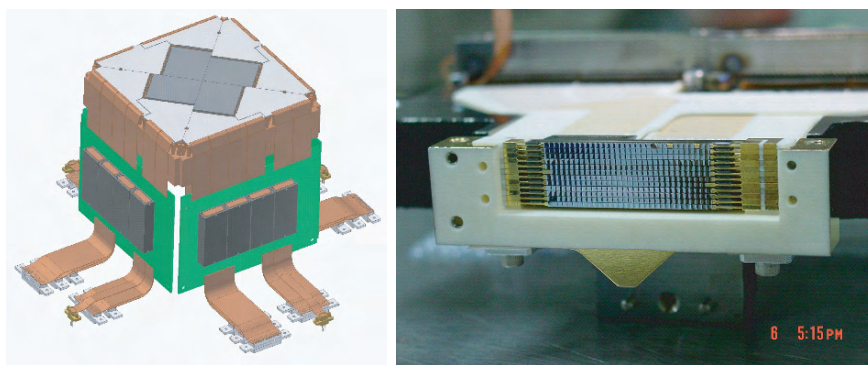
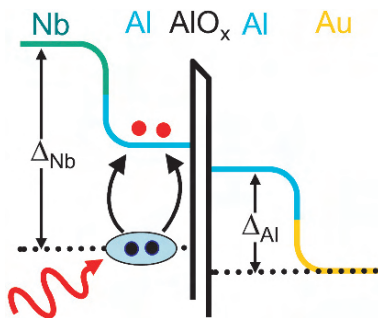


Fig. 21 Photographs of instruments using TES bolometers and silicon thermistors. SCUBA II multiplexed bolometer array (left) and SHARC II array with implanted silicon thermistors

Fig. 22 Operation of superconducting tunnel junctions (STJs) which convert incoming submillimeter-wave photons to current through photon assisted tunnelling



In superconducting tunnel junction detectors (STJ), incoming photons break the Cooper pairs and are filtered out through the tunnel junction [82]. Multiplexed read-out of these detectors can be accomplished using RF single-electron transistors. Figure 22 shows the operation of superconducting tunnel junctions.

The quasi-particles produced by photons when they break the Cooper pairs may also be detected by measuring the complex ac surface impedance of the superconductor. At finite frequencies, the surface impedance is nonzero and is in fact largely inductive; this is known as the kinetic inductance effect. This effect can be used to make very simple detectors where the resonance frequency of a superconducting resonator will change when a photon is detected, and can be monitored with microwave readout circuits [83]. Fig. 23 shows the operation of kinetic inductance detectors.

There are also a few other direct detectors which provide very good sensitivity for detecting photons at submillimeter wavelengths such as the SIS photon detectors

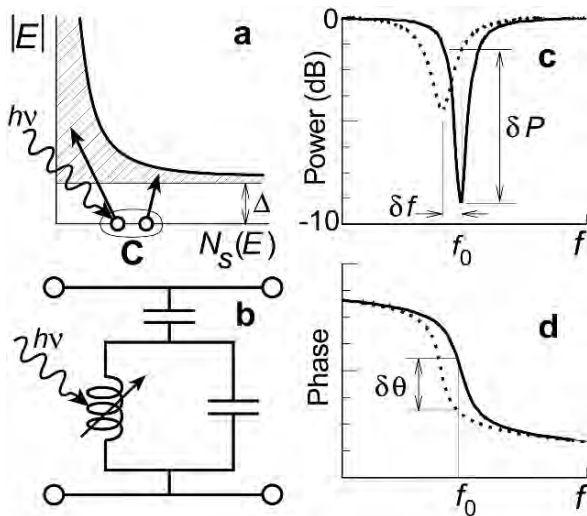


Fig. 23 Operation of a microwave kinetic inductance detector when the change of resonance of a superconducting resonator is detected when a photon is absorbed

where superconducting tunnel junctions directly convert submillimeter photons to electrical current through the process of photon assisted tunneling. There is also another kind of detector known as superconductor insulator normal metal (SIN) sensors which is similar to SIS detectors except one of the metals is a normal metal [84]. There are also room temperature direct detectors which have limited applications because of their poor sensitivity compared to superconducting detectors. However, for applications where cryogenic detectors are not feasible, they play an active role. Small area GaAs Schottky diodes, composite bolometers with bismuth or tellurium, and Golay cells are all used for room temperature direct detectors. Reference [12] has a good review on these detectors.

4 Future Trends

Receivers at submillimeter wavelengths are either waveguide based or quasi-optical, although multi-pixel direct detectors with planar architecture are being used in recent years [85]. At frequencies beyond a few hundred gigahertz, the feature sizes of all but the simplest waveguide circuits are too small and the required tolerances are too demanding to be fabricated using even the best state-of-the-art conventional

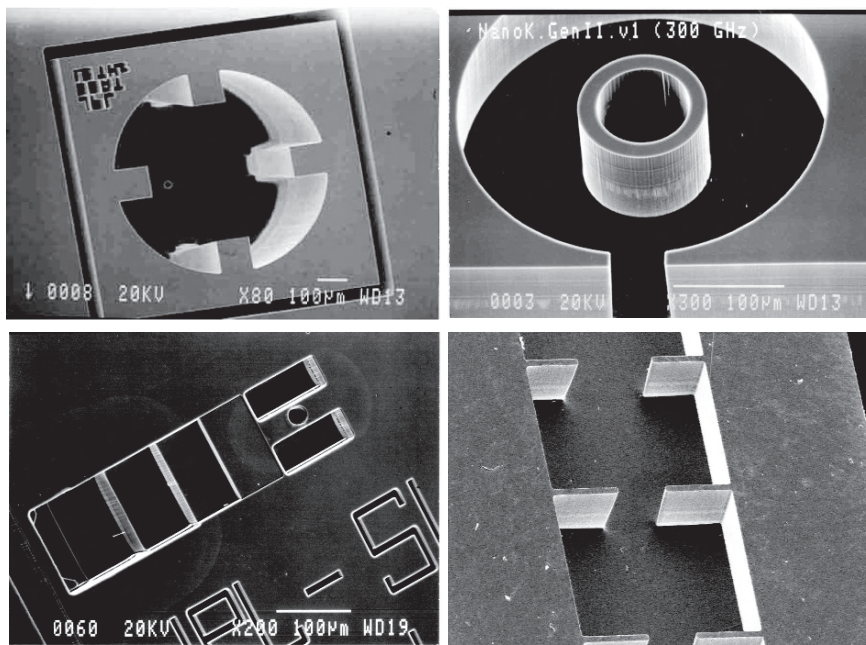


Fig. 24 Photograph of some of the silicon micromachined components fabricated at the Jet Propulsion Laboratory at submillimeter wavelengths

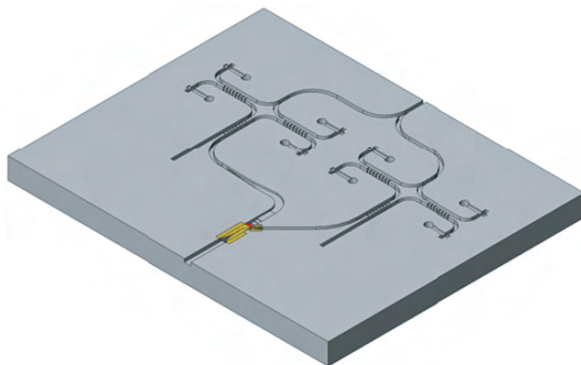


Fig. 25 Schematic of a silicon micromachined dual-polarized sideband separating balanced heterodyne receiver at submillimeter wavelengths.

machining. To mitigate this situation, silicon micromachining is becoming very popular. Silicon micromachining shatters the barriers of conventional machining, and brings the added benefits of rapid turn-around time and excellent process control. Furthermore, silicon micromachined circuits are light weight and capable of achieving a high degree of integration on a single chip. While there have been several demonstrations of waveguide circuits fabricated with silicon micromachining, as shown in Fig. 24, few if any of these circuits have been subjected to any significant electrical testing. It is quite possible that the future multi-pixel, multi-functional, high performance astrophysics and planetary instruments at submillimeter wavelengths will depend on the success of silicon micromachined waveguide components [86]. Figure 25 shows schematic of a multi-functional dual-polarized, sideband separating, balanced receiver architecture based on silicon micromachining. On the other hand, the direct detector community is primarily focusing their attention to large antenna coupled planar arrays which will require minimal hand assembly. These techniques are going to drive the development of submillimeter sensors in the coming years.

5 Conclusion

Sensors at submillimeter wavelengths have a wide range of applications, from astrophysics, planetary, Earth-observing, plasma diagnostics to medical imaging and homeland security. Due to severe atmospheric attenuation at these frequencies except for a handful available of windows, most of the sensors at these frequencies are used for space-borne applications. Sensors at these wavelengths provide unprecedented sensitivity. Coherent detectors are reaching quantum-noise limit, and the sensitivity direct detector instruments are not limited by the background noise. This is an exciting time for submillimeter-wave sensors.

Acknowledgements The author wants to thank the members of the Submillimeter-Wave Advanced Technology (SWAT) Group at the Jet Propulsion Laboratory (JPL). This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, under contract with National Aeronautics and Space Administration.

References

1. P. H. Siegel, "Terahertz Instruments for Space," *IEEE Trans. on Ant. and Prop.*, vol. 55, pp. 2957–2965, Nov. 2007.
2. R. R. Cladwell and M. Kamionkowski, "Echoes from the Big Bang," *Scientific American*, pp. 38–43, 2001.
3. R. D. Lorenz, G. Biolluz, P. Encrenaz, M. A. Janssen, R. D. West, and D. O. Muhelman, "Cassini Radar: Prospects for Titan Surface Investigations Using the Microwave Radiometer," *Planetary and Space Science*, vol. 51, no. 4–5, pp. 353–364, April-May 2003.
4. J. Waters, "Submillimeter-Wavelength Heterodyne Spectroscopy and Remote Sensing of the Upper Atmosphere," *Proc. IEEE*, vol. 80, no. 11, Nov. 1992.
5. P. H. Siegel and R. J. Dengler, "Terahertz Heterodyne Imaging: Instruments," *Int. Journal of Infrared and Millimeter Waves*, vol. 27, no. 5, pp. 631–656, May 2006.
6. D. L. Woolard, E. Brown, M. Pepper, M. Kemp, "Terahertz Frequency Sensing and Imaging: A Time of Reckoning Future Applications?," *Proc. IEEE*, vol. 93, no. 10, pp. 1722–1743, Oct. 2005.
7. D. Mittleman, "Terahertz Imaging", in *Sensing with Terahertz Radiation*, D. Mittleman Ed. Berlin, Germany: Springer-Verlag, pp. 117–153, 2003.
8. K. B. Cooper, R. J. Dengler, G. Chattopadhyay, E. Schlecht, J. Gill, A. Skalare, I. Mehdi, and P. H. Siegel, "A High Resolution Imaging Radar at 580 GHz," *IEEE Microwave and Wireless Comp. Lett.*, vol. 18, no. 1, pp. 64–66, Jan. 2008.
9. R. Appleby and H. B. Wallace, "Standoff Detection of Weapons and Contraband in the 100 GHz to 1 THz Region," *IEEE Trans. on Ant. and Prop.*, vol. 55, pp. 2944–2956, Nov. 2007.
10. P. H. Siegel, "Terahertz Technology in Biology and Medicine," *IEEE Trans. Microwave Theory Tech.*, vol. 52, no. 10, pp. 2438–2447, Oct. 2004.
11. M. Tonouchi, "Cutting Edge THz Technology," *Nature Photonics*, vol. 1, pp. 97–105, Feb. 2007.
12. P. H. Siegel, "Terahertz Technology," *IEEE Trans. on Microwave Theory and Tech.*, vol. 50, no. 3, pp. 910–928, March 2002.
13. G. Chattopadhyay, "Sensor Technology at Submillimeter Wavelengths for Space Applications," *Proceedings of the Second International Conference on Sensing Technology*, Palmerston North, New Zealand, Nov. 2007.
14. G. L. Pilbratt, "The Herschel mission, scientific objectives, and this meeting," *Proc. Eur. Space Agency Symp.*, Dec. 2000, ESA SP-460, pp. 13–20.
15. R. L. Brown, "Technical specification of the millimeter array," *Proc. SPIE-Int. Soc. Opt. Eng.*, no. 3357, pp. 231–441, 1998.
16. E. F. Erickson and J. A. Davidson, "SOFIA: The Future of Airborne Astronomy," *Proc. Airborne Astronomy Symp. on the Galactic Ecosystem: From Gas to Stars to Dust*, eds., M. R. Haas, J. A. Davidson, and E. F. Erickson, San Francisco, April 1995.
17. D. J. Fixsen, C. L. Bennett, and J. C. Mather, "COBE Far Infrared Absolute Spectrophotometer Observations of Galactic Lines," *Astrophysical Journal*, 526: 207–214, 1999 Nov. 20.
18. T. G. Phillips and J. Keene, "Submillimeter Astronomy", *Proc. IEEE*, vol. 80, pp 1662–1678, Nov. 1992.
19. G. Melnick et al., "The Submillimeter Wave Astronomy Satellite: Science Objectives and Instrument Description," *Astrophys. J. Lett.*, pt. 2, vol. 539, no. 2, pp. L77–L85, Aug. 2000.

20. W. Wild, et al., "ESPRIT: A Space Interferometer Concept for the Farinfrared", *Proc. SPIE*, 6255, p. 62651Z, 2006.
21. Space Science Enterprise 2000: Strategic Plan (spacescience.nasa.gov/roadmap/pdffiles/2000/2-3.pdf).
22. G. Schwehm and R. Schulz, "The International Rosetta Mission," Ehrenfreund et al., eds: *Laboratory Astrophysics and Space Research*, 1999, pp. 537–546.
23. S. Gulkis, et. al., "MIRO: Microwave Instrument for Rosetta Orbiter," *Space Science Reviews*, vol. 128, no. 1-4, pp. 561–597, Feb. 2007.
24. M. C. Gaidis, "Space-Based Applications of Far Infrared Systems," *Eighth Intl. Terahertz Electron. Conf.*, Darmstadt, Germany, Sept. 28–29, 2000, pp. 125–128.
25. Earth Science Enterprise Strategic Plan, "Exploring Our Home Planet," *NASA Headquarters*, November 2000 (www.earth.nasa.gov).
26. S. Solomon, "Stratospheric Ozone Depletion: A Review of Concepts and History," *Reviews of Geophysics*, vol. 37, pp. 275–316, 1999.
27. World Meteorological Organization, "Tropospheric Ozone and Related Processes," Chapter 8, *Scientific Assessment of Ozone Depletion*: 1998, Rep. 44, Geneva, Switzerland, 1998.
28. A. M. Thomson, "The Oxidizing Capacity of the Earth's Atmosphere: Probable Past and Future Changes," *Science*, vol. 256, pp. 1157–1165, 1992.
29. M. A. Janssen, editor, *Atmospheric Remote Sensing by Microwave Radiometry*, John Wiley, 1993.
30. J. W. Waters, "Submillimeter-Wavelength Heterodyne Spectroscopy and Remote Sensing of the Upper Atmosphere," *Proc. IEEE*, vol. 80, no. 11, pp. 1679–1701, November 1992.
31. J. W. Waters, et. al., "The UARS and EOS Microwave Limb Sounder Experiments," *Journal of Atmospheric Science*, vol. 56, pp. 194–218, 1999.
32. D. T. Petkie, F. C. DeLucia, C. Casto, P. Helminger, E. L. Jacobs, S. K. Moyer, S. Murrill, C. Halford, S. Griffin, and C. Franck, "Active and Passive Millimeter and Submillimeter-Wave Imaging," *Proc. SPIE*, vol. 5989, pp. 598918–1 to 598918–8, 2005.
33. J. C. Dickinson, T. M. Goyette, A. J. Gatesman, C. S. Joseph, Z. G. Root, R. H. Giles, J. Waldman, and W.E. Nixon, "Terahertz Imaging of Subjects with Concealed Weapons," *Proc. SPIE*, vol. 6212, pp. 62120Q–1 to 62120Q–12, 2006.
34. M. C. Kemp, P. F. Taday, B. E. Cole, J. A. Cluff, A. J. Fitzgerald, and W. R. Tribe, "Security Applications of Terahertz Technology," *Proc. SPIE*, vol. 5070, pp. 44–52, 2003.
35. G. Chattopadhyay, K. B. Cooper, R. J. Dengler, E. Schlecht, A. Skalare, I. Mehdi, and P. H. Siegel, "A 675 GHz FMCW Radar with Sub-Centimeter Range Resolution," *Proceedings of the Eighteenth International Symposium on Space Terahertz Technology*, Pasadena, CA, USA, March 2007.
36. R. J. Dengler, K. B. Cooper, G. Chattopadhyay, I. Mehdi, E. Schlecht, A. Skalare, C. Chen, and P. H. Siegel, "600 GHz Imaging Radar with 2 cm Range Resolution," *2007 IEEE MTT-S Intl. Microwave Symp. Digest*, Honolulu, Hawaii, June 2007, pp. 1371–1374.
37. N. C. Luhmann and W. A. Peebles, "Instrumentation for Magnetically Confined Fusion Plasma Diagnostics," *Rev. Sci. Instrum.*, vol. 55, no. 3, pp. 279–331, March 1984.
38. P. de Bernardis, P. Ade, J. Bock, J. Bond, J. Borrill, A. Boscaleri, K. Coble, B. Crill, G. De Gasperis, P. Farese, P. Ferreira, K. Ganga, M. Giacometti, E. Hivon, V. Hristov, A. Iacoangeli, A. Jaffe, A. Lange, L. Martinis, S. Masi, P. Mason, P. Mauskopf, A. Melchiorri, L. Miglio, T. Montroy, C. Netterfield, E. Pascale, F. Piacentini, D. Pogosyan, S. Prunet, S. Rao, G. Romeo, J. Ruhl, F. Scaramuzzi, D. Sforna, and N. Vittorio, "A Flat Universe from High-Resolution Maps of the Cosmic Microwave Background Radiation," *Nature*, vol. 404, no. 6781, pp. 955–959, 2000.
39. S. Hanany, P. Ade, A. Balbi, J. Bock, J. Borrill, A. Boscaleri, P. De Bernardis, P. Ferreira, V. Hristov, A. Jaffe, A. Lange, A. Lee, P. Mauskopf, C. Netterfield, S. Oh, E. Pascale, B. Rabii, P. Richards, G. Smoot, R. Stompor, C. Winant, and J. Wu, "MAXIMA-1: A Measurement of the Cosmic Microwave Background Anisotropy on Angular Scales of 10^{-5} degrees," *Astrophys. J.*, vol. 545, no. 1, pp. L5–L9, 2000.

40. N. Halverson, E. Leitch, C. Pryke, J. Kovac, J. Carlstrom, W. Holzapfel, M. Dragovan, J. Cartwright, B. Mason, S. Padin, T. Pearson, A. Readhead, and M. Shepherd, "Degree Angular Scale Interferometer First Results: A Measurement of the Cosmic Microwave Background Angular Power Spectrum," *Astrophys. J.*, vol. 568, no. 1, pp. 38–45, 2002.
41. D. Spergel, L. Verde, H. Peiris, E. Komatsu, M. Nolta, C. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. Meyer, L. Page, G. Tucker, J. Weiland, E. Wollack, and E. Wright, "First-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters," *Astrophys. J. (Suppl.)*, vol. 148, no. 1, pp. 175–194, 2003.
42. I. Smail, R. J. Ivison, and A.W. Blain, "A Deep Submillimeter Survey of Lensing Clusters: A New Window on Galaxy Formation and Evolution," *Astrophys. J.*, vol. 490, pp. L5–L8, 1997.
43. A. Barger, L. Cowie, D. Sanders, E. Fulton, Y. Taniguchi, Y. Sato, K. Kawara, and H. Okuda, "Submillimeter-Wavelength Detection of Dusty Star-Forming Galaxies at High Redshift," *Nature*, vol. 394, no. 6690, pp. 248–251, 1998.
44. D. Hughes, S. Serjeant, J. Dunlop, M. Rowan-Robinson, A. Blain, R. Mann, R. Ivison, J. Peacock, A. Efstathiou, W. Gear, S. Oliver, A. Lawrence, M. Longair, P. Goldschmidt, and T. Jenness, "High-Redshift Star Formation in the Hubble Deep Field Revealed by a Submillimeter-Wavelength Survey," *Nature*, vol. 394, no. 6690, pp. 241–247, 1998.
45. A. W. Blain, I. Smail, R. J. Ivison, J.-P. Kneib, and D. T. Frayer, "Submillimeter Galaxies," *Phys. Rep.*, vol. 369, pp. 111–176, 2002.
46. S. C. Chapman, A.W. Blain, R. J. Ivison, and I. R. Smail, "A Median Redshift of 2.4 for Galaxies Bright at submillimeter Wavelengths," *Nature*, vol. 422, pp. 695–698, Apr. 2003.
47. J.-L. Puget, A. Abergel, J.-P. Bernard, F. Boulanger, W. B. Burton, F.-X. Desert, and D. Hartmann, "Tentative Detection of a Cosmic Far infrared Background with COBE," *Astron. Astrophys.*, vol. 308, pp. L5–L8, Apr. 1996.
48. D. J. Fixsen, E. Dwek, J. C. Mather, C. L. Bennett, and R. A. Shafer, "The Spectrum of the Extragalactic Far-Infrared Background from the COBE FIRAS Observations," *Astrophys. J.*, vol. 508, pp. 123–128, Nov. 1998.
49. M. G. Hauser, R. G. Arendt, T. Kelsall, E. Dwek, N. Odegard, J. L. Weiland, H. T. Freudenreich, W. T. Reach, R. F. Silverberg, S. H. Moseley, Y. C. Pei, P. Lubin, J. C. Mather, R. A. Shafer, G. F. Smoot, R. Weiss, D. T. Wilkinson, and E. L. Wright, "The COBE Diffuse infrared Background Experiment Search for the Cosmic Infrared Background. I. Limits and detections," *Astrophys. J.*, vol. 508, pp. 25–43, Nov. 1998.
50. J. C. Webber and M. W. Pospieszalski, "Microwave Instrumentation for Radio Astronomy," *IEEE Trans. Microwave Theory Tech.*, vol. 50, pp. 986–995, Mar. 2002.
51. P. Richards, "Bolometers for Infrared and Millimeter Waves," *J. Appl. Phys.*, vol. 76, no. 1, pp. 1–24, July 1994.
52. A. Turner, J. Bock, J. Beeman, J. Glenn, P. Hargrave, V. Hristov, H. Nguyen, F. Rahman, S. Sethuraman, and A. Woodcraft, "Silicon Nitride Micromesh Bolometer Array for Submillimeter Astrophysics," *Appl. Opt.*, vol. 40, no. 28, pp. 4921–4932, 2001.
53. C. D. Dowell, J. E. Groseth, T. G. Phillips, C. A. Allen, S. R. Babu, M. D. Jhabvala, S. H. Moseley, and G. M. Voellmer, "The 12×32 Pop-up Bolometer Array for the SHARC II camera," *Proc. Far-IR, Sub-MM, and MM Detector Workshop*, vol. NASA/CP-2003-211 408, J. Wolf, J. Farhoomand, and C. R. Mc-Creight, Eds., 2003, pp. 251–254.
54. G. B. Heim, M. L. Henderson, K. I. Macfeely, T. J. McMahon, D. Michika, R. J. Pearson, G. H. Rieke, J. P. Schwenker, D.W. Strecker, C. L. Thompson, R. M. Warden, D. A. Wilson, and E. T. Young, "Multiband Imaging Photometer for SIRTF," *Proc. SPIE, Space Telescopes Instrum. V*, vol. 3356, pp. 985–1000, Aug. 1998.
55. J. Zmuidzinas and P. L. Richards, "Superconducting Detectors and Mixers for Millimeter and Submillimeter Astrophysics," *Proc. IEEE*, vol. 92, no. 10, pp. 1597–1616, Oct. 2004.
56. J. Zmuidzinas, "Thermal Noise and Correlations in Photon Detection," *Appl. Optics*, vol. 42, no. 25, pp. 4989–5008, Sept. 2003.

57. J. Horn, O. Siebertz, F. Schmulling, C. Kunz, R. Schieder, and G. Winnewisser, "A 4×1 GHz Array Acousto-Optical Spectrometer," *Exp. Astron.*, vol. 9, no. 1, pp. 17–38, 1999.
58. J. Hagen and D. Farley, "Digital-Correlation Techniques in Radio Science," *Radio Sci.*, vol. 8, no. 8–9, pp. 775–784, 1973.
59. W. Urry, D. Thornton, and J. Hudson, "The Hat Creek Millimeter-Wave Hybrid Spectrometer for Interferometry," *Publ. Astron. Soc. Pac.*, vol. 97, no. 594, pp. 745–751, 1985.
60. E. Crete, M. Giard, L. Ravera, J. Noullet, P. d'Aguerre, M. Torres, and J. Mayvial, "Digital GaAs Autocorrelator Developed for Space Borne Submillimeter Astronomy," *Exp. Astron.*, vol. 8, no. 3, pp. 239–255, 1998.
61. M. Torres, "A Frequency-agile Hybrid Spectral Correlator for mm-Wave Radio Interferometry," *Rev. Sci. Instrum.*, vol. 65, no. 5, pp. 1537–1540, 1994.
62. S. Padin, T. Clark, M. Ewing, R. Finch, R. Lawrence, J. Navarro, S. Scott, N. Scoville, C. Seelinger, and T. Seling, "A High-Speed Digital Correlator for Radio Astronomy," *IEEE Trans. Instrum. Meas.*, vol. 42, pp. 793–798, Aug. 1993.
63. C. M. Bradford, B. J. Naylor, J. Zmuidzinas, J. J. Bock, J. Gromke, H. Nguyen, M. Dragovan, M. Yun, L. Earle, J. Glenn, H. Matsuhara, P. A. R. Ade, and L. Duband, "WaFIRS: A Waveguide Far-IR Spectrometer: Enabling spectroscopy of High- z Galaxies in the Far-IR and Submillimeter," *Proc. SPIE, IR Space Telescopes Instrum.*, vol. 4850, pp. 1137–1148, Mar. 2003.
64. T. Nikola, S. Hailey-Dunsheath, G. J. Stacey, D. J. Benford, S. H. Moseley, and J. G. Staguhn, "ZEUS: A Submillimeter Grating Spectrometer for Exploring Distant Galaxies," *Proc. SPIE, Millimeter Submillimeter Detectors Astronomy*, vol. 4855, pp. 88–99, Feb. 2003.
65. J. G. Staguhn, D. J. Benford, F. Pajot, T. J. Ames, J. A. Chervenak, E. N. Grossman, K. D. Irwin, B. Maffei, S. H. Moseley, T. G. Phillips, C. D. Reintsema, C. Rioux, R. A. Shafer, and G. M. Voellmer, "Astronomical Demonstration of Superconducting Bolometer Arrays," *Proc. SPIE, Millimeter Submillimeter Detectors Astronomy*, vol. 4855, pp. 100–107, Feb. 2003.
66. A. Poglitsch, C. Waelkens, and N. Geis, "The photodetector array camera and spectrometer (PACS) for the Herschel Space Observatory," *Proc. SPIE, IR Space Telescopes Instrum.*, vol. 4850, pp. 662–673, Mar. 2003.
67. C. M. Bradford, G. J. Stacey, M. R. Swain, T. Nikola, A. D. Bolatto, J. M. Jackson, M. L. Savage, J. A. Davidson, and P. A. R. Ade, "SPIFI: A Direct-Detection Imaging Spectrometer for Submillimeter Wavelengths," *Appl. Opt.*, vol. 41, pp. 2561–2574, May 2002.
68. T. de Graauw, N. D. Whyborn, H. van de Stadt, G. Beaudin, D. A. Beintema, V. Belitsky, P. Cais, E. Caux, M. Gheudin, A. Cros, P. de Groene, A. Emrich, N. R. Erickson, T. C. Gaier, and J. D. Gallego-Puyol, "Heterodyne Instrument for FIRST (HIFI): Preliminary Design," *Proc. SPIE, Adv. Technol. MMW, Radio, Terahertz Telescopes*, vol. 3357, pp. 336–347, July 1998.
69. M. L. Edgar and J. Zmuidzinas, "CASIMIR: A Submillimeter Heterodyne Spectrometer for SOFIA," *Proc. SPIE, Airborne Telescope Syst.*, vol. 4014, pp. 31–42, June 2000.
70. R. Guesten, I. Camara, P. Hartogh, H. Huebers, U. U. Graf, K. Jacobs, C. Kasemann, H. Roeser, R. T. Schieder, G. Schnieder, O. Sievertz, J. Stutzki, G. Villanueva, A. Wagner, P. van der Wal, and A. Wunsh, "GREAT: The German Receiver for Astronomy at Terahertz Frequencies," *Proc. SPIE, Airborne Telescope Syst. II*, vol. 4857, pp. 56–61, Feb. 2003.
71. J. C. Pearson, I. Mehdi, E. Schlecht, F. Maiwald, A. Maestrini, J. J. Gill, S. C. Martin, D. Pukala, J. Ward, J. Kawamura, W. R. McGrath, W. Hatch, D. G. Harding, H. G. LeDuc, J. A. Stern, B. Bumble, L. A. Samoska, T. C. Gaier, R. Ferber, D. Miller, A. Karpov, J. Zmuidzinas, T. G. Phillips, N. R. Erickson, J. Swift, Y. Chung, R. Lai, and H. Wang, "THz Frequency Receiver Instrumentation for Herschel's Heterodyne Instrument for Far Infrared (HIFI)," *Proc. SPIE, IR Space Telescopes Instrum.*, vol. 4850, pp. 650–661, March 2003.
72. K. Shimoda, H. Takahasi, and C. H. Townes, "Fluctuations in Amplification of Quanta with Application to Maser Amplifier," *J. Phys. Soc. Japan*, 12, 686–700.
73. H. A. Haus and J. A. Muller, "Quantum Noise in Linear Amplifier," *Phys. Rev.*, 128, 2407–2413.

74. J. Zmuidzinias, "Progress in Coherent Detection Methods," *Physics and Chemistry of the Interstellar Medium (V. Ossenkopf et al., ed.)*, pp. 423-430, U. Cologne, GCA-Verlag Herdecke, 1999.
75. J. Zmuidzinias, J. W. Kooi, J. Kawamura, G. Chattopadhyay, J. A. Stern, B. Bumble, and H. G. LeDuc, "Development of SIS Mixers for 1 THz," *Proc. SPIE*, T. G. Phillips, ed., vol. 3357, Kona, Hawaii, March 1998, pp. 53-61.
76. G. Chattopadhyay, D. Miller, H. G. LeDuc, and J. Zmuidzinias, "A Dual-Polarized Quasi-Optical SIS Mixer at 550 GHz," *IEEE Trans. Microwave Theory and Tech.*, vol. 48, no. 10, pp. 1680-1686, October 2000.
77. E. M. Gershenzon, G. N. Gol'tsman, I. G. Gogidze, Y. P. Gusev, A. I. Elant'ev, B. S. Karasik, and A. D. Semenov, "Millimeter and Submillimeter Range Mixer Based on Electronic Heating of Superconducting Films in Resistive State," *Sov. Phys. Superconductivity*, vol. 3, 1582, 1990.
78. M. C. Gaidis, H. M. Pickett, C. D. Smith, S. C. Martin, R. P. Smith, and P. H. Siegel, "A 2.5 THz Receiver Front End for Space Borne Applications," *IEEE Trans. Microwave Theory and Tech.*, vol. 48, no. 4, pp. 733-739, April 2000.
79. J. L. Hesler, W. R. Hall, T. W. Crowe, R. M. Weikle, B. S. Deaver, R. F. Bradley, and S-K Pan, "Fixed-Tuned Submillimeter Wavelength Waveguide Mixers Using Planar Schottky-Barrier Diodes," *IEEE Trans. Microwave Theory and Tech.*, vol. 45, no. 5, pp. 653-658, May 1997.
80. K. D. Irwin and G. C. Hilton, "Transition-Edge Sensors," *Top. Appl. Phys.*, vol. 99, pp. 63-149, 2005.
81. M. Tinkham, *Introduction to Superconductivity*, 2nd ed. New York: McGraw-Hill, 1996.
82. E. Burstein, D. N. Langenberg, and B. N. Taylor, "Superconductors as Quantum Detectors for Microwave and Sub-millimeter-Wave Radiation," *Phys. Rev. Lett.*, vol. 6, no. 3, pp. 92-94, February 1961.
83. P. Day, H. LeDuc, B. Mazin, A. Vayonakis, and J. Zmuidzinias, "A broadband superconducting detector suitable for use in large arrays," *Nature*, vol. 425, no. 6960, pp. 817-821, 2003.
84. D. Golubev and L. Kuzmin, "Nonequilibrium Theory of a Hot-Electron Bolometer with Normal Metal-Insulator-Superconductor Tunnel Junction," *J. Appl. Phys.*, vol. 89, no. 11, pp. 6464-6472, 2001.
85. G. Chattopadhyay, C-L. Kuo, P. Day, J. J. Bock, J. Zmuidzinias, and A. E. Lange, "Planar Antenna Arrays for CMB Polarization Detection," *Proc. of the 33rd International Conference on Infrared, Millimeter, and Terahertz Waves*, Cardiff, United Kingdom, September 2007.
86. G. Chattopadhyay, "Future of Heterodyne Receivers at Submillimeter Wavelengths," *Proc. of the Joint 30th. International Conference on Infrared and Millimeter Waves and 13th. International Conference on Terahertz Electronics*, Williamsburg, Virginia, pp. 461-462, September 2005.

Index

A

- Action selection, 362
- Adaptation, 362
- Adaptive tracking window
 - size, 340
- Ad hoc on demand distance vector (AODV), 203
- Air-coupled ultrasonics, 297
- Ambient intelligence, 204
- AMR, *see* Anisotropic magneto-resistors (AMR)
- Angular position sensor, 16
- Angular sensors, 11
- Anisotropic magneto-resistors (AMR), 28–29, 47
 - magnetic sensors, accuracy, 51–54
 - model, 30–41
 - sensors, 28–29
- AODV, *see* Ad hoc on demand distance vector (AODV)
- Argon fluoride (ArF) excimer lasers, 117
- ART, *see* Asymmetric and Reliable Transport (ART)
- Artificial emotion, 356–357
- Asymmetric and reliable transport (ART), 236–237
- Asynchronous cooperation, 357
- Atacama Large Millimeter Array (ALMA), 390
- Attenuation, 291–292
- Auction, 357
- Average delivery
 - overhead, 226
 - ratio, 226
- Average packet latency, 227
- 3-axis sensor, 15

B

- Band-aid
 - contact Hall, 25
 - device, 25
- Base station (BS), 209
- BAS sensor networks, 213
- Bates' OZ Project, 365
- Beer-Lambert Law, 181
- Bent fibre, 163
- Berger's experiment, 31
- Bidirectional dialog mode, 204–205
- Bioimpedance, 77
 - sensor, measurement constraints related to, 78–79
 - spectroscopy, metrological aspects of, 77–78
- Biological sensors, 126
- Biological tissues, dielectric characterization of, 75–77
 - anisotropic medium, measurements on, 84–86
 - electrodes configuration, 86
 - bioimpedance
 - sensor, measurement constraints related to, 78–79
 - spectroscopy, metrological aspects of, 77–78
 - biological tissue sample, measurement constraints related to, 79
 - bone anisotropy measurements at low frequency, 86–87
 - ex vivo* results, 80, 81
 - discussion, 81–82
 - modelization approach, 82
 - permittivity and conductivity, relation between, 86
 - simulation, 83–84

human blood, measurements on, 80
 Biological tissue sample, measurement constraints related to, 79
 Blob identification, 340–341
 Bluetooth, 201–202
 Bone anisotropy measurements at low frequency, 86–87
 Bottom-up nanotechnology, 154
 Bounding box, 276
 Bragg wavelength, 119
 Broadcast data mode, 204
 Broadcasting of local eligibility, 358
 Buffer rod sensor, 288, 300–301
 Building automation, 212–213
 Bulk (fundamental) waves, 290
 Bulk waves, 290

C

Cathexis model, 364
 CBM, *see* Condition-based maintenance (CBM)
 CCD camera, 249, 254–255
 CCF, *see* Congestion Control and Fairness (CCF)
 CC2431 location engine, 272
 Cell manager, 277
 Cell-to-geometry visibility, 276
 Channel sampling, 225
 Cladding modes, 165, 171
 Cluster-based architectures, 208–210
 CMOS Hall sensors, 3
 angular position sensor, 16
 contactless joystick sensor, 18
 contactless magnetic position measurement, 4–5
 electrical current sensor, 18–19
 electronic compass, 19–20
 Hall effect and Hall element
 galvano-magnetic sensor, 3–4
 hysteresis, 13–14
 IMC, 6–8
 layer, properties of, 9
 material properties, 9–10, 11–12
 process, 8–9
 linear position sensor, 17
 multi-axis sensing, 5–6
 saturation, 12–13
 sensor architecture, 14–15
 Coated FBGs based on chemo-mechanical and chemo-thermal effects, 146
 CODA, *see* Congestion Detection and Avoidance (CODA)
 Coherent detection, 402
 Coherent (heterodyne) sensors, 403–405

Color image sequence, 334
 Color space, 347–348
 transformation, 347
 Colour segmentation, area thresholding, blob merging, 334–337
 Commercial magnetic sensors, 23–24
 AMR model, 30–41
 AMR sensors, 28–29
 future progress, 41–42
 Hall sensor design, 24–27
 Computer-aided design (CAD) floor plan, 273
 Condition-based maintenance (CBM), 216
 Congestion bit, 230
 Congestion Control and Fairness (CCF), 237–238
 Congestion degree, 226
 Congestion Detection and Avoidance (CODA), 229–230
 Contactless joystick sensor, 18
 Contactless magnetic joystick, 18
 Contactless magnetic position measurement, 4–5
 Contactless position measurement, 4–5
 Contact-type transducers, 296
 Converge-cast data flow, 223
 Cosmic microwave background (CMB), 388–389, 398
 radiation, 402
 Cosmology, 391
 Counter-balancing electric field, 24
 Cricket-based location-aware system, 270
 Crossfield effect of, 51–55
 AMR, 51–54
 fluxgate, 54–55
 Curie temperature, 296
 Current sensors, 11, 58–59
 electrical, 18–19
 peculiarities of, 58–59
 Current transformers, 58–59

D

Data collection modes, 204
 Data fusion, 310
 Data integration modes, 204–205
 Deadline miss ratio, 226
 Delay-lines, 300
 Detection methods, comparison of, 402
 DI/dt sensors, 59
 Dielectric characterization of biological tissues, 75–77
 anisotropic medium, measurements on, 84–85
 electrodes configuration, 86
 bioimpedance

- sensor, measurement constraints related to, 78–79
 - spectroscopy, metrological aspects of, 77–78
 - biological tissue sample, measurement constraints related to, 79
 - bone anisotropy measurements at low frequency, 86–87
 - ex vivo* results, 80, 81
 - discussion, 81–82
 - modelization approach, 82
 - permittivity and conductivity, relation between, 86
 - simulation, 83–84
 - human blood, measurements on, 80
 - Differential equation, 63–64
 - Digitalization, magnetic sensors, 58
 - Direct detection and background noise, 401–402
 - Discrete YUV LUT, 344–345
 - color space, 347–348
 - transformation, 347
 - populating discrete YUV, 345
 - testing color class membership, 345–346
 - Distributed TCP caching (DTC), 233
 - 3D map
 - construction, 274
 - finalizing and visualizing, 274–275
 - 3D navigation viewer (3DNV), 278
 - system architecture of, 278–279
 - application interface module, 279–280
 - 3D world module, 280
 - 3DNV, 270, 276
 - Downstream, 223
 - Downstream reliability, 223–224
 - D-shaped optical fibers, 123–128
 - DSR, *see* Dynamic source routing (DSR)
 - DTC, *see* Distributed TCP caching (DTC)
 - Dual-tree complex wavelet transform (DT-CWT), 320
 - 3D world module, 280
 - Dynamic source routing (DSR), 203
 - Dynamic VRML-based navigable 3D map, 269–270
 - 3D navigation viewer (3DNV), 278
 - system architecture of, 278–280
 - indoor location-aware systems
 - cricket-based location-aware system, 270
 - RSSI-based location-aware system, 271–272
 - modeling approach, 272–273
 - acquiring data, 272–274
 - constructing 3D map, 274
 - finalizing and visualizing 3D map, 274–275
 - system implementation and evaluation, 281–284
 - visibility computations and determination, 274
 - portal culling algorithm, 276–278
- E**
- Earth's flux density, 19
 - EIS, *see* Electrical impedance spectroscopy (EIS)
 - Elastic waves, 288
 - Electrical current sensor, 18–19
 - Electrical impedance spectroscopy (EIS), 76, 77
 - Electromagnetic acoustic transducer (EMAT), 296–297
 - Electromagnetic dosimetry, 88
 - Electromagnetic field sensors
 - modelling of, 61–73
 - cell shape and size of, 71–72
 - convergence and execution consideration for, 72–73
 - differential equation methods for, 63–64
 - functions of, 69–71
 - integral equation methods, 64–65
 - integral equation solution, example of, 65–69
 - plane wave propagation, 62–63
 - proximity compensation for, 73
 - Electronic compass, 11, 19–20
 - Elliot's affective reasoner, 364
 - Elliott's affective reasoner, 364
 - EMAT, *see* Electromagnetic acoustic transducer (EMAT)
 - EM bio effects, 75–76
 - Emotional states, 361
 - Emotion-inducing factors, 367–368
 - End-to-end* reliability, 223–224, 228, 230
 - Energy loss*, 227
 - En-route*, 223–224
 - Environmental WSN applications, 216–217
 - ESRT, *see* Event-to-Sink Reliable Transport (ESRT)
 - Essential nodes (E-nodes), 236
 - Evanescent wave sensors, 122–123
 - coated FBGs based on chemo-mechanical and chemo-thermal effects, 146
 - micro-structured FBGs (MSFBG), 139–145
 - thinned FBGs (ThFBGs), 128–137
 - tilted FBGs, 137–139

written in D-shaped optical fibers, 123–128
Event reliability, 223–224, 226
 Event-to-Sink Reliable Transport (ESRT),
 230–231
 Experimental hardware setup, 333–334
 Extra urban driving cycle (EUDC), 185

F

Fast access color look-up-table (LUT)
 inspecting, 343–344
 membership testing, 342
 one-dimensional, 342
 posting, 342–343
 RGB color space, limitation, 340–341
 YUV thresholds, definition, 341
 Fast image capture and vision processing for
 robotic applications, 329–331
 colour segmentation, area thresholding,
 blob merging, 334–337
 discrete YUV LUT, 344–345
 color space, 347–348
 color space transformation, 347
 populating discrete YUV, 345
 testing color class membership,
 345–346
 experimental hardware setup, 333–334
 fast access color look-up-table (LUT)
 inspecting, 343–344
 membership testing, 342
 one-dimensional, 342
 posting, 342–343
 RGB color space, limitation, 340–341
 YUV thresholds, definition, 341
 full tracking vs. incremental
 tracking, 338–339
 adaptive tracking window size, 340
 global vision
 inherent sensor noise, 333
 odd and even scan fields of interlaced
 bit mapped image, 331–332
 variation of light intensity, 332–333
 interrupt based multi-buffered image
 capture, 337
 FBG, *see* Fiber Bragg gratings (FBG)
 Feedback compensation, 50, 53, 54
 FG, *see* Fiber grating (FG)
 FG based 3D vision sensor system and its
 hardware configuration, 254–256
 FG-CCD sensor installation, spot pattern
 projection and effective visible area,
 258–259
 Fiber Bragg gratings (FBG), 113–114
 evanescent wave sensors, 122–123

coated FBGs based on chemo-
 mechanical and chemo-thermal
 effects, 146
 micro-structured FBGs (MSFBG),
 139–145
 thinned FBGs (ThFBGs), 128–137
 tilted FBGs, 137–139
 written in D-shaped optical fibers,
 123–128
 history, 114–119
 perspectives and challenges, 146–147
 as sensors, 119–122
 technology evolution, 115
 Fiber grating (FG), 252, 257
 and operational principles, 252–254
 Fibre optic humidity sensors
 (FOHS), 154, 155
 Fidelity ratio, 227
 Field rotation vectors, 37
 FLAME model, 364–365
 Flux density, 5
 Fluxgate, 54–55, 58
 current sensors, 59
 sensors, 47–48
 low offset of, 48
 FOHS, *see* Fibre optic humidity sensors
 (FOHS)
 Forty-eight pole pair magnet, 35
Fragmentation/reassembly, 231
 Fredholm integral equation, 64
 Frequency dispersion, 291
 Full tracking vs. incremental tracking, 338–339
 adaptive tracking window size, 340
Funneling effect, 225
 Fused image using fusion metrics,
 reconstruction of, 317
 FUSION, 236
 Fusion method using ICA
 fused image using fusion metrics,
 reconstruction of, 317
 region-based fusion of ICA coefficients,
 315–316
 separated training sets, 314–315
 Fusion metrics, 312
 Petrovic metric, 313–314
 Piella metric, 312–313
 Fusion rule, 312

G

Galerkin method, 70
 Galvano-magnetic sensor, 3
 Gap, 38
 GARUDA, 232–233
 Geometrical correction factor, 25

- Global vision
 - inherent sensor noise, 333
 - odd and even scan fields of interlaced bit mapped image, 331–332
 - variation of light intensity, 332–333
- GMI sensors, 48–49
 - offset stability, 49
- Gratch's Émile, 365–366
- Green's function, 63
- Guided waves, 290

- H**
- Hall current sensors, 59
- Hall effect, 3–4
- Hall-effect sensor, 23, 24
- Hall element, 3–4
- Hall plate in CMOS technology, 4
- Hall sensors, 3–4, 46
 - design, 24–27
 - of magnetic sensors, 46
 - Semiconductor materials of, 46
 - see also* CMOS Hall sensors
- Helmholtz coil, 33
- Heterogeneity, 211, 212, 213
- High refractive index (HRI), 125
- Hilbert space, 65
- Hollow core fibres, 166–169
- Home control/automation, 211–212
- Homogeneous sensor nodes, 205
- Hop-by-hop* reliability, 223–224, 228
- HP4191A, 80
- Human blood and dielectric
 - characterization, 80
- Humidity monitoring, 153
- Humidity sensitive nano-films, 154
- Humidity sensors, optical fibre, using
 - nano-films, 153–154
 - optical techniques, 158
 - reflective sensors, 158–162
 - transmissive evanescent wave sensors, 162–172
 - sensing materials and deposition techniques, 155–156
 - nanostructured films, 156–158
- Hydrophilic gels, 155
- Hysteresis, 10, 13–14
 - curve, 9–10

- I**
- ICA, *see* Independent component analysis (ICA)
- IEEE 802.15.4, 211
- IEEE 802.15.1/Bluetooth, 211
- IEEE 802.11 series/WiFi, 211
- IEEE 802.15.3/WiMax, 211
- IF, *see* Intermediate frequency (IF)
- Image analysis using ICA, 311–312
- Image fusion, 310, 311, 312, 313, 317
 - metrics, 313
 - standard ICA image fusion method, comparison with, 318
 - state-of-the-art image fusion methods, comparison with, 318–324
- Image patch, 311
- Imaging using focused sensors, 301–302
- IMC, *see* Integrated magnetic concentrator (IMC)
- IMC 3-axis sensor, 15
- IMC photolithographic postprocess, 9
- Incoherent (direct) detectors, 405–408
- Incremental tracking, 330
 - algorithm, 330
 - vs.* full tracking, 338–340
- Independent component analysis (ICA), 311
- Indoor location-aware systems
 - cricket-based location-aware system, 270
 - RSSI-based location-aware system, 271–272
- Industrial automation, 213–214
- In-fibre Bragg gratings, 193
- INSIGHT, *see* INternet Sensor InteGration for HabitaT monitoring (INSIGHT)
- Integral equation, 61, 64–69
 - methods, 64–65
 - solution, example of, 65–69
- Integrated magnetic concentrator (IMC), 6–8
 - layer, properties of, 9
 - material properties, 9–10
 - geometric, 11–12
 - process, 8–9
- Interdigital sensors, 92
- Interface echo, 302
- Interface module, application, 279–280
- Intermediate frequency (IF), 399
- Internal temperature distribution, 303–304
- INternet-Sensor InteGration for HabitaT monitoring (INSIGHT), 216
- Interrupt based multi-buffered image capture, 337
- Isotropic portion, 32–33

- J**
- Joystick sensor, 18

- K**
- Kinetic inductance effect, 407
- Krypton-fluoride (KrF) excimer lasers, 117

L

- Lamb waves, 290
- LAP, *see* Location address protocol (LAP)
- Laplace transform (LT), 320
- Laser-ultrasonics, 296
- Layer-by-layer (LBL) self-assembly, 154, 156
- LEACH, *see* Low-energy adaptive clustering hierarchy (LEACH)
- Linear position sensor, 17
- Location address protocol (LAP), 242
- Longitudinal waves, 290
- Long-period fiber gratings (LPFG), 123, 170, 172
- Long period fibre bragg grating, 170–172
- Long Scan, 262–263
- Lorentz equation, 24
- Loss aggregation, 232
- Low-energy adaptive clustering hierarchy (LEACH), 209–210
- Low-rate wireless communication technology, 200–202
- LPFG, *see* Long-period fiber gratings (LPFG)

M

- Magnetic contactless linear position sensor, 17
- Magnetic portion, 32–33
- Magnetic sensors, accuracy
 - crossfield effect of, 51–55
 - AMR, 51–54
 - fluxgate, 54–55
 - current sensors, peculiarities of, 58–59
 - digitalization, remarks on, 58
 - linearity, 49–51
 - measurement for, 50–51
 - noise, 55–57
 - noise value measurement, 56–57
 - perming effect, 57–58
 - position sensors, peculiarities of, 59
 - temperature stability of, 45–49
 - fluxgate sensors, 47–48
- Magnetization, 16
- Magnetoresistors, 47
- Magnetostrictive wire type displacement sensor, 379–380
 - displacement error, measurement of
 - measuring method, 381
 - result, 382–383
 - operation principle of, 380–381
 - reduction of displacement error, consideration on, 383–384
- ManageCell*, 277
- Managing social regulation, 362
- Many-to-one data flow, 223
- Markovian emotion model, 366–370

- Amask, 117
- Mass flow controller (MFC), 184
- Maximum transmission unit (MTU), 231
- Maxwell-Fricke model, 82, 83–84
- Maxwell's equations, 62, 63
- Membership testing, 342
- MEMS, *see* Micro-electromechanical systems (MEMS)
- MFC, *see* Mass flow controller (MFC)
- MHM, 166, 167
- Microelectrodes, 78, 88
- Micro-electromechanical systems (MEMS), 297
- Micro-structured FBGs (MSFBG), 139–145
- Microstructured optical fibers (MOF), 147
- Microwave Instrument for Rosetta (MIRO), 393
- Microwave Limb Sounder (MLS) instrument, 396
- Microwave limb sounding, 396
- Military sensing networks, 214–216
- MIRO, *see* Microwave Instrument for Rosetta (MIRO)
- Mixer, 399
- MMF, *see* Multimode fibre (MMF)
- Mode conversion, 293
- Mode of propagation, types of, 290–291
- MOF, *see* Microstructured optical fibers (MOF)
- Mohr's circle, 33
- Molar absorption coefficient, 181–182
- Motivation and learning, 362–363
- Motivation based strategy, 357
- MTU, *see* Maximum transmission unit (MTU)
- MULE, 208
- Multianalyte detection, 147
- Multicasting, 223
- Multi-hop planar networks, 206
- Multimode fibre (MMF), 166
- Multi-robot system, 356
- Multi-robot team (MRT), 356–357
 - computing, sensing, expressing and synthesis, 371–372
 - development of affection based, 363–370
 - emotional roles in, 362–363
 - emotional states, 361
 - emotion and emotional intelligence, 359–361
 - taxonomy of, 357–359
- Multi-tier wireless sensor network, 205
- MURDOCH, 357

N

- NACK-based protocol, 231, 232
- '2-neighbour' algorithm, 336

- '4-neighbour' algorithm, 336
- Network efficiency, 226
- Network fairness, 226
- Network lifetime, 227
- Network topology, 202–203
- Node efficiency, 226
- Noise, 55–57
 - measurement, 56–57
 - reduction, 56
- Non-contact techniques, 296–297
- Nondestructive evaluation
 - buffer rod sensors for high temperature monitoring, 300–301
 - imaging using focused sensors, 301–302
 - in-situ* monitoring of solid-liquid interface, 302–303
 - internal temperature distribution, monitoring of, 303–304
- Non-destructive testing (NDT) method, 92

- O**
- Object Manager, 278
- One-dimensional wave equation, 62
- One-to-many data flow, 223
- Optical arrangement and working principles, 256–258
- Optical fibre humidity sensors using
 - nano-films, 153–154
 - humidity sensing materials and deposition techniques, 155–156
 - nanostructured films, 156–158
 - optical techniques, 158
 - reflective sensors, 158–162
 - transmissive evanescent wave sensors, 162–172
- Optical spectrum analyzer (OSA), 170
- OPTO-EMI-SENSE Project, 180–181
 - experimental results
 - gas measurement in mid infra red range, 183–187
 - gas measurement in ultra violet range, 187–193
 - optical fibre temperature measurement, 193–195
 - theoretical background, 181–183
- OSA, *see* Optical spectrum analyzer (OSA)
- Output voltage, 3
- Ozone hole, 395

- P**
- Packet latency*, 227
- Packet loss ratio, 226–227
- Packet reliability*, 223–224
- Parallel transmission line (PTL), 65
- ParleE, 366
- PCCP, *See* Priority-based Congestion Control Protocol (PCCP)
- PCS, *see* Plastic cladding fibres (PCS)
- PDMS, *see* Polydimethylsiloxane (PDMS)
- Per-child packet queues, 228
- Per-child subtree size, 238
- Perming effect, 57–58
- Petrovic metric, 313–314
- Phase mask, 117, 118
- Photometry, 398
- Photon and quasi-particles, 407
- Photosensitivity, 115
- Piella metric, 312–313
- Piezoelectric transducers, 295
- Pitch-catch configuration, 299
- Planar interdigital sensor, 108
- Planar wireless sensor network, 205–206
- Plane wave propagation, modelling of, 62–63
- Plastic cladding fibres (PCS), 162
- Plate waves, 290
- POF, *see* Polymer optical fibers (POF)
- Poisson's ratio, 291
- Polyanion-polycation multilayer, 157
- Polydimethylsiloxane (PDMS), 128
- Polymer optical fibers (POF), 147
- Poly methyl methacrylate (PMMA) tube, 129
- Populating discrete YUV, 345
- Portal culling algorithm, 276–278
- Position sensors, 59
- Position sensors, peculiarities of, 59
- Priority-based Congestion Control Protocol (PCCP), 238–239
- Propagation distance of elastic wave, 381
- Proton magnetometer, 49
- PSFQ, *See* Pump Slowly Fetch Quickly (PSFQ)
- PTL, *see* Parallel transmission line (PTL)
- Pump Slowly Fetch Quickly (PSFQ), 232
- PVDF (polyvinylidene fluoride), 295

- Q**
- Quantum limit, 401, 402
 - for coherent detection, 400–401
- Query-metadata, 232

- R**
- RAP, 241–243
- Rayleigh wave, 290
- RBC, *see* Reliable Bursty Convergecast (RBC)
- Real-time processing, 261–263
- Real time tracking and monitoring of human behavior, 250–251

- FG based 3D vision sensor system and its hardware configuration, 254–256
 - FG-CCD sensor installation, spot pattern projection and effective visible area, 258–259
 - fiber grating and its operational principles, 252–254
 - motivation and objectives, 251–252
 - operational software development application, 263–266
 - real-time processing, 261–263
 - reference spots frame generation, 259–261
 - optical arrangement and working principles, 256–258
 - Received signal strength indication (RSSI), 271
 - Redundancy*, 227
 - Reference forward model (RFM), 182
 - Reference node, 272
 - Reference spots
 - frame data file 261
 - frame generation, 259–261
 - Reflection and transmission, 292–293
 - Reflection coefficient, 292
 - Refraction and mode conversion, 293–294
 - Region-based fusion of ICA coefficients, 315–316
 - Relative humidity (RH), 155
 - Reliability semantics, 223–224
 - Reliable Bursty Convergecast (RBC), 240–241
 - Reliable Multi-Segment Transport (RMST), 231
 - Residual energy, 227
 - Resistance equation, 52
 - RGB color space, limitation, 340–341
 - RLE, *see* Run length encoding (RLE)
 - RMST, *see* Reliable Multi-Segment Transport (RMST)
 - RMST entity, 231
 - Rogowski coils, 59
 - Rooftop function, 71
 - Route-thru*, 223–224
 - Routing protocols, 203
 - RSS, 272
 - RSSI, *see* Received signal strength indication (RSSI)
 - RSSI-based location-aware system, 271–272
 - Run length encoding (RLE), 330
- S**
- Saturation, 12–13
 - Saturation magnetization, 10
 - SAW, *see* Surface acoustic waves (SAW)
 - Screw lash, 33
 - Self-term singularities ($1/r$), 64
 - SensEye, 208
 - Sensitivity
 - of coherent (heterodyne) and incoherent (direct) detection, 400–402
 - direction information, 5
 - Sensor(s)
 - architecture, 14–15
 - interrogation, 130
 - noise, 45
 - see also specific sensors*
 - Sensor TCP (STCP), 234–235
 - Sensor-to-sink data flow, 223
 - Sensory integration, 362
 - SenTCP, 235
 - Sequential component labeling algorithm, 334–335
 - Shear waves, 290
 - Sheep skin, tanning process of, 91–92, 96–97
 - embedded controller based sensing system, 101–103
 - experimental setup, 97–98
 - results, 98–101, 103–108
 - interdigital sensors, operating principle of, 93–96
 - motivation, 92
 - Short (fast) scan, 361–362
 - Silicon micromachined circuits, 409
 - SIN, *see* Superconductor insulator normal metal (SIN)
 - Sink, 222
 - Sink-received throughput, 226
 - Sink-to-sensor data flow, 223
 - Siphon, 240
 - Smart lighting, 204
 - Sobel edge operator, 313
 - Source nodes, 205–206
 - Specific acoustic impedance, 289
 - Spin-Dependent Tunneling (SDT)
 - magnetoresistors, 47
 - Spinning current operation, 14–15
 - SPR, *see* Surface plasmon resonance (SPR)
 - SRI, *see* Surrounding-medium refractive index (SRI)
 - Standard ICA image fusion method, 318
 - Star topology network, 202–203
 - State-of-the-art image fusion methods, 318–324
 - STCP, *see* Sensor TCP (STCP)
 - Submillimeter-wave coherent and incoherent sensors, 387–390
 - applications
 - astronomy and astrophysics, 390–392
 - earth sciences, 394–396

- planetary sciences, 393–394
 - future trends, 408–409
 - submillimeter-wave sensors, 398–400
 - coherent (heterodyne) sensors, 403–405
 - incoherent (direct) detectors, 405–408
 - sensitivity of coherent (heterodyne) and incoherent (direct) detection, 400–402
 - Submillimeter-wave sensors
 - coherent (heterodyne) sensors, 403–405
 - incoherent (direct) detectors, 405–408
 - sensitivity of coherent (heterodyne) and incoherent (direct) detection, 400–402
 - Superconducting transition edge sensors (TES)
 - thermistors, 406
 - Superconductor insulator normal metal (SIN), 408
 - Supporting dynamic routing, 214–217
 - Supporting static routing, 211–214
 - Surface acoustic waves (SAW), 290
 - Surface plasmon resonance (SPR), 126
 - Surrounding-medium refractive index (SRI), 123, 125
- T**
- Tanning, 92
 - Tanning process of sheep skin, 91–92
 - embedded controller based sensing system, 101–103
 - experimental setup, 97–98
 - results, 98–101, 103–108
 - interdigital sensors, operating principle of, 93–96
 - motivation, 92
 - 8 step process, 96–97
 - Tapered optical fibres, 163–166
 - TCP, *see* Transport control protocol (TCP)
 - Team agreement, 357–358
 - Temcos, 46
 - Temperature dependence of sensitivity, 46
 - Temperature sensor, 193–194
 - Temperature stability, 59
 - magnetic sensors, 45–49
 - fluxgate sensors, 47–48
 - GMI sensors, 48–49
 - Hall sensors, 46
 - magnetoresistors, 47
 - proton magnetometer, 49
 - of magnetic sensors, 45–49
 - Testing color class membership, 342–344
 - Theoretical signal propagation, 272
 - Thin-film InSb Hall sensors, 46
 - Thinned FBGs (ThFBGs), 128, 136
 - Three-tier sensor network architectures, 207–208
 - Tilted FBGs, 137–139
 - Tiny TCP/IP, 233–234
 - Topology, 202
 - Traffic monitoring system, 208
 - Traffic semantics, 223
 - Transducer element, 163
 - Transducers, 294–296
 - Transit* traffic, 223, 224
 - Transmission coefficient, 292
 - Transport control protocol (TCP), 227, 228
 - Transverse holographic method, 117
 - Trickle, 235
 - Tunneling magnetoresistors (TMR), 47
 - Two-tiered sensor network architectures, 206–207
- U**
- Ultrasonic sensing, 287–288
 - nondestructive evaluation, applications to
 - buffer rod sensors for high temperature monitoring, 300–301
 - imaging using focused sensors, 301–302
 - in-situ* monitoring of solid-liquid interface, 302–303
 - internal temperature distribution, monitoring of, 302–304
 - ultrasound, fundamentals of
 - features of, 289–294
 - in media, 288–289
 - ultrasound, measurement of
 - generation and detection of, 294–297
 - instrumentation, basics of, 298–299
 - Ultrasonic transducers, limitation of, 296
 - Ultrasonic waves
 - features of
 - attenuation, 291–292
 - reflection and transmission, 292–293
 - refraction and mode conversion, 293–294
 - types of (mode of propagation), 290–291
 - velocity, 291
 - wavelength, 292
 - generation and detection of
 - non-contact techniques, 296–297
 - transducers, 294–296
 - Upstream, 223
 - Upstream* reliability, 223–224
 - Upstream sensory data traffic, 223

V

- Velocity, 291
- Visible volume area (VVA), 278
- Vision systems, 329
- VOC, *see* Volatile organic compounds (VOC)
- Voigt- Thompson equation, 31
- Volatile organic compounds (VOC), 128
- VRML, 270, 274
- VVA, *see* Visible volume area (VVA)

W

- Wait-for First Packet (WFP), 233
- Wave equations for isotropic solid, 291
- Waveguides, 300
- Wavelength, 292
- Wavelength consideration of, 72
- Wi-Fi characteristics, 201, 202
- Wireless FieldBus, 212, 213
- Wireless industrial Ethernet, 213
- Wireless sensor networks (WSN), 199–200
 - applications, 210
 - supporting dynamic routing, 214–217
 - supporting static routing, 211–214
 - architectures, 205
 - cluster-based, 208–209
 - tier-based, 205–208
 - fundamentals
 - data integration modes, 204–205
 - network topology, 202–203
 - routing protocols, 203
 - low-rate wireless communication
 - technology, 200–202

WSN, *see* Wireless sensor networks (WSN)

- WSN transport layers, 222
 - congestion detection, mitigation, and control, 224–225
 - loss recovery, 224
 - performance metrics, 225–228
 - protocols, 228–229
 - ART, 236–237
 - CCF, 237–238
 - CODA, 229–230
 - ESRT, 230–231
 - FUSION, 236
 - GARUDA, 232–233
 - PCCP, 239
 - PSFQ, 232
 - RAP, 241–243
 - RBC, 240–241
 - RMST, 231
 - SenTCP, 235
 - Siphon, 240
 - STCP, 234–235
 - Tiny TCP/IP, 233–234
 - Trickle, 235
 - reliability semantics, 223–224
 - traffic semantics, 223

Y

YUV thresholds, definition, 341

Z

- Zigbee, 201–202, 211, 213
 - home automation systems, 212