Youn-Long Lin · Chong-Min Kyung
Hiroto Yasuura · Yongpan Liu *Editors*

Smart Sensors and Systems

Springer

Smart Sensors and Systems

Youn-Long Lin • Chong-Min Kyung
Hiroto Yasuura • Yongpan Liu
Editors

Smart Sensors and Systems

Springer

Editors

Youn-Long Lin
National Tsing Hua University
Hsichu, Taiwan

Hiroto Yasuura
Kyushu University
Fukuoka, Japan

Chong-Min Kyung
Korea Advanced Institute of Science
and Technology
Daejeon, Korea

Yongpan Liu
Tsinghua University
Beijing, China

ISBN 978-3-319-14710-9 ISBN 978-3-319-14711-6 (eBook)
DOI 10.1007/978-3-319-14711-6

Library of Congress Control Number: 2015938153

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Ever advancement of semiconductor processing, information and communications technology has made possible proliferation of electronics devices and systems into our daily life. Over the past decade, smart phones have become mobile terminals of the Internet. In the future, everything will be connected to the Internet for more intelligent use and management. Intelligent electronics devices will be implanted into human body as well as embedded in automobiles and appliances, plants, animals, and manufactured goods. Hence, both the academy and industry are addressing issues facing the Internet of Things (IoT) era.

One of the essential issues in the IoT era is sensing. A sensor measures its ambient parameters such as temperature, pressure, motion, acceleration, position, pH value, and gas. It converts various measurements into electronics signal for later processing, storage, and transmission. Its design constraints are more stringent. It calls for interdisciplinary research because it interacts with photonic, mechanical, chemical, biological, and medical interfaces.

In recognition of the growing significance of smart sensors, the editors started a workshop for sharing information and views on the status/direction of R&D, societal acceptance/impact, and promising business models on smart sensors among four countries in Asia. The first one, 1st Asian workshop on smart sensor systems, was held in Jeju, Korea, on March 28–30, 2013, and the second one was held on March 20–22, 2014, in Hualien, Taiwan, where the editors agreed to publish selected presentations from the last two workshops. It is strongly believed that smart sensor business will become very important in the IT industry and future society. The R&D in Asian academia is expected to strongly reflect and meet this societal demand. It seems very timely and fortunate that we were able to solicit and publish in this volume selected research works in Asia. This book consists of 18 chapters grouped into six parts: (1) Biochemical sensing mechanism and devices, (2) Imaging, photography, and video analytics, (3) Gas and odor sensing, (4) Energy harvesting, (5) SoCs and equipment for biomedical sensing, (6) Deployment and service of smart sensors in the society.

Part I: Biochemical Sensing Mechanism and Devices

In the first chapter, Young June Park et al. contribute "C-chip: Platform for Electrical Biomolecular Sensors." They present a new CMOS platform for real-time, multiple molecules detection of biomolecules. In the second chapter, Junghoon Lee and Jun-kyu Choi contribute "Chemomechanical Transduction Systems: A Sensing Platform by Surface Force Measurement." They show how to transduce a chemical binding event into a mechanical deformation and demonstrate several applications. In the third chapter, Xiaojun Guo et al. present a "Fully Printable Organic Transistor Technology for Sensor Transducer."

Part II: Imaging, Photography, and Video Analytics

The fourth chapter entitled "The Three-Dimensional Evolution of Hyperspectral Imaging" by Min H. Kim provides a brief overview on the foundations of hyperspectral imaging and introduces advanced applications of hyperspectral 3D imaging. It surveys the fundamentals of optics and calibration processes of hyperspectral imaging and then studies two typical designs of hyperspectral imaging. In the fifth chapter, Hajime Nagahara and Rin-ichiro Taniguchi contributed "Computational Photography Using Programmable Aperture." They propose a programmable aperture camera using a Liquid Crystal on Silicon (LCoS) device. Sixth chapter is entitled "Exploratory Visual Analytics for Winter Road Management Using Statistically Preprocessed Probe-Car Data." Yuzuru Tanaka et al. describe a winter road management system deployed in Sapporo, snow capital of Japan. They show that statistically preprocessed probe car data over road links enables visualization of the dynamic change of the traffic flow in terms of the divergence and flow vector field, hotspots of traffic, main traffic streams, and route selection preference.

Part III: Gas and Odor Sensing

In the seventh chapter, Ho Won Jang et al. contribute "Novel Metal Oxide Gas Sensors for Mobile Devices." They discuss critical issues for the realization of miniaturized and integrated chemoresistive thin film sensors based on metal oxides and introduce notable recent achievements. Shih-Wen Chiu et al. present a "Handheld Gas Sensing System" in the eighth chapter. Based on an array of surface acoustic wave (SAW) gas sensors, they realize a bioinspired gas sensing system (also called electronic nose) to construct a robust system to identify gases. They introduce nanocomposites of polymers and ordered mesoporous carbons (OMCs), grew polymers directly on the carbon material through a radical polymerization process, thus forming interpenetrating and inseparable composite frameworks with carbon, and developed several odor classification algorithms to perform gas classification. In the ninth chapter, Chuanjun Liu and Kenshi Hayashi contributed "Odor Sensing Technologies for Visualization of Odor Quality and Space." Their biological-inspired odor sensing technologies based on various molecular recognition technologies, such as partial structure recognized water membrane/Pt electrodes, benzene-patterned self-assembled monolayer (SAM) layers, size and polarity selected molecular sieve materials, and molecularly imprinted polymer (MIP) adsorbents, construct an artificial odor map.

Part IV: Energy Harvesting

Sehwan Kim and Pai H. Chou contributed "Energy Harvesting with Supercapacitor-Based Energy Storage" in the tenth chapter. They first review ambient energy sources and their energy transducers for harvesting, followed by descriptions of harvesters with low-overhead efficient charging circuitry and supercapacitor-based storage. In the eleventh chapter, Yongpan Liu et al. describe "Power System Design and Task Scheduling for Photovoltaic Energy Harvesting Based Nonvolatile Sensor Node."

Part V: SoCs and Equipment for Biomedical Sensing

The twelfth chapter is titled "Basic Principle and Practical Implementation of Near-Infrared Spectroscopy (NIRS)." This chapter provides a brief overview of the basic principle of NIRS, the imaging technique of diffuse optical tomography, and the superficial noise reduction method. Then, it introduces three modulation methods for realizing the multichannel CW NIRS and illustrated the implementation of spread-spectrum-code-based CW NIRS. In the thirteenth chapter, Shey-Shi Lu and Hsiao-Chin Chen contribute "Wireless CMOS Bio-medical SoCs for DNA/Protein/Glucose Sensing." They present the design concepts of cantilever-based DNA sensors, poly-silicon nanowire-based protein/DNA sensors, a hydrogel-based glucose sensor, an ISFET-based pH sensor, and a bandgap-reference-based temperature sensor. They also described the fabrication processes and sensor interface readout circuits that can deal with voltage, current, capacitive, and resistive sensing signals. Xiaoyang Zhang et al. contribute the fourteenth chapter entitled "Design of Ultra-Low-Power Electrocardiography Sensors." In the fifteenth chapter, Chen-Yi Lee et al. present "A Sensor-Fusion Solution for Mobile Health-Care Applications." They introduce a sensor-fusion approach towards an energy-efficient and data-reliable solution. By exploiting event-driven architecture, energy efficiency can be enhanced, and analysis accuracy improved with the support of multi-data sets.

Part VI: Deployment and Service of Smart Sensors in the Society

In the sixteenth chapter, Wen-Tsuen Chen et al. contribute "An IoT Browsing System with Learning Capability." They propose a novel IoT browsing system with a learning capability middleware for IoT browser integrated with context-aware services. The system adopts a service-oriented architecture and provides device interoperability, resource reusability, and spatial-temporal awareness. With the help of the IoT browsing system, heterogeneous devices can cooperate with each other to provide IoT services in accordance with context inference results. In the seventeenth chapter, Rin-Ichiro Taniguchi present "Toward Social Service Based on Cyber Physical Systems." Ashir Ahmed et al. contribute the eighteenth chapter "Portable Health Clinic: A Tele-Healthcare System for UnReached Communities." They have prototyped "portable health clinic (PHC)," a portable health clinic box with diagnostic equipment and a software tool, "GramHealth," for archiving and searching patients' past health records. Doctors at the medical call center access GramHealth data cloud through the Internet or have a copy of the database in the call center server. Upon receiving a multimedia call from a patient, the doctor can find that patient's previous EHR record and then create and send an e-Prescription.

The Asian Workshop on Smart Sensor Systems will be continuously held annually. We expect to publish more volumes in the future. Comments or questions can be sent to ylin@cs.nthu.edu.tw. We would be happy to hear from you.

Hsichu, Taiwan Youn-Long Lin
Daejeon, Korea Chong-Min Kyung
Fukuoka, Japan Hiroto Yasuura
Beijing, China Yongpan Liu

Contents

Part I Biochemical Sensing Mechanism and Devices

"C-chip" Platform for Electrical Biomolecular Sensors 3
Young June Park, Jinhong Ahn, Jaeheung Lim, and Seok Hyang Kim

Chemomechanical Transduction Systems: A Sensing Platform by Surface Force Measurement 25
Junghoon Lee and Jun-kyu Choi

Fully Printable Organic Thin-Film Transistor Technology for Sensor Transducer 47
Xiaojun Guo, Linrun Feng, Wei Tang, Cheng Jiang, Jiaqing Zhao, and Wenjiang Liu

Part II Imaging, Photography, and Video Analytics

The Three-Dimensional Evolution of Hyperspectral Imaging 63
Min H. Kim

Computational Photography Using Programmable Aperture 85
Hajime Nagahara and Rin-ichiro Taniguchi

Exploratory Visual Analytics for Winter Road Management Using Statistically Preprocessed Probe-Car Data 109
Yuzuru Tanaka, Hajime Imura, and Jonas Sjöbergh

Part III Gas and Odor Sensing

Novel Metal Oxide Gas Sensors for Mobile Devices 131
Ho Won Jang, You Rim Choi, and Yeon Hoo Kim

Handheld Gas Sensing System 155
Shih-Wen Chiu, Hsu-Chao Hao, Chia-Min Yang, Da-Jeng Yao, and Kea-Tiong Tang

Odor Sensing Technologies for Visualization of Odor Quality and Space 191
Chuanjun Liu and Kenshi Hayashi

Part IV Energy Harvesting

Energy Harvesting with Supercapacitor-Based Energy Storage 215
Sehwan Kim and Pai H. Chou

Power System Design and Task Scheduling for Photovoltaic Energy Harvesting Based Nonvolatile Sensor Nodes 243
Yongpan Liu, Huazhong Yang, Yiqun Wang, Cong Wang, Xiao Sheng, Shuangchen Li, Daming Zhang, and Yinan Sun

Part V SoCs and Equipment for Biomedical Sensing

Basic Principle and Practical Implementation of Near-Infrared Spectroscopy (NIRS) 281
Hyeonmin Bae

Wireless CMOS Bio-medical SoCs for DNA/Protein/Glucose Sensing 303
Shey-Shi Lu and Hsiao-Chin Chen

Design of Ultra-Low-Power Electrocardiography Sensors 359
Xiaoyang Zhang, Yongfu Li, Lei Wang, Wei Zou, Yinan Sun, Yongpan Liu, Huazhong Yang, Yong Lian, and Bo Zhao

A Sensor-Fusion Solution for Mobile Health-Care Applications 387
Chen-Yi Lee, Kelvin Yi-Tse Lai, and Shu-Yu Hsu

Part VI Deployment and Service of Smart Sensors in the Society

An IoT Browsing System with Learning Capability 401
Wen-Tsuen Chen, Chih-Hang Wang, Yen-Ju Lai, and Po-Yu Chen

Toward Social Services Based on Cyber Physical Systems 427
Rin-ichiro Taniguchi, Kauzaki Murakami, Atsushi Shimada, Shigeru Takano, Akira Fukuda, and Hiroto Yasuura

Portable Health Clinic: A Telehealthcare System for UnReached Communities 447
Ashir Ahmed, Andrew Rebeiro-Hargrave, Yasunobu Nohara, Rafiqul Islam Maruf, Partha Pratim Ghosh, Naoki Nakashima, and Hiroto Yasuura

About the Editors

Youn-Long Lin is a Chair Professor in the Department of Computer Science, National Tsing Hua University, Taiwan. His research interests include High-Level Synthesis and Physical Design Automation of VLSI, SOC Design Methodology, Video Coding Architecture Design, and Software-Defined Networking. He coauthored the book *High-Level Synthesis: Introduction to Chip and System Design*. He has served on editorial boards of ACM TODAES and TECS. He is a cofounder of Global Unichip Corp. Professor Lin obtained his PhD in Computer Science from the University of Illinois, Urbana-Champaign, IL, USA, in 1987.

Chong-Min Kyung received his B.S. in EE from Seoul National University in 1975, and M.S. and Ph.D. in EE from KAIST in 1977 and 1981, respectively. Since 1983 he has been working at the KAIST on CAD, computer graphics, microprocessors, DSP cores, and SoCs. His current research interest is system-level optimization of smart sensors, especially 3-D smart cameras. In 2011 he founded and currently leads Center for Integrated Smart Sensors (CISS), a Global Frontier Project supported by Korean Government. He received Best Paper/Design Awards in numerous international conferences including ASP-DAC 1997 and 1998, DAC in 2000, ICSPAT in 1999, ICCD in 1999, and ISQED in 2014. He is a member of the National Academy of Engineering of Korea and Korean Academy of Science and Technology and is an IEEE fellow.

Hiroto Yasuura is an Executive Vice President of Kyushu University and in charge of Finance (CFO), Academia-Industry relationship, and Chief Information Officer (CIO) of Kyushu University. Prof. Yasuura received his B.E., M.E., and Ph.D. degrees in computer science from Kyoto University, Kyoto, Japan, in 1976, 1978, and 1983 respectively. Prof. Yasuura developed several EDA systems for VLSI and hardware algorithms in Kyoto University. In Kyushu University, Prof. Yasuura has conducted research projects on the system LSI design methodology. He also developed an educational microprocessor, KUE-CHIP2, and promoted education of VLSI design in computer science area in Japan with VDEC in the University of Tokyo. Prof. Yasuura served as Technical Program Chair and General Chair of

ICCAD in 1997 and 1998, respectively. He served as a Vice President of IEEE CAS Society, an ACM SIGDA advisory board member, and General Chair of ASP-DAC 2003. He is also the Steering Committee Chair of ASP-DAC and a fellow of IEICE and IPSJ.

Yongpan Liu is an associate professor in the Department of Electronic Engineering, Tsing Hua University. His research is supported by NSFC, 863, 973 Program and Industry Companies such as Intel, Rohm, Huawei, and so on. He has published over 50 peer-reviewed conference and journal papers and led over 6 SoC design projects for sensor applications and has received the ISLPED2012/2013 Design Contest Award and several Best Paper Candidates. He is an IEEE, ACM, and IEICE member and has been invited to serve on several conference technical program committees (DAC, ASP-DAC, ISLPED, ICCD, A-SSCC, etc.).

Part I
Biochemical Sensing Mechanism and Devices

“C-chip” Platform for Electrical Biomolecular Sensors

Young June Park, Jinhong Ahn, Jaeheung Lim, and Seok Hyang Kim

Abstract In this chapter, a new CMOS platform for the real time, multiple molecules detection of the bio molecules will be introduced. The semiconductor channel of the sensor device is composed of the carbon nanotube network (CNN) decorated with the gold nano particles (GNP) as the docking agents for the probe molecules and integrated onto the CMOS chip which receives the modulation of the channel resistance and performs the necessary signal processing. The number of the sensor devices on a chip is in the range of a few thousands and statistical analysis of the multiple sensing is possible in addition to multiple targeting by adding different probe molecules. In addition, the charge screening effects and non selectivity problem, which have been regarded as the show stoppers of the electrical sensing of the bio molecules can be widely avoided by introducing the electrical pulse technique. The architecture of the platform, salient feature of the sensor devices and the behavior of the sensor devices on the electrical pulse agitation will be reviewed.

1 Introduction; Affinity Based Electrical Biochip

1.1 General Introduction

Electrical biosensors have many advantages compared with the sensor devices based on other principles such as optical, mechanical method. One of the salient advantages is easy integration in the CMOS chip, thereby signal processing and communication through existing IT infra-structure are readily available, not to mention the label free detection. The (bio) chemical sensors [1, 2] used for electrical signal detection either detect the electric charges directly or detect the electric field caused by electric charges. The former type is called the “amperometric sensor,”

Y.J. Park (✉) • J. Lim • S.H. Kim
Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
e-mail: ypark@snu.ac.kr

J. Ahn
NANO Systems Institute, Seoul National University, Seoul, Korea

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_1

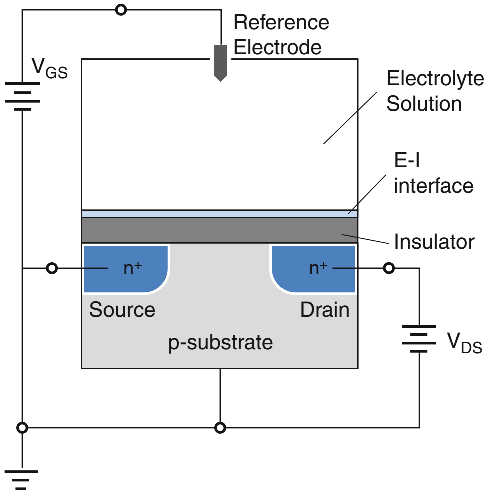


Fig. 1 EISFET structure with the n-type FET. The device can be divided into three important regions: the gate-electrolyte region, the FET region, and the EI interface region. The FET device is an n-channel device, and V_{GS} is applied to the electrode in the electrolyte with respect to the source voltage (Reproduced by permission of Pan Stanford Publishing [4])

and the latter is the "potentiometric sensor." The electric charges have different forms depending on the (bio) chemical reactions involved in the specific sensing. The charges can be electrons, protons (H^+), or those from the backbone of DNA (nucleotide) or proteins. In the amperometric sensor, electrons resulting from the (bio) chemical reactions are measured directly by a metal electrode in the form of the electrical current.

Even though amperometry is used for commercial glucose sensor, most of the electrical biosensor chip is based on the FET principle. The reason is that the FET device is easy to be integrated on to the CMOS chip where the electrical signal can be processed on the chip without the need of complicated electrical lines. The advantages of integration on to the CMOS chips are many: small size, more immunity to external noise, potential to multiple targeting of the biomolecules, to list a few.

MOST of the FET based biosensor have root in the electrolyte-insulator-semiconductor FET (EISFET) structure (Fig. 1), where the FET device is directly contacted to the electrolyte solution [3]. After the EISFET device was introduced, its applications have been extended from pH sensing to sensing of various biological molecules contained in the electrolyte. Underlying principle of the molecular sensing is the selective affinity between the probe molecules and the target molecules. As the target molecules have electrical charges in the electrolyte solution (buffer solution, serum, etc.), the electric charges are supposed to affect the nearby channel conductance through the field effect. Figure 2 shows the general setting of the affinity-based electrical sensor where the target molecules are contained in the electrolyte. It is noted that other molecules (called the noisy molecules here) are contained in the electrolyte system as well. Usually concentration of the noisy

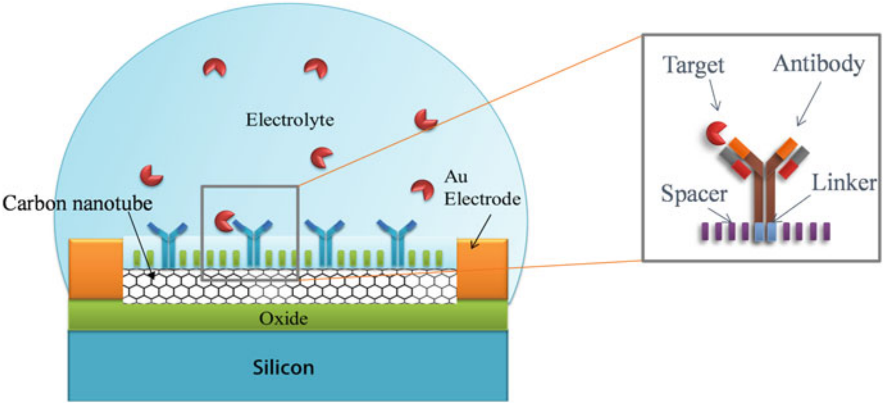


Fig. 2 A schematic diagram of the affinity-based electrical biosensor

molecules are much bigger than that of the target molecules. The binding between the noisy molecules and the target molecules, and the attachment of the molecules on the bare channel are the source of the noise. The noise associated with these 'non selective' binding events is one of the most serious problems in the affinity-based sensor. The challenges will be detailed in the next session.

One of the typical applications is DNA sensing, where single strand (ss) DNA, called the probe DNA, is immobilized onto the insulator, and the change in the charge due to hybridization with other oligomer DNA strands in the electrolyte sample is sensed by the change in the EISFET current [1]. The structure with the probe DNA is called the DNAFET. The molecular structure of the probe DNA should be designed and fabricated in such a way that it is matched with the target DNA to be sensed.

Instead of the conventional planar type ion-sensitive FET (ISFET) structure, silicon nanowires (Si-NWs) have been introduced as FET devices [2]. The principle of operation of the silicon nanowire DNAFET is similar to the operation of the conventional DNAFET. Figure 3 shows a DNAFET based on the (a) planar FET, (b) nanowire FET, both of which share similar operating principles; the binding of the target complementary ssDNA with the probe DNA attached on the gate insulator modulates the NW surface potential through a field effect, which in turn produces a change in the current of the FET. The motivation for using the NW FET instead of the planar FET structure is that it may achieve better sensitivity, commonly attributed to the fact that the surface-to-volume ratio of the device is bigger for the nanowire FET. Thus, if the same change in the surface charges occurs in the EI interface, the change in the nanowire FET current is bigger.

Natural extension after the DNA sensing is the immunoassay where the target molecules are the proteins which are generally the indicator of the diseases.

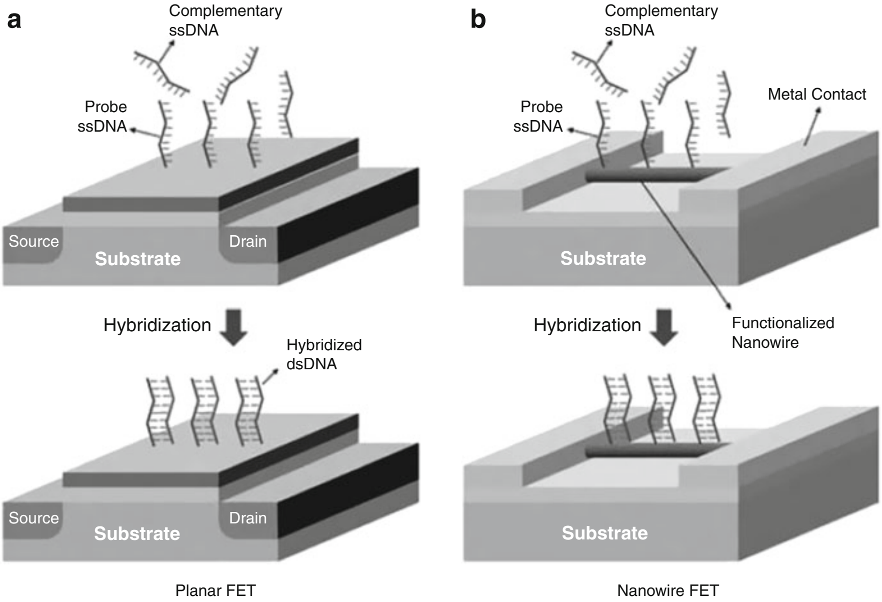


Fig. 3 DNAFET based on a silicon (**a**) planar FET and (**b**) nanowire FET. Notice that both the probe DNA and the target DNA have negative charges in the electrolyte solution (Reproduced by permission of Pan Stanford Publishing [4])

1.2 *Sensor Characteristics of the Affinity-Based Electrical Biochip*

The equation for the binding event between the ss probe DNA and target DNA can be written as

$$\frac{dN_{ds}}{dt} = k_F\left(N_p - N_{ds}\right)C_t - k_R N_{ds}, \tag{1}$$

where N_{ds} and N_p are the areal densities of the hybridized molecules and the probe molecules, respectively. In Eq. (1), C_t is the volume density of the target at the surface and k_F and k_R are the surface kinetic constant of the binding event and dissociation event, respectively.

After the steady state is reached, N_{ds} can be written as, by letting Eq. (1) to be zero.

$$N_{ds} = \frac{N_p}{1 + \left(\frac{k_R 1}{k_F C_t}\right)} \tag{2}$$

If the distribution of the probe molecules on the surface of the gate insulator is not uniform either due to non-uniform attachment of the probe molecules or the non-uniform orientation of the probe DNA toward the electrolyte solution [5], more general expression for N_{ds} can be used. One possible expression is written as

$$N_{ds} = \frac{N_p}{1 + \left(\frac{k_R 1}{k_F C_t}\right)^{-\gamma}}, \tag{3}$$

where γ is a parameter indicating the uniformity of the probe molecules in density, orientation, etc.

The ratio between the N_p and N_{ds} are the signal. In the FET based sensor, the molecular affinity (N_{ds}/N_p) is reflected as the change in the electrical conductance of the channel. The objective of the sensor design is to obtain a faithful transduction of the surface chemistry (affinity) to the electrical signal. If the transduction is faithfully tracing the binding event, the sensitivity of the sensor may be written as, $Sensitivity = N_{ds}/N_p$.

The general sensor characteristics may be best described by three parameters; sensitivity, limit of detection (LOD) and noise. LOD is defined as the minimum concentration of the target molecules to be recognized by the sensor device. From Eq. (1), the minimum detectable C_t is determined by k_R/k_F ratio, if the same noise floor is assumed. For probe molecules with the larger k_F and the smaller k_R with the target molecules, the smaller LOD may be obtained. It can be seen from Eq. (1), LOD and sensitivity are related and the surface kinetics parameters, k_R and k_F, are important parameters to determine the both.

Noise can originate from many sources. One of the serious noise sources is the non selective binding between the noisy molecules and the probe molecules. In the practical applications, the noisy molecules are far more than the target molecules in number and prevention of the non selective binding is the key success factor for the biosensor. In equation, the situation may be described as k_F/k_R (between the target molecules and probe molecules) $\gg k_F/k_R$ (between the noisy molecules and probe molecules).

1.3 Challenges

The challenges are the degradation of the electrical signal due to the charge screening effect, selectivity is degraded by the noisy molecules included in the real samples (such as serum) and the large statistical fluctuation among different measurements (device to device variation).

The limitations on the electrical bio FET structure come from the fact that change in the charges caused by the binding events between the probe and target molecules is screened out by the counter ions existing in the electrolyte. The Debye length determines the scale over which mobile charges screen out electric fields generated

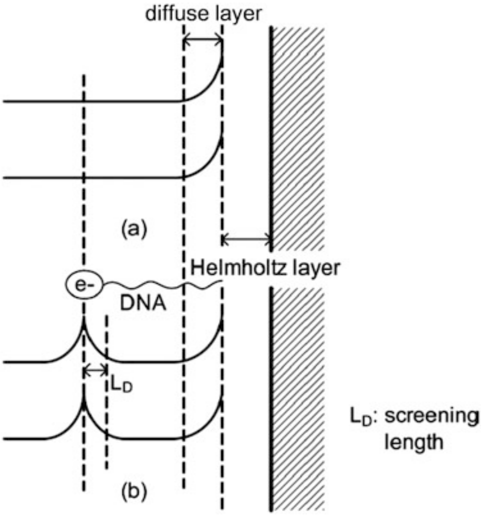


Fig. 4 The schematic diagram of the screening effect. (**a**) The energy diagram without the molecular charge. The effect of the molecular charge outside of the diffuse layer is considered in (**b**). If the Debye length is shorter than the distance from the diffuse layer, the effect of the molecular charge is screened (Reproduced by permission of Pan Stanford Publishing [4])

by the charges in the DNA molecules to be sensed. As the Debye length becomes shorter, the extent over which the change in the charges gives an effect becomes smaller. If the charge is outside the electrical double layer (EDL), the field effect will be nullified before it reaches to the EI interface, giving effect to the surface potential of the FET. The best way to understand the effects of the screening is to refer to Fig. 4. In Fig. 4, we assumed that Q is located between the Helmholtz layer and the diffuse layer in a delta function. However, if the charge is distributed within the diffuse layer, the situation may be depicted as the equivalent circuit shown in Fig. 4b. As the molecular charge and its extent of its effect (the Debye length) get close to the diffuse layer boundary with the electrolyte or outside of the diffuse layer, the effect to the conductance of the channel becomes negligible. Figure 4b is a schematic diagram showing that the mobile charges in the electrolyte screen out the change in the charges from the biomolecules. It should be noticed that the Debye length becomes shorter as the ionic concentration increases so that the screening effects may be more serious as the ionic concentration of the electrolyte increases.

In the explanation of the DNAFET, we neglected to consider the charges at the insulator surface. Since the isoelectric points of silicon oxide and nitride are quite far from the pH values of pure water and serum, the insulator film may be charged as is the case with an EISFET. In this case, the situation is far more complicated than the case we considered in the previous section. The detailed understanding of the Q_{DNA} and its interaction with the EDL, and thereby the modulation of the surface potential of the FET, is by no means fully understood and will be topic of the further research.

Together with the charge screening effect, the challenges in the electrical biochip is the noise and statistical errors. The noise sources are many, out of which the most important challenge comes from the non selective binding of the probe molecules with the noising molecules. Other than selecting the probe molecules having high affinity constant [large k_F in Eq. (2)], electrical method may be devised to help the situation. The electrical pulse method introduced in the next section may help the situation.

Also, large statistical fluctuation among devices and among different samples are the challenges of the electrical biosensors. The fluctuations are mostly originated from the fluctuations in the areal density in probe molecules, orientation and associated fluctuations in affinity constant, k_F. The statistical fluctuation can be largely remedied by using multiple device measurements rather than a single device measurement, thereby statistical signal processing follows with understanding of the average signal, fluctuation, etc.

2 “C-chip” Platform

2.1 *Introduction*

“C-chip” [6] is the FET based biosensor chip with the carbon nanotube network (CNN) built on the metal layers of the CMOS chip. The metal layers work as the metal electrodes and the conductance of the CNN channel is modulated by the affinity reaction between the probe molecules on the channel and the target molecules in the solution once the solution is applied on top of the chip (Fig. 5).

Salient features of the “C-chip” can be summarized as follows. Firstly, the concentric arrangement of the metal electrodes allow effective electrical shielding the channel from the neighbor channels by the surrounding electrode (grounded), so the lithography process can be skipped to etch away the CNN channels between the cells after deposition of the carbon nanotubes (CNT)s. Secondly, reference electrode used to stabilize the potential of the solution can be removed, since the solution potential traces the surrounding electrode potential by way of capacitance coupling (large contact area between the surrounding electrode and solution) [7]. Thirdly, thanks to the simple electrode structures and easy connection of the electrodes to the embedded circuit in the CMOS chip, electrical manipulations are possible in addition to a simple monitoring of change of the conductance of the channel with the molecular binding events. Electrical manipulation includes application of the voltage pulse signal to the electrode so that the electrical field in the electrolyte system may be largely modified. The potential of the electrical pulse method to overcome the general challenges in the affinity based electrical biosensors will be the topics of the next chapter.

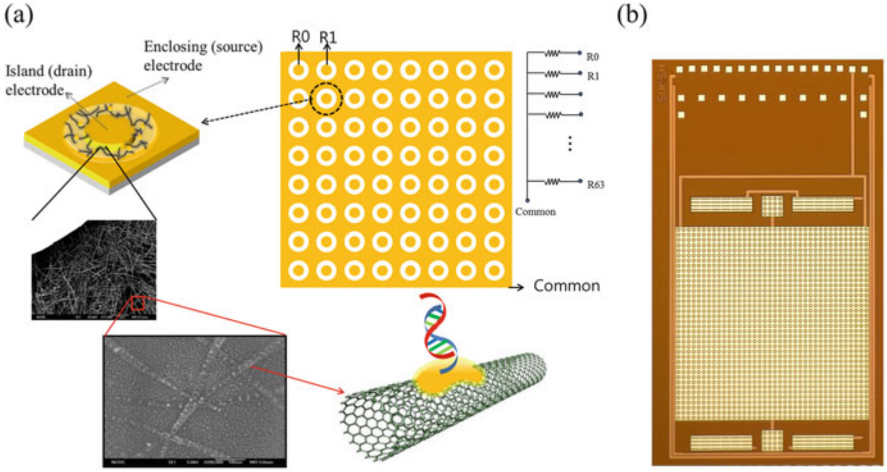


Fig. 5 (**a**) The electrode structure of “C-chip”. Surrounded electrodes (each electrode is drain and connected to the digital to analog (D/A) converter embedded in the CMOS chip) are connected to the surrounding electrode (common ground) by the CNN channel (Reproduced by permission of The Royal Society of Chemistry [8]). (**b**) Fabricated chip photo

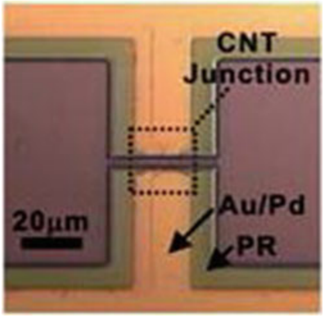


Fig. 6 The conventional electrode. There are two symmetrical electrodes and the non-metallic part is placed between the two electrodes as the channel [11]

2.2 *Concentric Electrode*

Generally, the electrode is an electrical conductor used to make contact with a nonmetallic part of a circuit such as semiconductor or an electrolyte. When the voltage is applied between the two electrodes, current flows from the terminal at higher electric potential to the terminal at lower electric potential.

Most conventional electrode scheme is shown in Fig. 6, in which the CNN is formed between the two symmetrical electrodes. The interdigitated electrode scheme, which is derivative of conventional electrode scheme is also popular in various applications [9]. With respect to electrical properties, one important issue for the CNT device fabrication can be the device-to-device performance variations after the completion of CNT coating. Therefore, for the high-precision assembly of a large number of CNTs on the SiO_2, the confinement of channel area is essential,

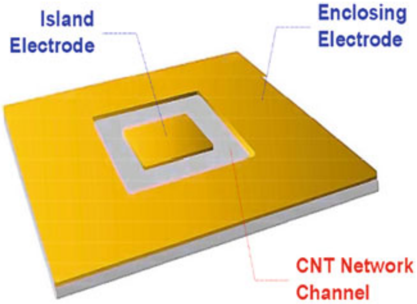


Fig. 7 The conceptual diagram of the proposed concentric electrode. The concentric electrode has an inner electrode (named Island Electrode, IE) enclosed by an outer electrode (named Enclosing Electrode, EE) [11]

requiring an additional photo-lithography process [10]. However, it can be a time-consuming task to produce a large number of CNT devices and, in addition, can cause a side-effect such as photo-resist (PR) residue.

To overcome the above-mentioned problems, the concentric electrode which is requiring no additional photo-lithography process to etch away the unwanted channel area formed by the CNT network is proposed. Figure 7 shows the concept of the concentric electrode scheme. The concentric electrodes have an inner electrode (Island Electrode, IE) enclosed by an outer electrode (Enclosing Electrode, EE) in which the CNT network channel is confined to the area between the two electrodes without additional process [7]. The concentric electrode itself is used as a kind of mask, called the self-aligned channel formation, where the channel is formed automatically by the electrode formation and no additional removal step is necessary to etch away unnecessary channel region. This self-alignment is simple and is very useful because we thereby guarantee that we can coat the CNTs on the entire test chip without additional activity, such as photo-lithography.

Figure 8a, b shows the schematics diagram of the bare CNT network device and its equivalent circuit which is basically a two-terminal resistor with the electrolyte droplet as the gate. While the electrolyte gate is floated, however, it is capacitive coupled with the Island Electrode (drain) and Enclosing Electrode (source). The electrostatic potential of the electrolyte follows the potential of the source as the source-electrolyte capacitance is much larger than the drain-electrolyte capacitance ($C_S \gg C_D$). Notice that butting area between the electrolyte and source is much larger than that between the electrolyte and drain. The source electrode is shared with other devices.

The unique asymmetric feature of two electrodes system is manifested as the asymmetric current versus voltage ($I-V$) characteristics between two electrodes. We have previously reported this “*self gating effect*” using the unique biosensor architectures consisting of two gold electrodes with concentric structures [7]. Figure 8c shows the $I-V$ characteristics of a representative bare CNN device in which the drain voltage is swept from –0.5 V $\sim$ to +0.5 V, while the source voltage is fixed at 0 V. Due to the aforementioned *self gating effect*, the measured current

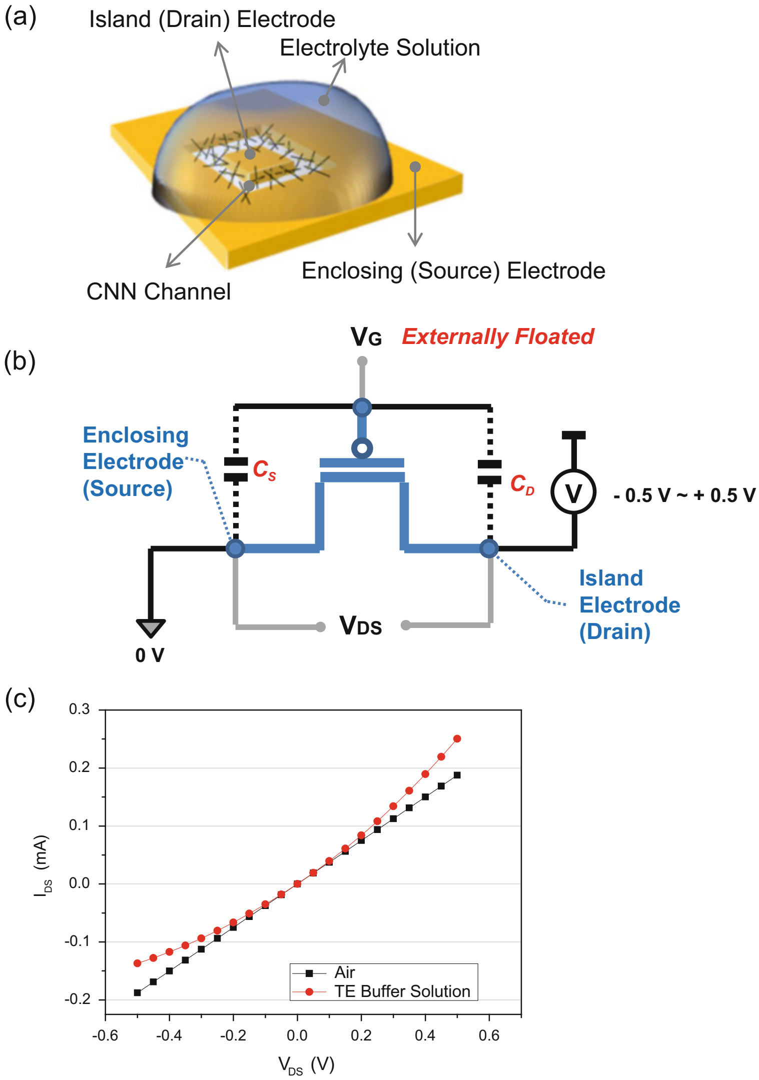


Fig. 8 (**a**) The schematics of concentric structure in electrolyte. (**b**) The equivalent circuit diagram of the concentric structure which has two electrodes. (**c**) The current versus voltage characteristics of a CNN fabricated on concentric structure with the floated gate (Reproduced by permission of The Royal Society of Chemistry [8])

shows well known asymmetric behavior with respect to the polarity of the V_{DS}. It is a diode characteristic for positive V_{DS} and a saturation pMOSFET characteristic for negative V_{DS}. Notice here that our CNN channel is a p-type semiconductor with positive threshold voltage.

2.3 *Carbon Nanotube Network with the GNP (Gold Nano Particles)*

In the "C-chip", the electrical channel is the CNN. Gold nano particles (GNP) decorated on the CNN channel work as the docking places for the probe molecules which are properly treated by thiol chemistry. Thanks to the signal processing circuitry connected with the drain electrodes (surrounded electrodes in Fig. 7), the electrical modulation of each cell can be massively transferred from the CMOS chip and data processing can be performed. Also, by introducing different probe molecules for each cell (or group of cells), multiple assays are possible.

In addition to the CNN with GNP, other parasitic component may influence the channel conductance. They are the defects in CNN and SiO_2. The unwanted charging and discharging of the defects are the main sources of the transient instability of the channel conductance. As the defects are coupled with the ions (mostly H^+ in the electrolyte system) with their own affinity, thereby own time constants, the charging/discharging of the defects modulate the nearby channel through the field effects. They are partly avoided by passivation of the defects in the channel formation fabrication process [12].

Figure 9 shows a schematic diagram of the defects residing on the SiO_2 (where the CNN channel sits) and the CNN channel and the measured current vs. time after the buffer solution is applied to the CNN channel. Two regimes in time are clearly defined in the data. In the 1st transient, current increases and whereas current decreases in the 2nd transient. The 1st transient is attributed to charging of the SiOH sites to negative charge, namely SiO^- in the buffer solution and increases the channel current by accumulating the holes, which are the electrical carriers in the CNTs.

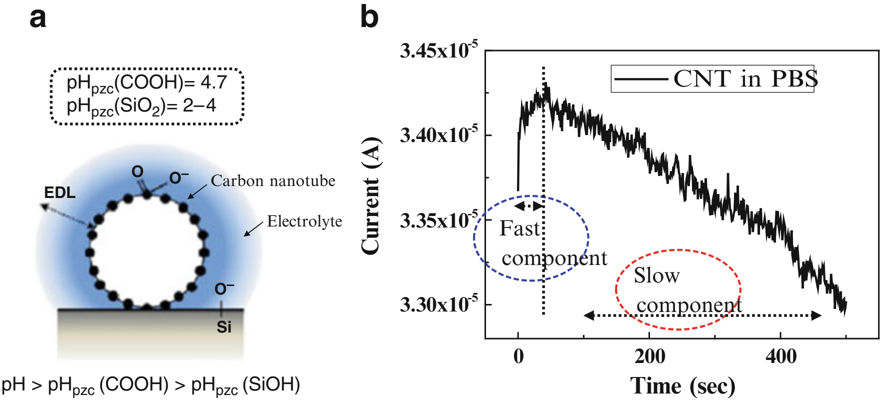


Fig. 9 (**a**) The scheme of SiOH group on the SiO_2 and COOH group on the CNT surface are deprotonated in PBS. (**b**) Real-time measurement of current change of CNT-FET in PBS [12]

However, the slow component, as shown in Fig. 9, shows a decrease in the current with long time constant after the fast current increase. This result can be explained by the negatively charged carboxyl group (COO^-) on the CNT surface. The pH-sensing mechanism for a CNN-based chemical sensor is generally known to be the protonation/deprotonation status of COOH on the CNT surface [13]. Carboxyl group partially dissociates into H^+ and $RCOO^-$ in phosphate buffered saline (PBS) (pH 7.4), and the H^+ dissociation (deprotonation) which occurs in the carboxylated CNT can be [14] considered as the undoping of the hole. The protonation/deprotonation of carboxylic groups on the surface of CNT is considered as hole doping/undoping [15].

The transient effects, especially the slow component originated from the defects on the channel, may be harmful to reliable sensing of biomolecules. One of the methods to avoid the instability is the passivation of the defects by a self-assembled monolayer (SAM) formation. Usually it is used to modify surface properties in the electrochemical sensors applications, and one of the most commonly used SAM is organosilane monolayer on hydroxyl surfaces such as oxide surface. A chemical modification of SiO_2 surface by organo-functional silane is a well-known technique for biosensor application [16, 17]. Among them, 3-aminopropyltriethoxysilane (APTES) is widely used as interfacing molecules. In aqueous solution, the amino-silanes undergo hydrolysis and polymerization in the bulk phase before forming bonding with SiOH of oxide surface [18]. Therefore, we propose an anhydrous silanization which can offer a homogenous monolayer of the amino-silane on the oxide surface. In our experiment, the oxide surface of CNN-based sensor may have different charge densities depending on the ionization of silanol groups (–SiOH). Figure 10 shows the time dependence of the normalized current for pristine CNT and the SiO_2 surface after being passivated with APTES. As shown in Fig. 10a a surface modification with APTES must consume SiOH functional groups on the oxide surface and should change the charge density at the surface. To modify the oxide surface of the channel region, a CNN-based sensor was soaked in an

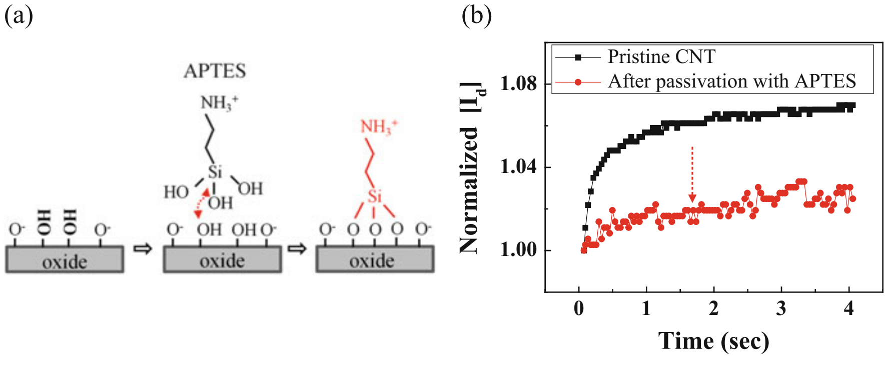


Fig. 10 (**a**) The scheme of the SiO_2 surface passivation with APTES and (**b**) time dependence of the normalized current for the pristine CNT [12]

20 anhydrous ethanol solution containing 2 % APTES for 1 h, followed by a rinsing process under an alcohol flow [19]. APTES has an amine group that will be protonated and positively charged in an aqueous solution. An APTES treatment can induce a positive charged site on the oxide surface and decrease the negative net charge from the silanol groups on the oxide surface as shown in Fig. 10a. After the surface modification of the CNN-based sensor with APTES, we observe that a current increase is reduced appreciably, as shown in Fig. 10b. Therefore, it is clearly confirmed that the fast component of the current signal is due to the effect of the silanol group on the oxide surface.

2.4 Passivation of the CNN Defects

The pH-sensing mechanism adopted in a carboxylated CNT-based chemical sensor has been generally known to be due to a defect functional group, especially COOH [20]. The CNN used in our experiment is treated in HNO_3 for 30 min, so it is estimated that the carboxyl group (COOH) is introduced on the CNT surface. In general, COOH group is deprotonated to COO^- and negative charged correspondingly to the difference between the bulk pH and the isoelectric pH (pH_{iep}) of the COOH.

The carboxylated CNT is a p-type semiconductor and the protonation/deprotonation is interpreted as the hole doping/undoping [21]. To investigate the relation between the deprotonation status of COOH groups on the CNT surface [22] and conductance change of CNN, we measured the current change of CNN channel with several devices for each different pH values. Figure 11a shows the real-time current measurement of CNN-based sensor with different pH buffer solutions. This suggests that the conductance property of the CNT is expected to be affected by the pH value of solution. The CNT-FET conductance decreases as pH value increases. The current is significantly dependent on pH and protonation/deprotonation of COOH plays a role of the conductance change of CNN channel. It can be explained by the fact that the COOH group is fully associated at pH 3, and COOH partially dissociates into H^+ and $RCOO^-$ at pH 7, while H^+ is fully dissociated from the COOH group at pH 9.

In order to verify the role of COOH group, we carried out a capping treatment for COOH group on the CNT surface. We passivated the COOH defects of the CNT by the 1-(3-(dimethylamino)-propyl)-3-ethyl carbodiimide hydrochloride (EDC)/N-hydroxy-succinimide (NHS) treatment. EDC is a cross-linking agent, used to convert COOH groups to amine-reactive NHS ester in the presence of NHS [23]. Therefore, EDC/NHS treatment is usually used to prepare surface modification in biosensor application. After this treatment, a significant current decrease disappears as shown in Fig. 11b. This result is in sharp contrast to our previous observations in Fig. 11a. We attributed the pH independent behavior of CNN to a capping of defect. This experiment clearly demonstrates the role of the COOH group on the conductance property of the carboxylated CNN channel in aqueous solution [24].

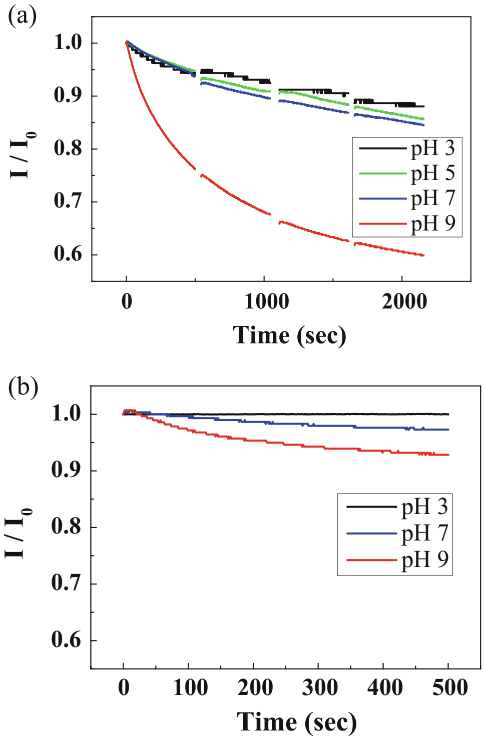


Fig. 11 (**a**) The real-time measurement for a pristine CNT-FET in various pH solutions and (**b**) after the capping of the COOH group on the CNT surface [12]

This pH-dependent behavior of the CNN is an undesirable side reaction in a biosensor application. Therefore, we suggest that a capping treatment of COOH defects on CNT surfaces is a promising method to obtain a reliable operation of the biosensor adopting CNN channel. Furthermore, we propose that the defect of CNT may be useful for binding sites of antibody immobilization. COOH may be reacted to NHS in the presence of an EDC, resulting in an intermediate species which can react with amine group of protein surface. After covalent attachment of IgG antibodies to COOH groups (defects), the change in channel current was measured by the time–current measurement before and after antibody immobilization, as shown in Fig. 12. After antibody immobilization on the COOH, we would like to emphasize that the CNN channel became fully insensitive with pH condition.

2.5 Electrical Pulse Method: The SNU (Seoul National University) Approach

In various medical and biological applications, the label-free detection of the charged biomolecules such as DNA/RNA and proteins has a number of advantages over the well-established optical methods [25, 26]. However, the degradation of the

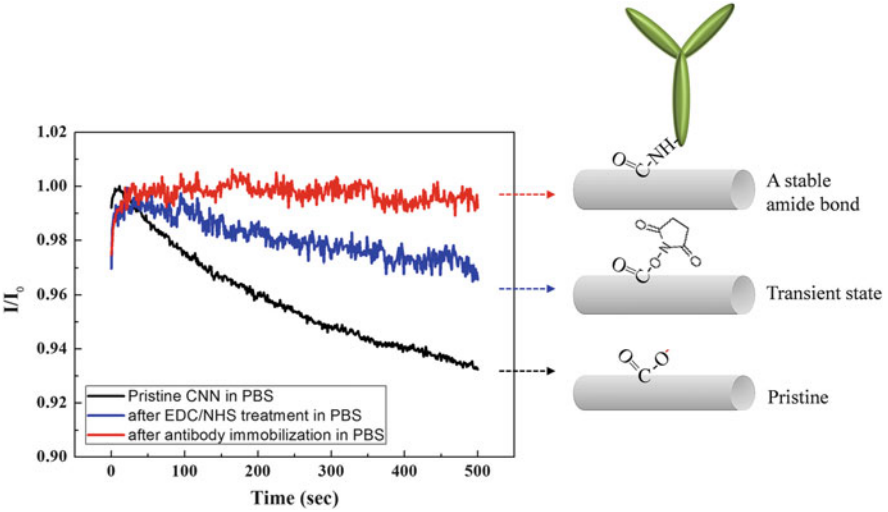


Fig. 12 Real-time measurement of the normalized current after application of unit voltage of 0.1 V to the drain for various treatment of the CNN channel [12]

signal due to non-specific binding with the background molecules and the charge screening effect due to the ions within the EDL have been the roadblocks for the success of the electrical biosensor [27, 28].

Our group applied an electrical pulse bias between the electrical channel, to which the probe molecules are attached, and the electrolyte containing the target molecules during the hybridization event. The method provides two important improvement; increase in the hybridization rate and the transient signal detection before a charge screening effect is resurrected. Firstly, the hybridization rates between the probe and target molecules are significantly improved due to the mechanical oscillations of the probe and target molecules caused by an external electric field around charge probe molecules. Also, the dynamic motion of the probe molecules could suppress the undesired binding event with nonspecific molecules (Fig. 13).

Secondly, it is possible to perform the transient measurement of the electrical channel current after pulse transition. By properly choosing the timing of the measurement after pulse transition, one can measure the electrical signal from the target molecule charges before the charges are screened by the ions in the system. Generally the real time monitoring of the molecular hybridization events is performed under the buffer solution, where the Debye length is ∼1 nm (comparable to 100 mM in ionic concentration), while the size of biomolecule is over several nanometers. In this case, large part of the charges from the biomolecules is screened and the sensitivity of the sensor is sacrificed. It takes a while for ions in the electrolyte system to redistribute to reach the steady state for another screened state after the pulse bias applied. The detection within the transient state can provide opportunities to the sensitivity enhancement (Fig. 14).

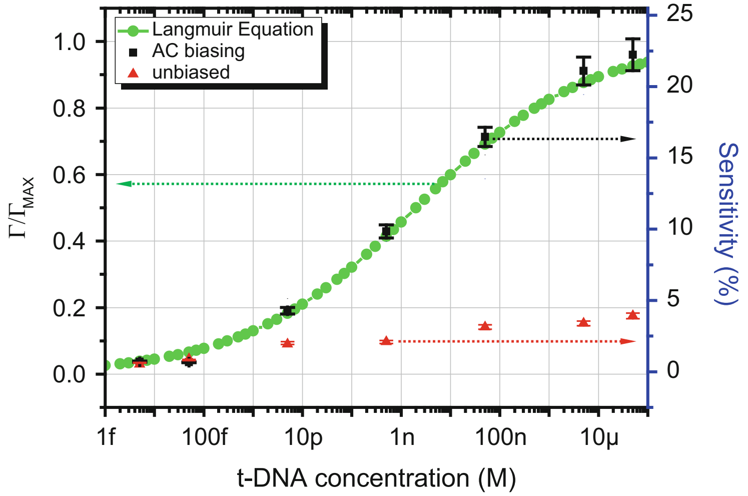


Fig. 13 Sensitivity of the electrical DNA sensor vs. target-DNA (t-DNA) concentration, for two groups of sensor devices after hybridization with t-DNA in the serum conditions; one group hybridized under the pulsed bias (±0.5 V in amplitude and 1 kHz in frequency), and the other group hybridized under the unbiased conditions. The pulse biasing during hybridization achieves remarkable improvement in sensitivity (Reproduced by permission of The Royal Society of Chemistry [8])

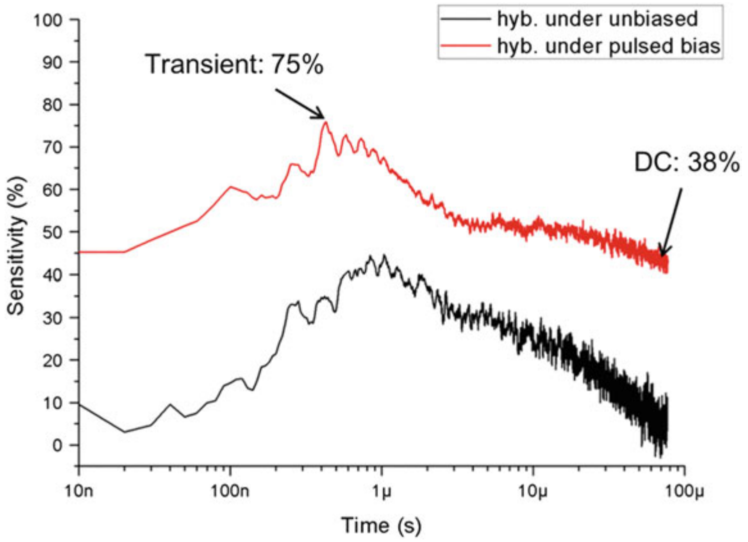


Fig. 14 Sensitivity vs. time after the unit pulse is applied (transient measurement), for two sensor devices, after hybridization of 1,000 s. The data denoted by the *black line* and by the *red line* are for the devices hybridized with the complementary t-DNA of 5 mM, under the unbiased conditions, and AC bias conditions, respectively (Reproduced by permission of The Royal Society of Chemistry [8])

Even though comprehensive simulation platform has been proposed in [Liu and Dutton, IEDM] [29] in two and three dimensions, we have made an in-house 1 dimensional simulator to consider the timing behavior of ions and charged molecules and their effects on the electrode. In this model, the charged molecules are assumed to be at fixed positions. The electrical potential and charged carriers in the electrolyte solution are governed by the Poisson equation,

$$\nabla \cdot \varepsilon \left(-\nabla \psi\right) = q \left(\left[n^{+}\right] - \left[n^{-}\right]\right), \tag{4}$$

where ε is the permittivity of electrolyte solution, ψ is the electrical potential, and $[n^{+}]$ and $[n^{-}]$ are the number density of the cation and anion in the electrolyte solution, respectively.

In the electrolyte solutions, ion carriers accelerate through the drift and diffusion similar to the transport of the electrons and holes in the semiconductor. Accordingly, the ion flux is represented by

$$F_{[n^{\pm}]} = \pm \mu_{n^{\pm}} \left[n^{\pm}\right] \nabla \psi - D_{n^{\pm}} \nabla \left[n^{\pm}\right], \tag{5}$$

where $\mu_{n\pm}$ and $D_{n\pm}$ are the mobility and the diffusion coefficient of cation and anion, respectively, in the electrolyte solution.

The continuity equation for ion carriers can be written as

$$\frac{\partial \left[n^{\pm}\right]}{\partial t} = -\nabla \cdot F_{[n^{\pm}]}. \tag{6}$$

In the steady state, the left term of Eq. (6) becomes zero, yielding the Boltzmann distribution for ions,

$$\left[n^{\pm}\right] = n_0 \ \exp\left(\mp \psi / V_t\right), \tag{7}$$

where n_0 is bulk ion concentration of the electrolyte solution. The diffuse layer derives from this distribution, suggested by Gouy and Chapman.

After Stern’s modification that considers the Helmholtz layer due to the minimum distance between the ions and the electrode surface, the EDL has been regarded as a series capacitance made up of the Helmholtz layer and diffuse layer.

The parameters and simulation conditions used in this work are summarized in Table 1 and Fig. 15. The electrical potential ψ_0 is applied to the electrode with respect to bulk solution, which is grounded by the reference electrode. Generally, the semiconductors are employed as the electrode in the affinity-based biosensors to detect the charge induced by charged molecules. The electrode surface charge induced by charged molecules is calculated from

$$Q = -\int \mathbf{D} \cdot d\mathbf{s}. \tag{8}$$

Table 1 Parameters used in this simulation [30]

Parameter	Value	Note
ε	$78 \times \varepsilon_0$	Aqueous solution in room temperature
μ_{H+}	33.3×10^{-4} cm^2/V·s	
μ_{OH-}	18.8×10^{-4} cm^2/V·s	
μ_{Na+}	5.9×10^{-4} cm^2/V·s	
μ_{Cl-}	7.0×10^{-4} cm^2/V·s	
$D_{n\pm}$	$\mu k_B T/q$	Einstein relation
n_0	0.1 M	In buffer solution
d_H	5 Å	Thickness of Helmholtz layer

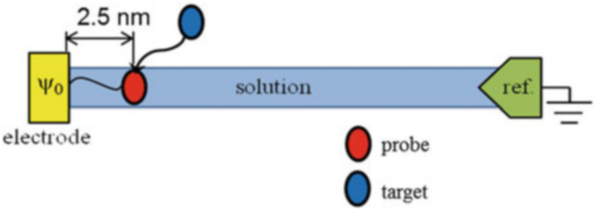


Fig. 15 Schematic diagram of the probe-target binding event in the affinity-based biosensor [30]

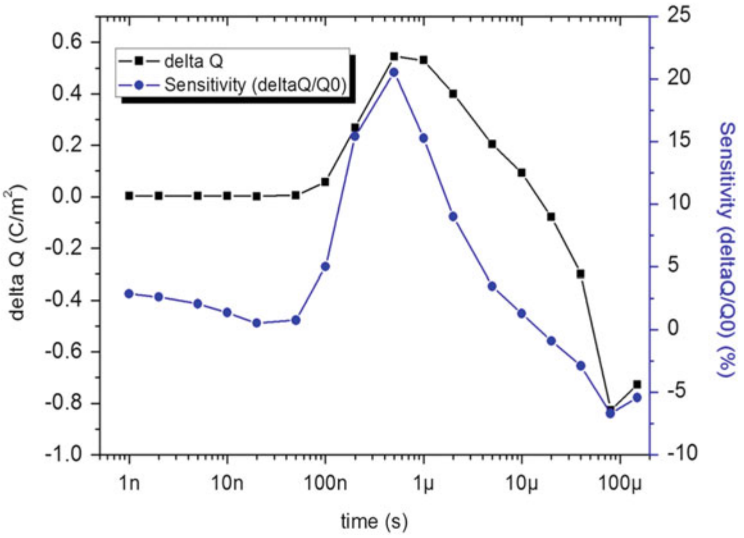


Fig. 16 Variance of induced surface charge and corresponding sensitivity with respect to time after step-pulse biasing in the buffer solution [30]

Figure 16 shows the change of the induced charge variation and the sensitivity with the time. At $t = 0$ s, step pulse voltage is applied to the electrode (distance = 0) in the buffer condition. High step pulse bias induces the electro-diffusion flow of mobile ions on transient state, so that the screening length is extended instantaneously. As the screening length is extended, the sensitivity increased to reach the maximum, and decreases again as the system settles down at the steady state.

3 Summary; Toward the General Immunoassay

In this chapter, we summarized the general challenges faced by the electrical biosensor chips. They are mostly related with the electrical signal degradation due to the charge screening effects and statistical uncertainty introduced from the fluctuations in the surface chemistry (density and affinity of the probe molecules).

In addition to efforts to develop the probe bio molecules with high affinity and selectivity with the target molecules, judicious selection of the electrical signal and understanding of the behavior of the electrical response of the channel may help the situation a lot.

In this chapter, we introduced the "C-chip" platform having electrical channel (CNN with the gold nano particle) built on the final metal layer (built on the top surface) of the CMOS chip. With multiple sensor devices and connection of the metal electrodes with the signal processing circuits built in the CMOS chip, massive statistical treatment of the multiple sensor chips are performed.

In addition, we have shown the potential of the electrical pulse techniques applied to the electrode of the sensor devices. Our strategy is very simple; adoption of the electrical pulse to the electrode. We have shown that electrical pulse method mitigates the signal degradation caused by the charge screening effects by agitating the screening ions. In addition, it has been shown by the DNA sensing experiments that the electrical pulse applied during the binding events enhances the hybridization affinity significantly.

Three time constants may be involved during the agitation. The first time constant is associated with the RC time constant before the EDL is stabilized. During this time period, large electric field is applied to the neutral solution and charge de screening effect can be observed. The second time constant is associated with modulation of the pH value at the surface and charging/discharging of the target molecules. The time constant may be determined by the mobility of proton and association constant of proton and radicals of the molecules. The third time constant is associated with the motion of molecules according to the transient electric field both in the EDL and the neutral solution region.

The electrical pulse methods proposed in this work together with the massive parallel measurements may help to shorten the realization of the general platform for the immunoassay chips.

Acknowledgments This work is supported by the Pioneer Research Center Program and by the Center for Integrated Smart Sensors as the Global Frontier Project through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning.

References

1. Cui Y, Wei Q, Park H, Lieber CM. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science. 2001;293(5533):1289–92. doi:10.1126/science.1062711.
2. Xie P, Xiong Q, Fang Y, Qing Q, Lieber CM. Local electrical potential detection of DNA by nanowire–nanopore sensors. Nat Nanotechnol. 2012;7(2):119–25. doi:10.1038/nnano.2011.217.
3. Bousse L, De Rooij N, Bergveld P. Operation of chemically sensitive field-effect sensors as a function of the insulator–electrolyte interface. IEEE Trans Electron Devices. 1983;30(10):1263–70.
4. Park BG, Hwang SW, Park YJ. Nanoelectronic devices. Singapore: Pan Stanford Publishing Pte. Ltd; 2012.
5. Liu Y, Dutton RW. Effects of charge screening and surface properties on signal transduction in field effect nanowire biosensors. J Appl Phys. 2009;106(1):014701.
6. Ko JW, Woo JM, Jinhong A, Cheon JH, Lim JH, Kim SH, Chun H, Kim E, Park YJ. Multi-order dynamic range DNA sensor using a gold decorated SWCNT random network. ACS Nano. 2011;5(6):4365–72. doi:10.1021/nn102938h.
7. Kim DW, Choe GS, Seo SM, Cheon JH, Kim H, Ko JW, Chung IY, Park YJ. Self-gating effects in carbon nanotube network based liquid gate field effect transistors. Appl Phys Lett. 2008;93(24):243115. doi:10.1063/1.2978095.
8. Woo JM, Kim SH, Chun H, Kim SJ, Ahn J, Park YJ. Modulation of molecular hybridization and charge screening in a carbon nanotube network channel using the electrical pulse method. Lab Chip. 2013;13(18):3755–63. doi:10.1039/c3lc50524c.
9. Li J, Lu Y, Ye Q, Cinke M, Han J, Meyyappan M. Carbon nanotube sensors for gas and organic vapor detection. Nano Lett. 2003;3(7):929–33.
10. Maeng S, Moon S, Kim S, Lee H-Y, Park S-J, Kwak J-H, Park K-H, Park J, Choi Y, Udrea F. Highly sensitive NO_2 sensor array based on undecorated single-walled carbon nanotube monolayer junctions. Appl Phys Lett. 2008;93(11):113111–3.
11. Seo SM. (A) CMOS integration of metal decorated carbon nanotube network (C chip) and its application to molecular sensing. PhD. Seoul National University; 2011.
12. Kwon HJ. Electrical detection of the microcystin-LR using carbon nanotube network channel. PhD, Seoul National University, 2013.
13. Ishikawa FN, Curreli M, Olson CA, Liao H-I, Sun R, Roberts RW, Cote RJ, Thompson ME, Zhou C. Importance of controlling nanotube density for highly sensitive and reliable biosensors functional in physiological conditions. ACS Nano. 2010;4(11):6914–22.
14. Choi Y, Moody IS, Sims PC, Hunt SR, Corso BL, Perez I, Weiss GA, Collins PG. Single-molecule lysozyme dynamics monitored by an electronic circuit. Science. 2012;335(6066):319–24. doi:10.1126/science.1214824.
15. Liao H-K, Chi L-L, Chou J-C, Chung W-Y, Sun T-P, Hsiung S-K. Study on pH_{pzc} and surface potential of tin oxide gate ISFET. Mater Chem Phys. 1999;59(1):6–11.
16. Zhao X, Kopelman R. Mechanism of organosilane self-assembled monolayer formation on silica studied by second-harmonic generation. J Phys Chem. 1996;100(26):11014–8.
17. Simon A, Cohen-Bouhacina T, Porte M, Aime J, Baquey C. Study of two grafting methods for obtaining a 3-aminopropyltriethoxysilane monolayer on silica surface. J Colloid Interface Sci. 2002;251(2):278–83.
18. Joshi M, Goyal M, Pinto R, Mukherji S. Characterization of anhydrous silanization and antibody immobilization on silicon dioxide surface. In: 2003 IEEE international workshop on computer architectures for machine perception. IEEE; 2004. p. 7–11.
19. Fu Q, Lu C, Liu J. Selective coating of single wall carbon nanotubes with thin SiO_2 layer. Nano Lett. 2002;2(4):329–32.

20. Lee D, Cui T. pH-dependent conductance behaviors of layer-by-layer self-assembled carboxylated carbon nanotube multilayer thin-film sensors. J Vac Sci Technol B: Microelectron Nanometer Struct. 2009;27(2):842. doi:10.1116/1.3002386.
21. Lee D, Cui T. Low-cost, transparent, and flexible single-walled carbon nanotube nanocomposite based ion-sensitive field-effect transistors for pH/glucose sensing. Biosens Bioelectron. 2010;25(10):2259–64.
22. Gao C, Guo Z, Liu J-H, Huang X-J. The new age of carbon nanotubes: an updated review of functionalized carbon nanotubes in electrochemical sensors. Nanoscale. 2012;4(6):1948–63.
23. Snow ES, Novak JP, Campbell PM, Park D. Random networks of carbon nanotubes as an electronic material. Appl Phys Lett. 2003;82(13):2145. doi:10.1063/1.1564291.
24. Bui M-PN, Pham X-H, Han KN, Li CA, Lee EK, Chang HJ, Seong GH. Electrochemical sensing of hydroxylamine by gold nanoparticles on single-walled carbon nanotube films. Electrochem Commun. 2010;12(2):250–3.
25. Daniels JS, Pourmand N. Label-free impedance biosensors: opportunities and challenges. Electroanalysis. 2007;19(12):1239–57.
26. Thévenot DR, Toth K, Durst RA, Wilson GS. Electrochemical biosensors: recommended definitions and classification. Biosens Bioelectron. 2001;16(1):121–31.
27. Stern E, Wagner R, Sigworth FJ, Breaker R, Fahmy TM, Reed MA. Importance of the Debye screening length on nanowire field effect transistor sensors. Nano Lett. 2007;7(11):3405–9. doi:10.1021/nl071792z.
28. Zhang G-J, Zhang G, Chua JH, Chee R-E, Wong EH, Agarwal A, Buddharaju KD, Singh N, Gao Z, Balasubramanian N. DNA sensing by silicon nanowire: charge layer distance dependence. Nano Lett. 2008;8(4):1066–70.
29. Liu Y, Lilja K, Heitzinger C, Dutton RW. Overcoming the screening-induced performance limits of nanowire biosensors: a simulation study on the effect of electro-diffusion flow. In: Electron devices meeting, 2008. IEDM 2008. IEEE International. IEEE; 2008. p. 1–4.
30. Woo J-M, Kim SH, Park YJ. Transient state in the affinity-based biosensor: a simulation and experimental study. In: IEEE international conference on simulation of semiconductor processes and devices, 2012; p. 177–80.

Chemomechanical Transduction Systems: A Sensing Platform by Surface Force Measurement

Junghoon Lee and Jun-kyu Choi

Abstract The transduction of chemical binding event into a mechanical deformation has been developed as a label free sensing platform with potentially mobile detection setups. In this chapter we will describe the basic and engineering principles of the chemomechanical transduction with key example applications.

1 Introduction

Chemo-mechanical transduction for biological and chemical events open up a new horizon for the expanding applications of diagnostics, drug discovery, security and threat evaluation. Although DNA and protein analysis are mature technologies, they usually utilize extrinsic labels which require expensive detection equipment to decide if the label is present after DNA–DNA hybridization or antigen–antibody binding. These works can be carried out on a chip platform without the use of exogenous labels. Micro- and nanometer scale cantilevers have been mainly used for those purposes. The micro cantilever and polymer demonstrated surface stress change caused by a specific bio molecular interaction such as self-assembled monolayer formation [1], DNA hybridization [2], antigen–antibody binding [3] and cellular binding [4]. These approaches are, however, hardly realized into a compact device due to the bulky optical detection equipment and it has trouble with various selectivity performance and robustness [5]. Although micro cantilever provides a viable alternative to resonant mass sensing techniques, the surface stress sensing mechanism is fundamentally different with membrane technology, where the resonance frequency detection decides the change of adsorption mass on the resonator [6]. For sensing surface stress, thin membrane transducer on the other hand has several advantageous characteristics. Firstly, shell structures are more

J. Lee (✉)
School of Mechanical and Aerospace Engineering, Seoul National University,
Seoul 151-744, South Korea
e-mail: jleenano@snu.ac.kr

J.-k. Choi
Small Machines Co., Ltd, Bldg#314 Rm#307, 1 Gwanak-ro, Gwanak-gu,
Seoul 151-742, South Korea

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_2

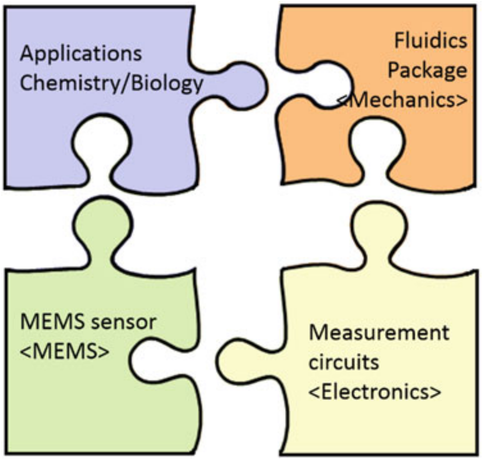


Fig. 1 Compact and portable chemical and biological detection system need broad field technologies such as MEMS, chemistry, biology, mechanics and electronics

robust and stable than cantilever beams, but it is very sensitive to surface reaction [7], therefore, can be easily functionalized and probed by using commercially available printing techniques [8]. Secondly, the detection surface is physically isolated from the electrical sensing surface, which can be easily used for low-noise precise capacitance measurement technology [9]. Also, the isolated surface is a diverse platform which is accessible by both liquid and gas samples. Finally, electric measurement of sensing structure can readily be scaled and multiplexed with each reaction chamber [10, 11].

In the previous research [12–14], a polymer based thin membrane transducers are introduced to show the possibility of bio molecular detection. To improve the lower sensitivity of the membrane transducer compared to cantilever type transducer, polymer material which has lower rigidity stiffness is used. The advantage of low rigidity stiffness is offset by the decrease of reliability in wet environments and fabrication uniformity compared with micromachined materials; silicon oxide, poly silicon, and silicon nitride. In addition to difficulties in fabricating a completely integrated device with polymer membrane, polymer based transducer has issues in the packaging of fluidic channel and measurement circuits because of high flexibility and low melting point. The reliability of chemical and biological diagnosis is crucial for commercial products, as reliability often dominates the device designs. MEMS devices are core products in future integrated systems which miniaturization and advanced functionalities are inevitable described in Fig. 1.

2 Theoretical Background

Berger et al. [1] demonstrated mechano-chemical transduction by using a long polymer strands, the trans and cis isomers of a 1,2-disubstituted benzocyclobutene. The polymer is subject to an ultrasound mechanical force caused by electrocyclic ring opening in a formally conrotatory and disrotatory process. During the chemical reactions, reactant molecules require certain amount of energy to overcome the energy barrier to allow their transformation. Usually, the energy is provided in the form of heat, light, pressure or electrical potential, however mechanical force moving the reactants from ground state to excited state potential energy is facilitated to surmount the energy barrier. An opposite case is a chemomechanical transduction. Chemical reactions induce the mechanical deformation which is inspired of nature nanotechnology. For example, ion channels in cell membrane are transmembrane proteins that regulate ionic permeability on lipid bilayer membrane. The operation of this transmembrane protein is controlled by mechanical feedback induced surface stress due to chemical reaction [2]. In Fig. 2, a calcium ion channel can pump the ions out of the cell membrane when two gramicidin monomer molecules are attached. This process involves mechanical stress since the length of this dimer molecule is smaller than the membrane thickness. When the calcium level build an addition in cell membrane, the surface stress of the curved lipid bilayer is decreased, and then the dimer will be separated into two monomers and stops functioning.

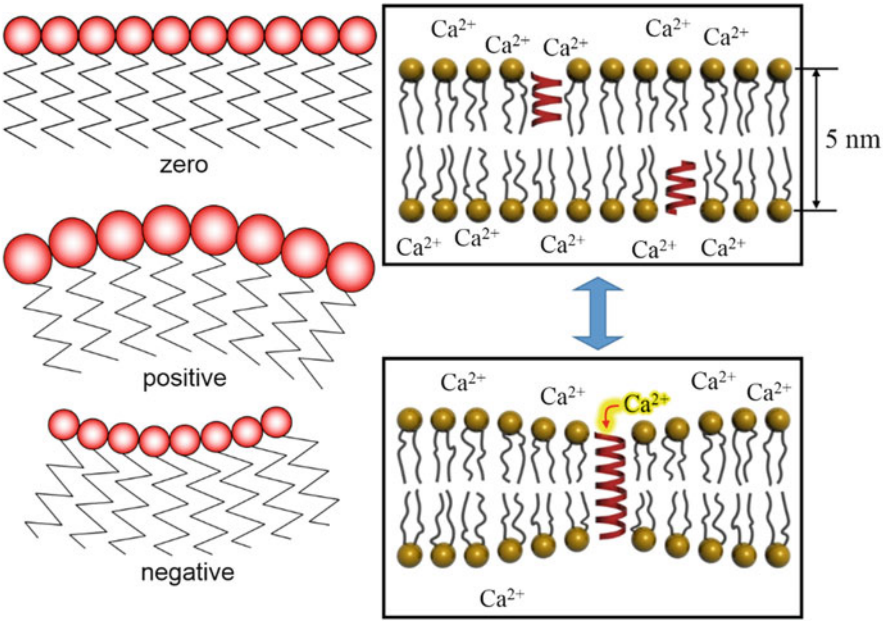


Fig. 2 Equilibrium curvatures of lipid monolayer which is dependent on surface free energy in a specific environment

2.1 Principle of Chemo Mechanical Transduction

AFM micro cantilevers are commonly used in various microscopy applications detecting electrostatic, spatial and electronic forces on the surfaces. Some researchers utilize the entire cantilever beam instead of the cantilever's tip as a probe for large range of detection [15]. The molecular interactions cause the cantilever to bend upon adsorption on the cantilever surface. Using the mechanism, physical, chemical and biochemical processes can be directly transduced into nanomechanical responses due to the surface stresses from the energy of molecular interactions. The surface energies are first defined by Josiah Willard Gibbs as the amount of reversible work per unit needed to elastically stretch a pre-existing surface. Surface stress occurs when surface atoms or thin films undergo some dynamic molecular interaction processes resulting in a change in density. If the bond strength between surface atoms is stronger than that of substrate atoms, a tensile surface stress will be generated by the attractive forces that result in a concave surface curvature.

2.1.1 Surface Stress

A nano- and microscopic processes induces the surface stress of a macroscopic quantity. Recent mechanochemical interactions have been investigated with molecular reconstruction, interfacial mixing, and self-organization at solid surfaces have renewed the interest in the study of surface stress [4, 5]. In fact, surface stress arises when surface atoms go through some dynamic micro-structural processes such as thermal reflow, high temperature deposition, and deposition of different stress materials. For example, surface atoms that tend to repel each other will collectively induce a compressive surface stress, resulting in a convex surface curvature. In contrast, when surface atoms are attracted to each other, a tensile surface stress will be resulted to form a concave surface curvature. Figure 3 illustrates the mechanism of compressive and tensile surface stress showing the surface curvature.

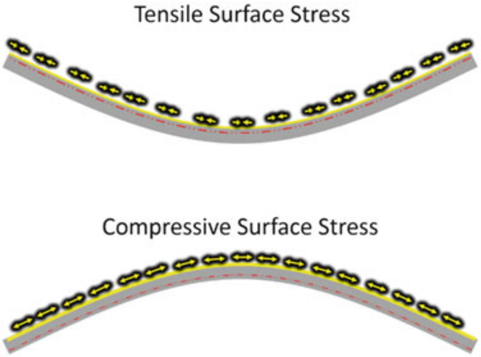


Fig. 3 Surface stress: A tensile surface stress (positive surface stress) contracts the top surface of a thin plate inducing a concave curvature; a compressive surface stress (negative surface stress) expands the tip surface of a thin plate inducing a convex curvature

When a clean metal surface is deposited, the electrons of a metal replace themselves in response to the absence of atoms on the surface; the distribution of metal atoms near the surface is distinguished with what it is in the bulk of the material. If the charge density of metal molecules would be the same at the surface with the bulk state of the material, no surface stress would be generated. Gold surface has inherent tensile surface stress because of the competition between the repulsive interactions between the filled d shells and an electron gas attraction from the mobiles sp electrons [16]. Therefore, the increases of charge density result in an increase in tensile surface stress, which can be large enough to initiate surface reconstruction. In the case of Au(111), surface atoms are reconstructed to compensate for this increase in surface charge density so as to reduce the surface free energy. As a result, the surface accommodates an additional Au atom in a $(23 \times \sqrt{3})$ reconstructed unit cell, which relieve the native tensile surface stress. The adsorption on a metal surface causes the change of electronic charge density which induce a compressive surface stress on the substrate resulted in deformation [17]. In an opposite case, the tensile surface is similar interpretation occurring at Si(100) surface [18]. Shuttleworth has defined the surface stress equation in terms of the surface stress tensor σ_{ij} and surface energy, γ; which named as the Shuttleworth equation [19].

$$\sigma_{ij} = \gamma\delta_{ij} + \frac{\partial\gamma}{\partial\varepsilon_{ij}} \tag{1}$$

In the equation, δij is the Kronecker delta and εij is the elastic strain tensor. On the liquid surface, surface stress is equal to surface energy, γ, in which the second term vanishes as there is no resistance acting the plastic deformation on the liquid surface. However, the surface energy is changed according to the change of energy during elastic stretching of a pre-existing surface at solid interfaces that make the change of the second term of Eq. (1).

2.1.2 Thin Plate Theory

Plate theory is suitable for the analysis of micromachined thin film diaphragms. In the plate theory, both small and large deflection has different characteristics, which conditions can be classified as shown in Table 1. Normally, membrane deflection belongs to the small deflection range due to a uniform axial surface stress.

Table 1 Conditions of plate theory for small and large deflection

Theory	Conditions
Small deflection	Plate deflection $\leq 0.2 \times$ plate thickness and/or Plate deflection $\leq 0.04 \times$ plate diameter [20]
Large deflection	Plate deflection $> 0.3 \times$ plate thickness [21]

However, large initial deflection must be considered in order to obtain the change in differential surface stress due to a variety of reasons including imperfections of the membrane layer, a thin stressed film or adsorbed species on the surface, or deflections due to hydraulic gravity.

A classical plate theory of Timoshenko [22] explains that the mechanical behavior of a plate is dominated by the resistance of the deflection, by measuring the flexural rigidity, D:

$$\mathrm{D} = \frac{Et^3}{12\left(1-\nu^2\right)}, \tag{2}$$

where E, t, and ν are Young's modulus, plate thickness, and Poisson's ratio, respectively. For a circular plate with clamped edges, the initial deflection [22] is

$$w_i(r) = w_\delta\left[1-\left(\frac{r}{a}\right)^2\right]^2, \tag{3}$$

where a is the plate radius and w_δ is the maximum initial deflection (at r = 0). The total deflection of both initial bending and a surface stress is [7, 23]

$$\mathrm{w}\,(\mathrm{r}) = \left[\frac{32w_\delta}{(\beta a)^2}\right]\left[\frac{J_0(\beta r)-J_0(\beta a)}{(\beta a)\,J_1(\beta a)} - \left(\frac{1}{2}\right)\left(1-\left(\frac{r}{a}\right)^2\right)\right], \tag{4}$$

where J_0 *and* J_1 are Bessel functions of the first kind, $\beta^2 = \frac{\sigma_s}{D}$.

The change in differential surface stress is then

$$\Delta\sigma_s \approx \left(\frac{80}{11}\right)\gamma\left(1+\frac{\sqrt{w_\delta\left(w_\delta+4\Delta w\right)}}{w_\delta}\right), \tag{5}$$

where $w_\delta < 0,\ \gamma = \frac{D}{a^2},\ and\ \Delta\sigma_s = (\sigma_s)_{t=t_f} - (\sigma_s)_{t=0}$.

For applications such as label-free bio sensing, the precise calculation of the binding induced surface stress change may not be necessary; however, choosing a receptor layer with the appropriate functional group that generates a large change in surface stress upon binding is crucial to optimize the signal-to-noise ratio.

Small Deflection Theory

Kirchhoff used several assumptions to develop linear plate theory, which is known as Kirchhoff's Hypothesis. Small deflection theory is based on the Kirchhoff-Love Hypotheses [20]:

1. The plate material is elastic, homogenous, continuous, and isotropic.
2. The bending deflections are small compared to the thickness of the plate. The slope of the deflected plate is also small; hence the square of the deflection is small.
3. Plane sections originally normal to the surface are presumed to be normal after bending. Hence shearing strain γ_{rz} and $\gamma_{\theta z}$ are negligible. The stresses σ_z are small and can be neglected.
4. The deflections of the plate are due to the displacements of points in the middle surface of the plate in a direction normal to the non-deflected middle surface.

The differential equation of a standard plate theory for the displacement w(x, y) with an orthographic plate (here $E_x = E_y = E$) including residual stress [23]

$$\mathrm{D}\frac{\partial^4 w}{\partial x^4} + 2H\frac{\partial^4 w}{\partial x^2 \partial y^2} + D_1\frac{\partial^4 w}{\partial y^4} + Th\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right) = P \tag{6}$$

With D (flexural rigidity), $D_1 = D\nu$, $D_{xy} = Gh^3/12$ and $H = D_1 + 2D_{xy}$, the thickness h, Poisson's ratio ν and the homogeneous in-plane stress T. For a square membrane of area a^2, this equation can only be solved numerically by satisfying the following conditions for a rigidly clamped plate. In the boundary condition, deflection and slope of plate edges are zero.

$$\mathrm{w}\,(\mathrm{x} = 0, \mathrm{a\ y}) = 0, \quad \mathrm{w}\,(\mathrm{x} \cdot \mathrm{y} = 0, \mathrm{a}) = 0, \quad \left.\frac{\partial w}{\partial x}\right|_{x=0,\ a\cdot y} = 0, \quad \left.\frac{\partial w}{\partial y}\right|_{x\cdot y=0,a} = 0 \tag{7}$$

For closed form solution, the deflection of a square diaphragm of area A can be approximated by using a circular model with the same area [24]. This approximation offers a simpler calculation that comes with sufficient accuracy. For an isotropic circular, clamped diaphragm with existing plate deflection and subject to additional residual stress, the static deflection amplitude of pure plates (low tension) and pure membrane (high tension) can be expressed by

For Pure plate deflection,

$$w_{plate}(r) = \frac{pa^4}{64D}\left[1 - \left(\frac{r}{a}\right)^2\right]^2 \tag{8}$$

For Pure membrane deflection,

$$w_{membrane}(r) = \frac{p}{\sigma t}\left[\frac{a^2 - r^2}{4}\right] \tag{9}$$

According to membrane theory, the residual stresses (σ) on the diaphragm are said to dominate the mechanical response. The total membrane deflection can be calculated with superposition solution that incorporates both plate and membrane theory

$$\frac{1}{w_{total}} = \frac{1}{w_{plate}} = \frac{1}{w_{membrane}} \tag{10}$$

So that the total deflection becomes

$$w_{total} = \frac{Pa^4}{64D} \cdot \frac{1}{1 + \frac{Th}{16D}\left(a^2 - r^2\right)}\left[1 - \left(\frac{r}{a}\right)^2\right]^2 \tag{11}$$

In this equation, we can define corner stress, σ_c, which is a critical point for plate behavior.

$$\sigma_c = \frac{16D}{a^2 t} \tag{12}$$

In this membrane, the value of corner stress is 0.4 MPa. When the stress on the plate is smaller than corner stress, the plate deflection is dominant with flexural rigidity, but when the stress of membrane is larger than corner stress, the membrane deflection is dominant with residual stress.

Large Deflection Theory

When the out-of-plane deflection of the plate is comparable to the thickness, the strain and the curvature throughout the plate are no longer uniform. The deformation including the coupling between axial and transverse motion is geometrically nonlinear. Timoshenko [22] explained the deflection of thin film with an approximate solution, based on the energy method, which is also suitable for small deflection range.

$$\mathrm{w} = w_0\left(1 - {r^2}\big/{a^2}\right)^2 \tag{13}$$

Bert et al. [25], solve the constant (w_0) of equation into variable factor form including stretching body of membrane from its original erroneous form.

$$\frac{\mathrm{P}a^4}{Et^4} = 4.20\frac{1}{(1-\nu^2)}\frac{w}{t} + 1.58\frac{1}{(1-\nu^2)}\left(\frac{w^3}{t^3}\right) \tag{14}$$

Even though the above formulations for large deflections are quite useful, they do not take into account the built-in stress, which is commonly present in thin films. To explain the effects of built-in stress, we consider this equation;

$$\frac{\mathrm{P}a^4}{Et^4} = \frac{4\sigma a^2}{Et^2}\left(\frac{w}{t}\right) \tag{15}$$

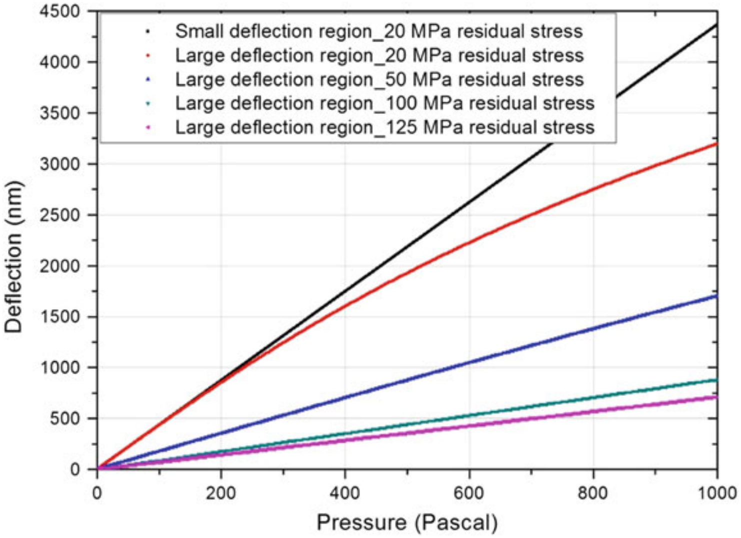


Fig. 4 Membrane deflections are corresponding to the applied pressures. For the radius of an assumed circular membrane is equal the area of square membrane with √((750)2/3.14). Also the silicon nitride is 0.5 μm thickness, 0.27 Poisson ratios, and 310 GPa Young's modulus

Rearranging the above equation with Eq. (15), one can obtain

$$\frac{\mathrm{P}a^4}{Et^4} = \left[4.20\frac{1}{(1-\nu^2)} + \frac{4\sigma a^2}{Et^2}\right]\frac{w}{t} + 1.58\frac{1}{(1-\nu^2)}\left(\frac{w^3}{t^3}\right) \tag{16}$$

Figure 4 demonstrates the difference of membrane deflection between small and large deflection region due to applied pressure. Also, the built-in stress in other words residual stress highly affects the membrane deflection. These differences generally cooperate with lower the overall sensitivity of a membrane sensor. From graph, the sensitivity of a diaphragm to mechanical response rapidly decreases depending on the residual stress. Therefore, it is of paramount importance to minimize if not eliminate the thin film stress to have a sensitive sensor.

2.2 Chemomechanical Transducer

Chemical sensors could be classified into several main fundamental transduction modes: (a) thermal, (b) mass, (c) electrochemical, and (d) optical. Each of these detection modes is related with features that are complementary rather than competitive with each other, and the search of an "ideal transducer" has continued [16]. In a biology, mechanical interactions play a crucial role for determining motility and adhesion on the cellular scale, and governing transportation and affinity on the molecular scale [17]. During the last two decades, advances in microelectromechanical systems (MEMS) have facilitated development of sensors that involve transduction of mechanical energy and rely heavily on mechanical phenomena.

Most MEMS sensors detect the mechanical movements and deformations of their micromachined components such as single clamped suspended beams like as cantilevers, double clamped suspended beams like as bridges, and suspended diaphragms. Due to atomic force microscopy, a single clamped suspended beam, which name it as nano or micro cantilever is the first device for chemo-mechanical transducer [18]. A small cantilever is the central element in many mechanical biosensors that offers several advantages such as label-free detection, high sensitivity, and real-time monitoring for detecting bio-molecule interaction.

The biomechanical signal transduction has been observed in a variety of binding assay, as summarized in Table 2. The sensor response to binding events can be attributed to a variety of mechanisms. In case of DNA hybridization, one of the mechanisms causing stress on sensor surface is electrostatic force due to the highly negative charge of the molecule. When short length ssDNA chains are involved, under 15-mer, the negative charge on the oligonucleotide backbone produce electrostatic force, and this is the dominant factor. By the action of hybridization, formation of double-stranded structure, this force becomes even greater [12]. When examining this situation from a thermodynamic point of view, long chain ssDNA behave as a worm-like chain (WLC), effectively and double-stranded DNA are more like a rod. The conformational change produced by hybridization leads to reduced steric effect and induces tensile surface stress [19]. When we consider binding phenomena involved in antigen–antibody [19, 20] or aptamer-protein [21, 22] reactions, in most of these cases, there is a net charge effect which could lead to electrostatic force, however the steric effect involved is more dominant. These molecules are much bigger and have stretched conformations; also the net charge further enhances the steric effect after binding occurs, leading to a compressive surface stress.

2.2.1 Cantilever Transducer

Atomic force microscopy was developed in 1986 from Gerd binning. It is able to detect atomic resolution surface with non-contact method by using van der Waals force between stylus and surface atom. Commercialization of AFM has initiated the development of micro-cantilevers for the surface stress sensors. In 1994, instead of the tip Ibach [18] utilize the whole body of cantilever beam as a probe to detect photo-thermal energy in the order of Pico joules (10^{-12} J). The cantilever-based sensors comprehend the transduction of chemical and physical interaction occurring on the surface into a mechanical deflection or resonant frequency shift of the cantilever depending on the respective working principle. The micro-cantilever generate a great deal of interest after it shows promising potential in the area of applications such as chemical/physical analysis and the detection of the freshness of food, charge, surface stress, IR radiation and heat flux. Therefore, some companies

Table 2 A variety of binding assay using chemo-mechanical transduction

Type	Material	Dimension (μm)	Molecule (concentration)	Deflection (nm)	Stress	Reason	Paper reference
Membrane	Silicon	Thickness = 2 Circle R = 400	Dodecanethiol (98 %, 400 μl)	33	Compressive	Surface free energy has been reduced	APL, 89, 173123, 2006
Membrane	Parylene	0.5 × 300 × 300	10 mJ/m^2	65	Compressive	Compressive stress	Sensors and actuators B, 115, 2006, 494–502
Membrane	PDMS	2 × 500 × 500	16mer DNA (1 μM)	20 (6 fF)	Compressive	Compressive stress	Lab chip, 2008, 8, 932–937
Membrane	PDMS	2 × 500 × 500	Thrombin (500 nM)	18 fF	Compressive		Lab chip, 2008, 8, 932–937
Cantilever	Silicon	1 × 350 × 35	Glucose oxidase-glucose (25 mg/ml)	155	Tensile		Anal. Chem. 2004, 76, 292–297
Cantilever	Silicon	1 × 500 × 100	12mer DNA (2 μM, 400 nM, 80 nM)	22, 15, 3	Compressive	Charge density from sugar-phosphate backbone	Science, 2000, 288, 316–318
Cantilever	Silicon	1 × 500 × 100	12 mer DNA (1 μM)	8			PNAS, 2002, vol. 99. no. 15, 9783–9788
Cantilever	Silicon	1 × 350 × 30	MDTP (1 mM)	250	Tensile	π–π interaction of aromatic rings of pyridine in MDTP	Mendeleev Commun., 2010, 20, 329–331
Cantilever	Silicon	2 × 350 × 30	$CuCl_2 \cdot 2H_2O$ (1 mM)	600 @ 80 min	Compressive	Copper chelation	Mendeleev Commun., 2010, 20, 329–332
Cantilever	Silicon	3 × 350 × 30	Histidine (1 mM)	250 @ 70 min	Compressive	Repulsions of electronegative chloride atoms	Mendeleev Commun., 2010, 20, 329–333
Cantilever		0.65 × 600 × 20	Free prostate specific antigen (fPSA)	40 @ 0.5 ng/ml, 70 @ 1 ng/ml, 110 @ 10 ng/ml			MRS Bulletin, vol. 34, June 2009

such as Concentris [23] and Veeco [24]. Especially, Fritz demonstrate biomolecular recognition can be translated into nanomechanical response [25]. In this paper [2], the specific DNA hybridization is shown to take place on the surface of cantilever beam, surface stress that resulted in the deflection of cantilever inducing for differential sensing they used an array of sensors which are differently functionalized, which implicate a real molecular recognition signal rejecting common nonspecific adsorption of an array of cantilevers. Also, single mismatch oligonucleotides with hybridization of 12-mer complementary oligonucleotides are clearly detected implying that this chemo-mechanical transduction has improved the selectivity of molecular binding. The specificity of this sensing layer determines the application with the difference of the chemical end-group present at the mono layer/gas/liquid interface.

The sensing layers of the functionalized SAM are both receptive sensitized to react with target molecules and responsive by allowing the transduction of the surface stress to the cantilever beam [26–28]. The micro-cantilever has been observed in a variety of molecule binding applications such as detection of mercury gas [29], humidity sensing [30], explosive gas [31, 32], volatile organic compound (VOC) sensing [3, 33, 34], cell [35], and cocaine [36]. The nematic-states of molecule crystal induce the stress on the surface of cantilever beam. This surface stress owing to electrostatic repulsion or attraction, steric interactions, hydration and entropic effects make the deflection. Each parameter of these applications becomes more relevant according to the functionality of the particular device.

However, the micro-cantilevers are extreme flexible in a liquid that cause the undesirable deflection and deformation. Also, most cantilevers utilize the bulky optical measurement system that has the detecting limits of non-opaque analyte, and the back-side of cantilevers is vulnerable to nonspecific adsorption. Thundat et al. [29] suggest oscillating mode for alternative measurement system depict in Fig. 5. The mechanism has potential of system miniaturization, but this technology has also limitations such as analyte viscosity, thermal noise and low sensitivity.

2.2.2 Membrane Transducer

Membrane base chemo-mechanical transducers overcome the challenge of cantilever as well as enhance the reliability. Membrane is the another configuration of structure member for detecting surface stress changes using silicon polymer bimorph to sensing VOC and humidity [26]. Membranes offer many advantages, some of which include easy electronic readout and sample isolation from detection systems.

Satyanarayana et al. [12] publish the parylene micro membrane capacitive sensor for chemical and biological detection. This paper reports the design parameter with FE analysis simulation which demonstrates sensor response highly affected by gold coverage ratio comparing membrane area. The salient advantages of the membrane-based chemomechanical transducer: (1) is label-free; (2) is a universal platform

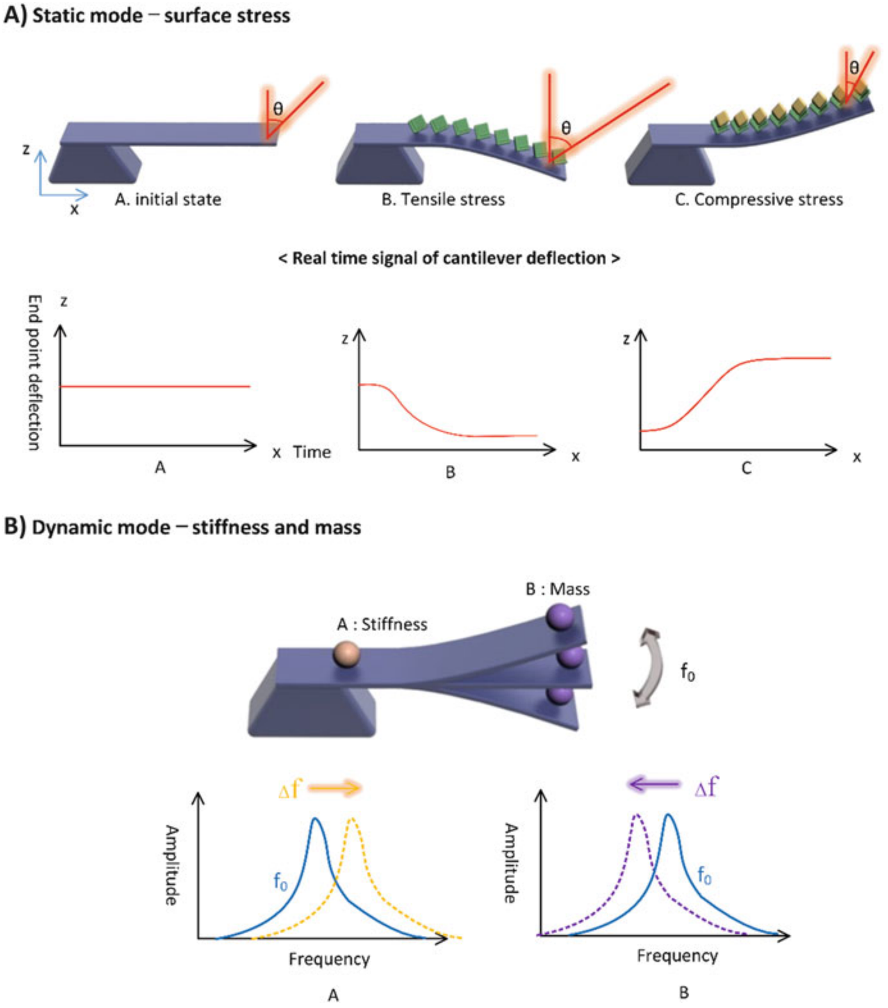
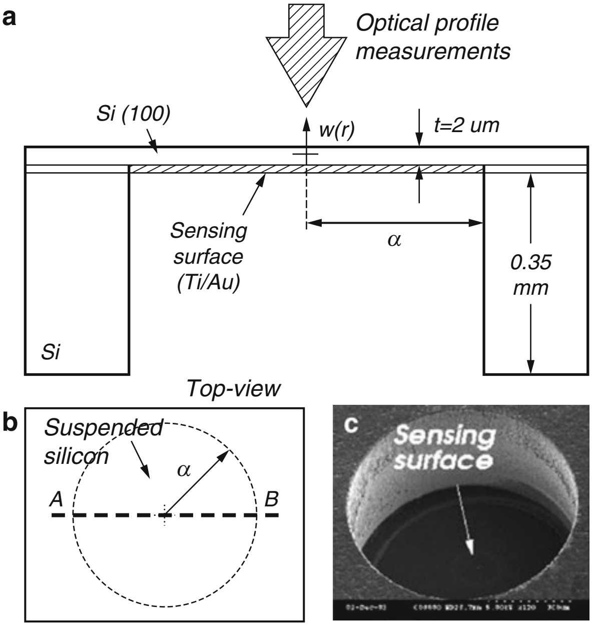


Fig. 5 Main operation modes of nanomechanical sensors: static mode. When the static mode is used, the measurement of the full cantilever profile, or at least a referenced z position, is needed for the end-point assay. The dynamic mode can follow changes in the added mass (i) and also changes in the stiffness (ii). When the adsorption is restricted either to the free-end or to the clamped end, both contributions can be disentangled

suitable for both chemical and biological sensing; (3) uses electronic readout; (4) has integrated microfluidics for addressing individual sensors on the chip; (5) is capable of handling both liquid and gas sample; (6) is made using standard low temperature micro-fabrication processes; (7) can readily be scaled and multiplexed. The sensor

Fig. 6 (100) silicon plate device for chemomechanical transducer sensing molecular interaction: (**a**) device cross section used for optical interferometry measurements, (**b**) plate deflection profiles form line A–B, and (**c**) scanning electron microscopy of exposed sensing surface before Ti/Au sputtering



response to organic vapors like isopropyl alcohol and toluene are measured. They fabricate an array of sensor to spontaneously detect various gas molecules of optical images of the parylene membrane sensor. Although a novel membrane-based nanomechanical sensor for chemo-mechanical transduction uses the parylene material which has a low mechanical stiffness of polymers to increase the sensitivity, the parylene is very susceptible of environmental noises such as fluidic pressure and high temperature.

The Charles Stark Draper Laboratory group shows the micromachined silicon plate membrane sensor for sensing molecular interactions shown in Fig. 6 [7]. They mentions that the ratio of deflection to a certain surface stress for membrane structures is smaller than cantilever structures by a factor of about 10–100 times [1, 27, 28], but the electronic displacement detection resolution exceeds 100 times over that of reported optical detection techniques [29, 30]. A thin suspended crystalline silicon circular plate is fabricated for surface stress sensor and the surface stress changes associated with vapor phase chemisorption of a self-assembled monolayer with 1-dodecanethiol vapor shown in Fig. 7. The isolated face of the suspended silicon plate serves as the sensing surface treated with a receptor layer sensitive to a target molecule such as gold, antibody and aptamer. The chemisorption of an Alkanethiol on the gold surface make 361 nm deformation after 300 s of the chemical evaporation which indicate compressive 0.72 ± 0.02 N m^{-1} surface stress change. However, the silicon material is vulnerable to crack the surface of thin film.

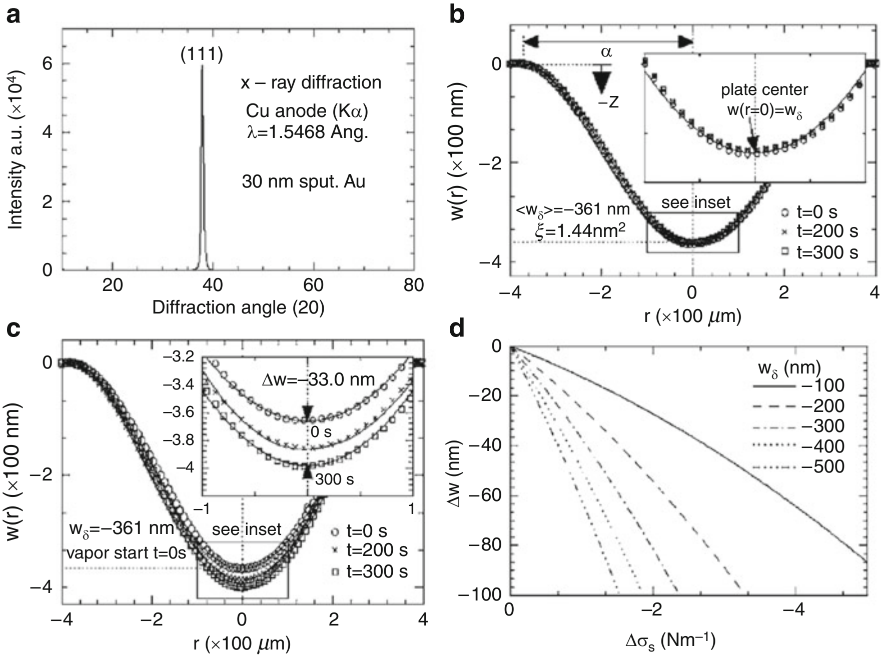


Fig. 7 Results of deformation: (**a**) X-ray diffraction scan 30 nm sputtered Au layer (with 8 nm Ti layer), (**b**) measured plate bending At t = 0, 200 and 300 s (10 s intervals) prior to vapor exposure, (**c**) bending during 1-dodecanethiol vapor exposure and (**d**) calculated Δw as a function of Δσs and wδ

2.3 Surface Sensitization Technique

Chemical and biological sensors require specific molecule for capturing target molecule which is called as receptor. The receptor layer on a biosensor provides specific binding sites for the target analytes such as molecules, proteins and cells. The receptor layers determine the sensor's characteristics such as selectivity, sensitivity and repeatability. The selective layer can be designed utilizing principles of molecular interaction; recognition, for example, DNA hybridization, antigen–antibody binding and aptamer epitope binding. Also, the sensing layers on the solid surface of biosensor can be achieved using molecular self-assembly monolayers for specific surface selectivity. Since the technology of surface functionalization, specificity and selectivity of biological receptors are too complex, sensor researchers have to understand molecule interaction of interesting target molecule and optimize the molecule structure of receptor.

2.3.1 Surface Sensitization

The specificity and sensitivity of biosensor systems are highly relative with the interfacial properties where bioactive species are immobilized. The design of molecular structures includes both the immobilization of the bio-receptor itself and the overall chemical preparation of the transducer surface. Also, the inertness of the surface limits the nonspecific adsorptions which induce the background noise of the sensor. Lastly, a robust interface improves the stability and the reliability of biosensor system. The optimum elaboration and the use of biosensor require the background knowledge of the surface preparation and immobilization process of bio-species which is a detailed overview of the individual steps of surface modification; functionalization and immobilization of bio-receptors onto solid supports. Furthermore, the strength of intermolecular bonds is main factor for surface stress resulted in membrane deflection. The strength of bonds varies with different chemical bonds [31]. Table 3 lists the categorization of chemical bond based on their strength.

2.3.2 Molecular Immobilization on Solid Surface

In the 1970s, biological analysis techniques such as ELISA [32], Southern blot [33] and Western blot [34] based on the detection of solid supported biomolecular interactions are developed and still remain widely used. The principle is the specific recognition between a molecule in solution and a receptor molecule immobilized onto a solid support. This immobilization is concentrated on the specific area of surface which increases the receptor molecular density. According to the Langmuir isotherm model, the incensement of the density makes the detection easier and

Table 3 The categorization of chemical bonds

Strong	Covalent bonds & antibonding	**Sigma bonds**: 3c-2e, bent bond, 3c-4e(Hydrogen bond, Dihydrogen bond, Agostic interaction), 4c-2e **Pi bonds**: π backbonding, Conjugation, Hyperconjugation, Aromaticity, Metal aromaticity **Delta bond**: Quadruple bond, Quintuple bond, Sextuple bond Dipolar bond, Facticity
	Ionic bonds	Cation–pi interaction, Salt bridge
	Metallic bonds	Metal aromaticity
Weak	Hydrogen bond	Dihydrogen bond, Dihydrogen complex, Low-barrier hydrogen bond, Symmetric hydrogen bond, Hydrophile
	Other noncovalent	Van der Waals force, Mechanical bond, Halogen bond, Aurophilicity, Intercalation, Stacking, Entropic force, Chemical polarity
Other	Disulfide bond, Peptide bond, Phosphodiester bond	

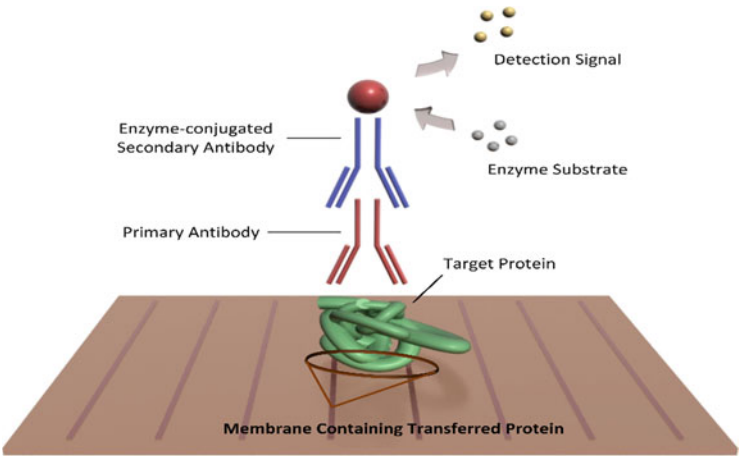

Detection Signal
Enzyme-conjugated Secondary Antibody
Enzyme Substrate
Primary Antibody
Target Protein
Membrane Containing Transferred Protein

Fig. 8 Detection in Western blots. In the original blotting techniques, the target is immobilized on the support with adsorption

more sensitive on the surface and the recognition events to be detected are precisely localized. In conventional blotting techniques in Fig. 8, target molecule is immobilized on a solid to find a group of probe molecules in solution. This process is conducted in parallel tests. The number of individual experiments corresponds to the number of available probe molecules. By developing of detection technology, known probes attach a multiple surface to contact target solution as a single test, which is called as arrays. There are a few materials to directly immobilize biological molecules onto native solid surface. For satisfying specificities of biosensor application, particular surface chemistries are required to control the immobilization in order to maintain biological activity and minimize nonspecific adsorption. Thin organic films on mineral surface are widely used to foment the immobilization of biomolecules to solid surface [35]. For appropriate immobilization, the organic films have some characteristics such as reproducibility, homogeneity, thermal and chemical stability and chemical reaction performed during the immobilization or functionalization process. Various organic molecules and chemical reactions are included depending on the characteristics of the solid surfaces such as hydroxyl-terminated surfaces by silane coupling agents and gold surfaces are modified by adsorption of alkanethiolate reagents.

2.3.3 Aptamer

Antibody is the most popular receptor for molecule recognition of a wide range of applications, so antibodies are routinely used for clinics and also contribute the advancement of most diagnostic test. The development of the systematic evolution of ligands by exponential enrichment (SELEX) process, however, made possible an

Aptamer for next generation of receptor with high affinity, low cost, and specificity. Although aptamers are mimic properties of antibodies in a variety of diagnostic formats, aptamer can detect small molecule such as ions, gases, and small particle like as cocaine which cannot detect with antibody receptor. Aptamers (synthetic antibodies) are (stable) single-stranded DNA, RNA, or peptide molecules capable of binding to its target antigen with high affinity and specificity. Aptamer have been discovered against a wide variety of target molecules from small organics to large proteins [36]. Compared with the bellwether antibody technology, aptamer research is still in its infancy, but it is progressing at a fast pace [37].

Aptamers that come out of a SELEX process are certain sequences containing the fixed sequences that were included to aid the amplification process. The SELEX process begins with a random sequence library obtained from combinatorial chemical synthesis of DNA as depicted in Fig. 9. The diversity of a library is related with the number of randomized nucleotide positions, which normally contain a 40-nucleotide random region represented by 1.2 × 1,024 individual sequences (420 = 1.2 × 1,024). The success of finding unique and rare aptamer that interact with a target parallels the diversity of the libraries used. These include peptide libraries used for phage display as well as the libraries made up of small organic molecules.

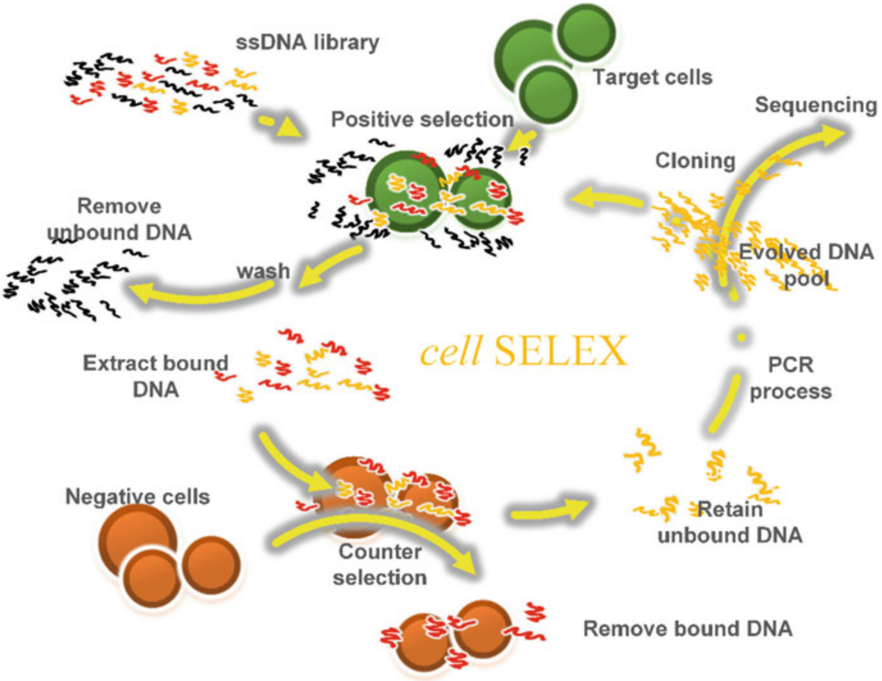


Fig. 9 The systematic evolution of ligands by exponential enrichment (SELEX)

Table 4 Benefits of aptamers over antibodies

	Aptamer	Antibody
Affinity	nM–pM	nM–pM
Specificity	High	High
Manufacturing	*In vitro chemical process*	*In vivo biological system*
Cost	*Low*	*High*
Target molecule	*Wide: ions, small organic molecules, proteins, whole cells, toxic molecules, etc.*	*Narrow: only immunogenic compounds*
Batch to batch variation	*No*	*Big problem; Significant*
Chemical modification	*Easy*	*Limited*
Thermal rigidity	*Small and reversible*	*Irreversible*
Life time	*Very long*	*Limited, storage on refrigerator*

The binding analysis of enriching populations of the target molecules is carried to determine the progress of the enrichment of high-affinity binders. Once affinity saturation is achieved after several rounds of selection/amplification, the enriched library is cloned and sequenced to obtain the sequence information of each member. Usually, the majority of individual sequences, over 90 %, in an enriched library are "winners", aptamers that bind to the target used for selection. Once the sequence information has been obtained through the SELEX process, the desired aptamer can be readily produced by chemical synthesis with low cost. For a given target molecule, these are two kinds aptamer, DNA and RNA. Since DNA lacks the $2'$ hydroxyl group of the RNA, they are quite difference in three-dimensional structure. In Table 4, I summarize the benefits of aptamers compared with antibodies and the italicized rows detail special advantages of aptamers to recognize target molecule for transducer.

Therefore, aptamers have numerous advantages over antibodies to be useful tools in analytical, diagnostics and therapeutic applications. The most important characteristic of aptamers is its ability to bind their target molecules with high specificity.

2.3.4 Dissociation Constant

The binding process is initiated with long-range electrostatic interactions, leading to the approach and combining of the ligand to appropriate intermolecular interactions of the complementary binding surfaces. During the molecular interaction, reversible and non-covalent binding are indispensable step in the most chemical and biological processes. Since binding interactions are so crucial factor for understanding of molecular interaction, binding processes can be studied at the structural, equilibrium-thermodynamic, and kinetic levels. Normally, characterizing the strength of the dynamic binding equilibrium determine dissociation constant, KD, which is experimentally determined quantity with units of molarity.

For example, in two state reversible interactions, dissociation constant equals between ligand and receptor concentration that gives half-maximal binding.

In a simple two state model, binding equilibrium model can be used to approximate most simple, reversible and non-covalent binding interactions [38].

$$A_xB_y \rightleftharpoons xA + yB \tag{17}$$

in which a complex A_xB_y break down into x A subunits and y B subunits. The dissociation constant KD, with units of molarity, is defined as

$$K_d = \frac{[A]^x \times [b]^y}{\left[A_xB_y\right]} \tag{18}$$

where [A], [B], and [AxBy] are the concentrations of A, B, and the complex AxBy, respectively.

Acknowledgments This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Berger R, Delamarche E, Lang HP, Gerber C, Gimzewski JK, Meyer E, Güntherodt H-J. Surface stress in the self-assembly of alkanethiols on gold. Science. 1997;276(27):2021–4.
2. Fritz J, Baller MK, Lang HP, Rothuizen H, Vettiger P, Meyer E, Güntherodt H-J, Gerber C, Gimzewski JK. Translating biomolecular recognition into nanomechanics. Science. 2000;288:316–8.
3. Yue M, Stachowiak JC, Lin H, Datar R, Cote R, Majumdar A. Label-free protein recognition two-dimensional array using nanomechanical sensors. Nano Lett. 2008;8(2):520–4.
4. Sang S, Witte H. A novel PDMS micro membrane biosensor based on the analysis of surface stress. Biosens Bioelectron. 2010;25(11):2420–4.
5. Datar R, Kim S, Jeon S, Hesketh P, Manalis S, Boisen A, Thundat T. Cantilever sensors: nanomechanical tools for diagnostics. In: MRS Bulletin. 2009. p. 449–54.
6. Gil-Santos E, et al. Mass sensing based on deterministic and stochastic responses of elastically coupled nanocantilevers. Nano Lett. 2009;9(12):4122–7.
7. Carlen ET, et al. Micromachined silicon plates for sensing molecular interactions. Appl Phys Lett. 2006;89(17):173123.
8. Wu Z, Choudhury K, Griffiths HR, Xu J, Ma X. A novel silicon membrane-based biosensing platform using distributive sensing strategy and artificial neural networks for feature analysis. Biomed Microdevices. 2012;14(1):83–93.
9. Zapata AM, Carlen ET, Kim ES, Hsiao J, Traviglia D, Weinberg MS. Biomolecular sensing using surface micromachined silicon plates. In: The 14th international conference on solid-state sensors, actuators and microsystems. Lyon, France; 2007. p. 831–4.
10. Lim S-HS, et al. Nano-chemo-mechanical sensor array platform for high-throughput chemical analysis. Sens Actuators B. 2006;119(2):466–74.
11. Xu T, et al. Micro-machined piezoelectric membrane-based immunosensor array. Biosens Bioelectron. 2008;24(4):638–43.
12. Satyanarayana S, McCormick DT, Majumdar A. Parylene micro membrane capacitive sensor array for chemical and biological sensing. Sens Actuators B. 2006;115(1):494–502.

13. Cha M, et al. Biomolecular detection with a thin membrane transducer. Lab Chip. 2008;8(6):932–7.
14. Kang TJ, Lim D-K, Nam J-M, Kim YH. Multifunctional nanocomposite membrane for chemomechanical transducer. Sens Actuators B. 2010;147:691–6.
15. Lang HP, et al. Nanomechanics from atomic resolution to molecular recognition based on atomic force microscopy technology. Nanotechnology. 2002;13(5):R29–36.
16. Heine V, Marks LD. Competition between pair-wise and volume forces at noble metal surfaces. Surf Sci. 1986;165:65.
17. Godin M. Surface stress, kinetics, and structure of alkanethiol self-assembled monolayers. In: Physics. Montreal, QC: McGill University; 2004.
18. Ibach H. The role of surface stress in reconstruction, epitaxial growth and stabilization of mesoscopic structures. Surf Sci Rep. 1997;29(5–6):195–263.
19. Shuttleworth R. The surface tension of solids. In: Proceedings of the physical society. Section A; 1950.
20. Finot M, Suresh S. Small and large deformation of thick and thin-film multi-layers: effects of layer geometry, plasticity and compositional gradients. J Mech Phys Solids. 1996;44(5): 683–721.
21. Ko WH, Qiang W. Touch mode capacitive pressure sensors for industrial applications. In: Tenth annual international workshop on micro electro mechanical systems, 1997. MEMS '97, Proceedings, IEEE; 1997.
22. Timoshenko S. Theory of plates and shells. McGraw-Hill Classic Textbook Reissue; 1959.
23. Szilard R. Theory and analysis of plates: classical and numerical methods. Englewood Cliffs: Prentice-Hall; 1974.
24. Maisano J. More advanced models for silicon condenser microphones. In: 92nd convention of the audio engineering society. Vienna, Austria; 1992.
25. Bert CW, Martindale JL. An accurate, simplified method for analyzing thin plates undergoing large deflections. Am Inst Aeronaut Astronaut (AIAA) J. 1988;26(2):235–41.
26. Chatzandroulis S, et al. Capacitive-type chemical sensors using thin silicon/polymer bimorph membranes. Sens Actuators B. 2004;103(1–2):392–6.
27. Wu G, et al. Bioassay of prostate-specific antigen (PSA) using microcantilevers. Nat Biotechnol. 2001;19(9):856–60.
28. Pei J, Tian F, Thundat T. Glucose biosensor based on the microcantilever. Anal Chem. 2004;76(2):292–7.
29. Rugar D, Mamin HJ, Guethner P. Improved fiber-optic interferometer for atomic force microscopy. Appl Phys Lett. 1989;55(25):2588–90.
30. Yaralioglu GG, et al. Analysis and design of an interdigital cantilever as a displacement sensor. J Appl Phys. 1998;83(12):7405–15.
31. Sang S. An approach to the design of surface stress-based PDMS micro-membrane biosensors-concept, numerical simulations and prototypes. Ilmenau: University of Bibliothek; 2010.
32. Engvall E, Perlmann P. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. Immunochemistry. 1971;8(9):871–4.
33. Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol. 1975;98(3):503–17.
34. Burnette WN. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate–polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal Biochem. 1981;112(2):195–203.
35. Dugas V, Elaissari A, Chevalier Y. Surface sensitization techniques and recognition receptors immobilization on biosensors and microarrays. In: Zourob M, editor. Recognition receptors in biosensors. New York: Springer; 2010. p. 47–134.
36. www.aptagen.com/home.aspx. Aptamer. 2013 [cited 2013 19th Oct.].
37. Jayasena SD. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. Clin Chem. 1999;45(9):1628–50.
38. WIKIPEDIA. Dissociation constant.

Fully Printable Organic Thin-Film Transistor Technology for Sensor Transducer

Xiaojun Guo, Linrun Feng, Wei Tang, Cheng Jiang, Jiaqing Zhao, and Wenjiang Liu

Abstract For many of future sensor applications, the sensors are required to be of low cost, easy production and to be able to work with multi-physics or bio/chemistry signals, and provide large area, flexible or comfortable surface coverage. All these bring challenges to the current silicon based manufacturing technology. This chapter introduces a hybrid integration concept combining the advantages of both the printed electronics and silicon technologies to address these issues. A fully printable low voltage organic thin-film transistor (OTFT) technology is developed for making the transducer in the hybrid sensor systems. A simple application example for pH sensor tag is demonstrated. This OTFT technology would provide a promising platform for developing general low cost and disposable multi-function integrated sensor systems.

1 Introduction

Sensors are becoming more and more important as part of global issue solutions for environment monitoring, digital health and Internet of Things [1]. Several organizations forecast the consumed sensor volumes per year to exceed trillions by 2022 [2]. Many of the sensors will be of shorter product life, and need to work with multi-physics or bio/chemistry signals. Therefore, for future sensor development, manufacturing processes with extremely low cost, high throughput and short production cycles are required. It is also important to have greater freedom to be able to easily integrate various kinds of functional materials, and realize "manufacturing-on-demand" for various different applications. Moreover, there are many emerging applications requiring sensors to be able to cover large area, and being flexible or comfortable. All these bring challenges to the current silicon based manufacturing technology.

Recently, the technology of printable electronics is attracting worldly-wide interests. With this technology, it is possible to integrate a wide range of organic, inorganic, nanostructure functional materials for electronics, battery, energy har-

X. Guo (✉) • L. Feng • W. Tang • C. Jiang • J. Zhao • W. Liu
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
e-mail: x.guo@sjtu.edu.cn

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_3

Table 1 Comparison with silicon microelectronics and printed electronics

	Si microelectronics	Printed electronics
Fab start-up and maintaining cost	Extremely high	Very low
Process condition	High temperature (>1,000 °C), high vacuum	Low temperature (<150 °C), ambient condition
Materials	Si, SiO_2/SiN_x	Various substrates, semiconductors and dielectrics
Device and integration structure	Almost fixed, multi-layer and high density	Versatile, but low density
Device performance	Excellent	Poor
Form factor	Rigid, small area	Flexible, large area
Production mode	Long period, filling up the fab capacity	Short period, "manufacturing as demanded"

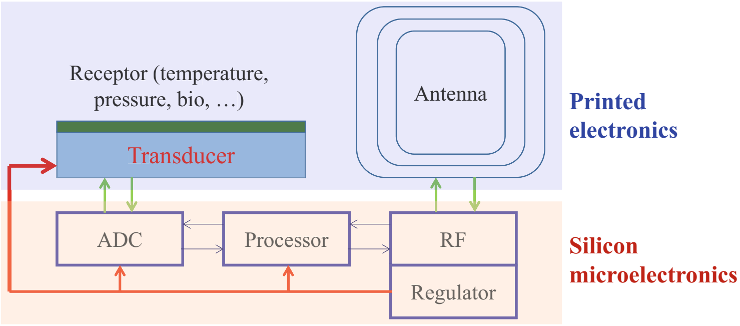


Fig. 1 A hybrid integration concept composed a printed transducer circuit and a common read-out and signal processing hardware or chip is proposed for future sensor development

vesting and sensor/display interfacing devices through simple printing or coating processes.

As shown in Table 1, despite of considerable compromise in the device performance and integration density, this technology owns several competitive advantages compared to silicon microelectronics, including low cost and high throughput manufacturing, being compatible with arbitrary substrates such as plastic and paper, and enabling easy hybrid integration of multi-functional materials. These features well match the above requirements. Therefore, it could be ideal supplement to conventional silicon technologies for future sensor development.

Figure 1 depicts a of a hybrid integration concept for future sensor development. For different types of sensing, a transducer circuit can be designed based on printed thin-film transistor (TFT) technology to convert the sensed signal to a standard voltage output. A common read-out and signal processing hardware or chip composed analog-to-digital converter, processor, radio frequency (RF) transceiver and regulator can thus be implemented for these different sensors. With printing processes, different sensing receptor materials and the antenna can be easily integrated with the transducer circuit. In such an integrated architecture, the printed TFT technology is important, and the following parts of this chapter focus on how to achieve fully printable low voltage transistors for the sensor applications.

2 Printable Thin-Film Transistors

Different from the silicon metal oxide semiconductor field effect transistors (MOSFET), thin-film transistors (TFTs) could be made on different substrates like glass, plastic, and paper, etc., instead of silicon substrate. Various device structures might also be used for TFTs, compared to the only one for silicon MOSFETs, as illustrated in Fig. 2. For printable TFTs, various choices of solution processable semiconductor materials are available, such as organic semiconductors [3], metal oxide [4], carbon nanotubes [5] and so on. The organic semiconductor is chosen for its superior mechanical flexibility, very low temperature with fast processing, great potential

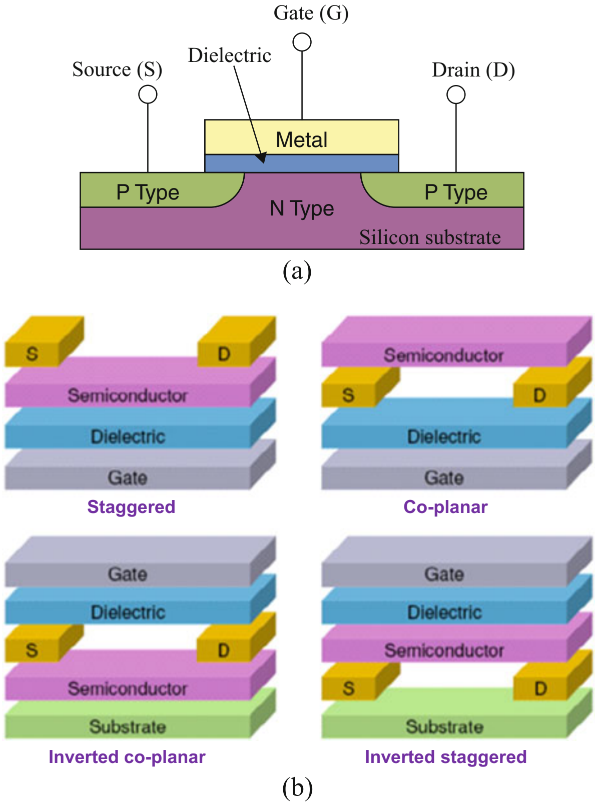


Fig. 2 Device structures used for (**a**) silicon MOSFETs and (**b**) thin film transistors

for further performance improvement and being functionalized by tailoring the chemical structure. In the past, lots of research has been put on developing high performance soluble organic small molecule and polymer semiconductors by via molecular design/engineering, as well as solution processed dielectric materials, conducting materials and interfacial materials [6]. The rapid progress of materials opens up the possibility of fabricating organic transistors and circuits with printing processes [7]. However, as shown in Fig. 3, the operation voltage currently with a typical value of tens and even more than 100 V, needs to be greatly reduced, since most of the targeted sensor applications are portable or mobile, and normally powered by battery or a.c. fields, and low voltage operated OTFTs are also required for interfacing silicon chips for the proposed integration architecture in Fig. 1.

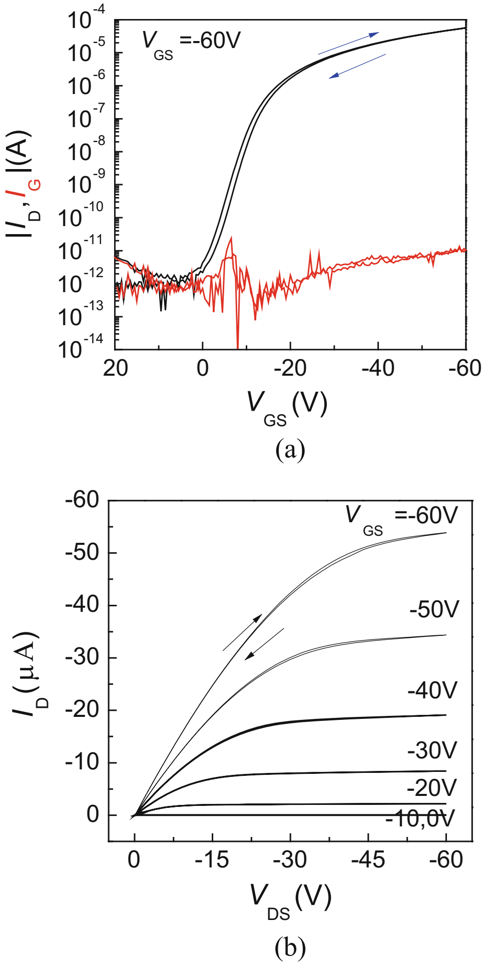


Fig. 3 Typical electrical characteristics of OTFTs: (**a**) the transfer characteristics (I_D-V_{GS}) and (**b**) the output characteristics (I_D-V_{DS})

3 Challenges for Low Operation Voltage Printable Transistors

The required operation voltage for an OTFT is determined by the subthreshold swing (SS), given by:

$$SS = \frac{kT\text{In}10}{q}\left(\frac{q^2 N_\text{t}}{C_\text{G}} + 1\right) \tag{1}$$

where q is the elementary charge, k is the Bolzmann's constant, T is the absolute temperature, N_t is the interface trap density, and C_G is the gate dielectric capacitance per unit area.

According to Eq. (1), a straightforward way for low voltage is to increase the effective gate insulator capacitance by dramatically decreasing the gate insulator layer thickness or using high dielectric constant (high-k) gate dielectric materials [8, 9]. However, many of these reported dielectric films in the literature are not based on solution processes or require high temperature post-annealing. Even for some other work with low temperature solution processed gate dielectric layers, the utilization of vacuum deposited channel layer and electrodes didn't consider the process orthogonality among the channel, the gate dielectric, which, however, is of utmost importance for realizing fully printable OTFTs. Besides the process orthogonality issues, for OTFTs over large area substrates, especially based on solution processed multi-layer integration, the gate insulator layer need to be thick enough (at least a few hundred nanometer thick) for the concern of device reliability and yield. Therefore, lots of work on using high-k gate dielectric material or even some solid electrolyte so that a thick dielectric layer can be used to realize low voltage [10]. The issue for high-k gate dielectric layer is that high-k dielectric materials could be unfavorable for carrier transport due to the broadening of the trap density of states at the semiconductor/dielectric interface by the formed dipole disorder [11]. The solid electrolyte can even induce electrochemical doping of the channel due to the presence of ions affect, causing reliability issues and slow operation speed.

As summarized in Fig. 4, for the conventional way of enlarging the gate dielectric capacitance to reduce the operation voltage, there is trade-off between processability and performance requirements.

4 Fully Printable Low Voltage OTFTs

Based on Eq. (1), high operation voltage with conventional organic transistors is also related to large trap density of states (DOS) at the semiconductor/dielectric interface. Therefore, if the interface trap DOS was minimized, low voltage OTFTs could be achieved without needing large gate dielectric capacitance. An approach

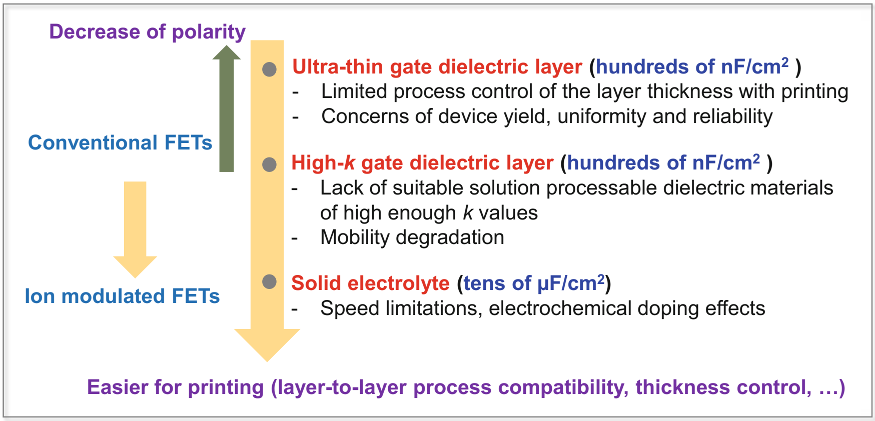

Decrease of polarity
Conventional FETs
Ion modulated FETs
Ultra-thin gate dielectric layer (hundreds of nF/cm²)
- Limited process control of the layer thickness with printing
- Concerns of device yield, uniformity and reliability
High-k gate dielectric layer (hundreds of nF/cm²)
- Lack of suitable solution processable dielectric materials of high enough k values
- Mobility degradation
Solid electrolyte (tens of μF/cm²)
- Speed limitations, electrochemical doping effects
Easier for printing (layer-to-layer process compatibility, thickness control, …)

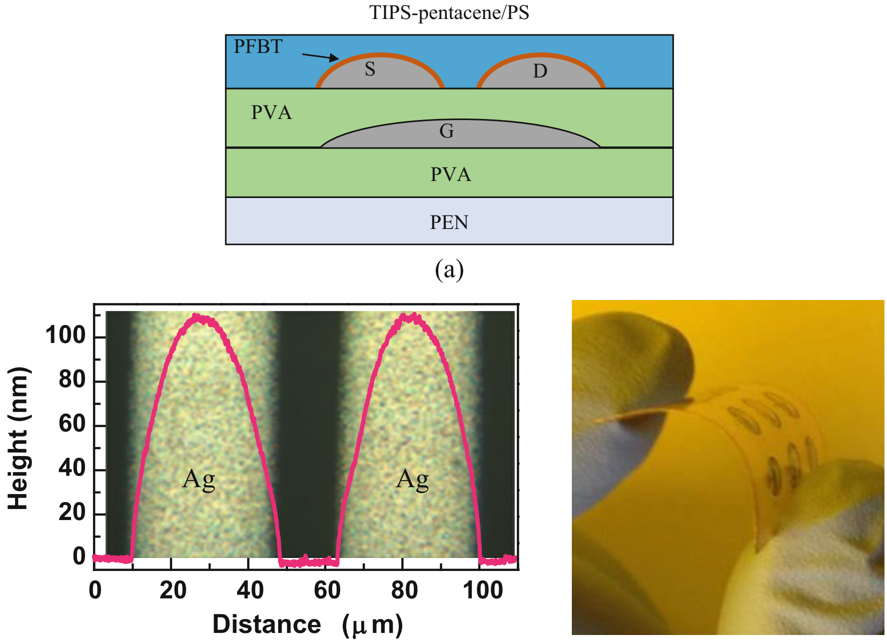
Fig. 4 Summary of the tradeoff between processability and performance requirements for enlarging the gate dielectric capacitance to reduce the operation voltage of printable OTFTs

Table 2 Comparison of the gate dielectric capacitance values for devices using different methods to achieve the similar subthreshold swing (*SS*)

Methods	*SS* (mV/decade)	C_G (nF/cm^2)
Ultra-thin gate dielectric [8]	∼100	600
High-*k* gate dielectric [9]	∼100	330
Channel engineering [12]	∼100	12.2

for that is through using blend of small molecule organic semiconductor with polymer binder in solution to form a low trap interface on a smooth polymer dielectric surface in a self-organization way [12]. Based on this approach, low voltage OTFTs can be realized with a very small gate dielectric capacitance of 12.2 nF/cm^2 [12], while in previous work using ultra-thin or high-k dielectric, gate dielectric capacitance of a few hundred nF/cm^2 is needed to achieve similar low operation voltages [8, 9], as compared in Table 2.

Figure 5a illustrates the device structure for the fully printable low voltage OTFTs in inverted coplanar (bottom-gate bottom-contact, BGBC) configuration fabricated on 125 μm thick polyethylene naphthalate (PEN) foil. Among various OTFT device structures, the BGBC structure is a preferred choice [13]. In this structure, the process-sensitive semiconductor layer is finally deposited for avoiding any undesired damages, and both the surface properties of the source/drain electrodes and the gate dielectric can be tailed to form controllable interfaces. More importantly, based on BGBC structure, the gate dielectric can provide a smooth surface to form interface with the semiconductor, which is beneficial for reducing the interface trap DOS. A metal-organic precursor type silver (Ag) ink (Jet-600C, Hisense Electronics, Kunshan, China) was printed with a piezoelectric inkjet printer (Dimatix, DMP 2831) using a 10 pL cartridge for the gate, source



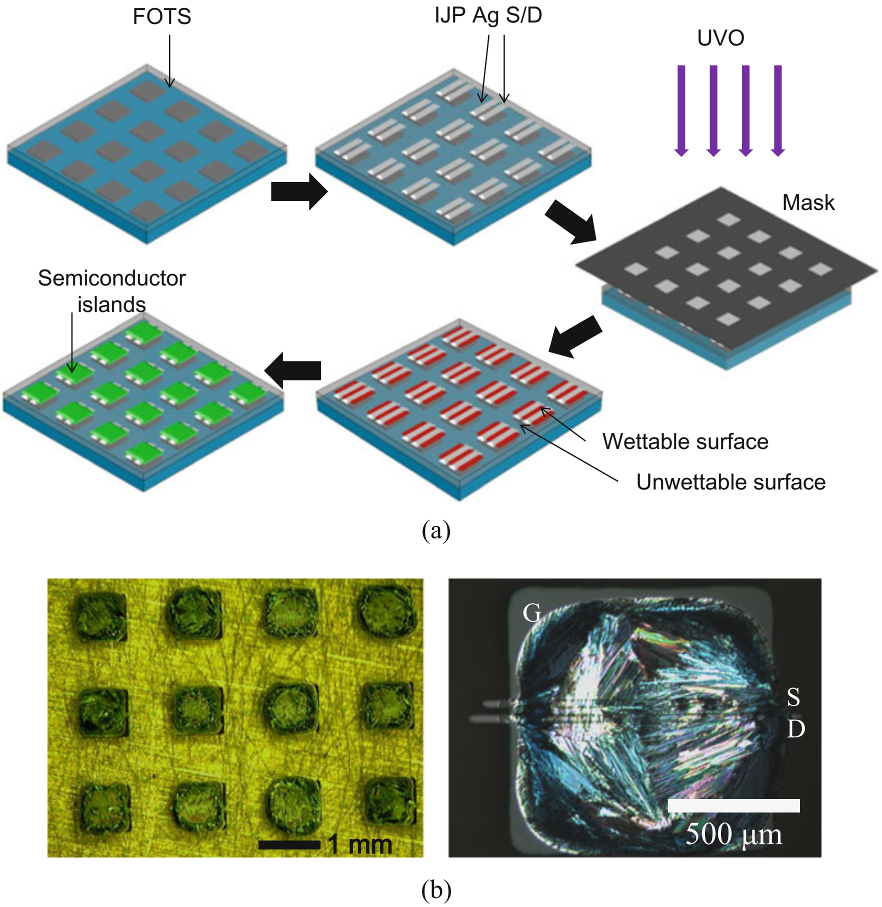
(a)

(b)

(c)

Fig. 5 (**a**) Cross-sectional structure of the fabricated low voltage OTFTs. (**b**) The surface profile of the inkjet printed source/drain Ag electrodes and the channel. (The channel length is 35 μm.) (**c**) The photograph of the fabricated device sample

and drain electrodes. Cross-linked polyvinyl-alcohol (PVA) was used for both the substrate buffer layer (∼400 nm thick) and the gate dielectric layer (∼250 nm thick) to provide the required surface for printing fine Ag electrodes. As shown in Fig. 5b, uniform fine Ag electrodes can be obtained to meet the strict requirements as the electrodes in fully printed OTFTs, by choosing suitable polymer dielectric as the substrate surface material, and optimizing the printing conditions [14, 15]. Before depositing the organic semiconductor layer, the inkjet printed Ag source and drain electrodes were modified with perfluorobenzenethiol (PFBT) self-assembled monolayers (SAMs) to tune the Ag electrode work function to enhance charge injection at the metal/semiconductor interface, and in the mean time help to form better crystallization of the semiconductor film near the contacts. The semiconductor layer was drop-casted from a solution made by mixing 6,13-bis(triisopropylsilylethynyl)-pentacene (TIPS-pentacene) and polystyrene (PS) at 10 g L^{-1} concentration of solids in chlorobenzene (3:1 ratio by volume). Figure 5c shows the photograph of the fabricated device sample. All the above processes were completed in the ambient, and would be easy to be scaled up for large area manufacturing. For circuit



(a)

(b)

Fig. 6 (**a**) Schematic illustration of the fabrication processes for fine inkjet printed (IJP) silver (Ag) electrodes and the patterned channel layer. The entire PVA surface was coated with fluoroalkylsilanes-trichloro(1H,1H,2H,2H-perfluorooctyl)silane (FOTS) self-assembled monolayer (SAMs) before inkjet printing of the Ag electrodes. The FOTS in the channel regions was then selectively removed by ultraviolet ozone (UVO) with a shadow mask, and the drop-casted TIPS-pentacene/PS solution self-assembled into the patterned wettable channel regions. (**b**) Microscope image of the patterned TIPS-pentacene/PS islands in the wettable areas of PVA dielectric surface [15]

integration, patterning the semiconductor layer is generally required to eliminate parasitic leakage current paths outside of the gate controlled area. A method used here is by selectively removing pre-coated SAM layer by ultraviolet ozone treatment to create patterned wettable and less wettable regions to form self-assembled organic semiconductor islands, as illustrated in Fig. 6 [15].

Figure 7 shows the measured transfer (I_D-V_{GS}) and output (I_D-V_{DS}) electrical characteristics of the devices with channel width and length of 2,000 μm

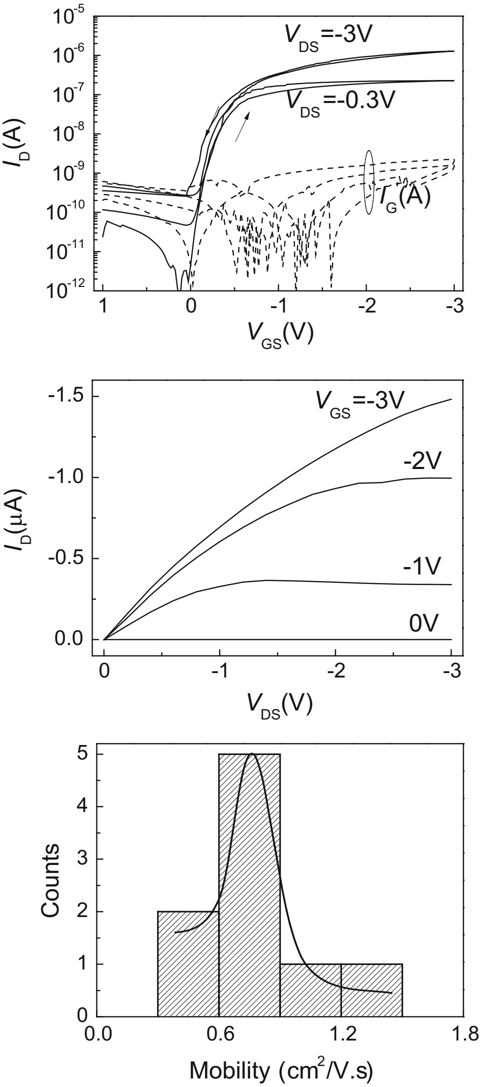
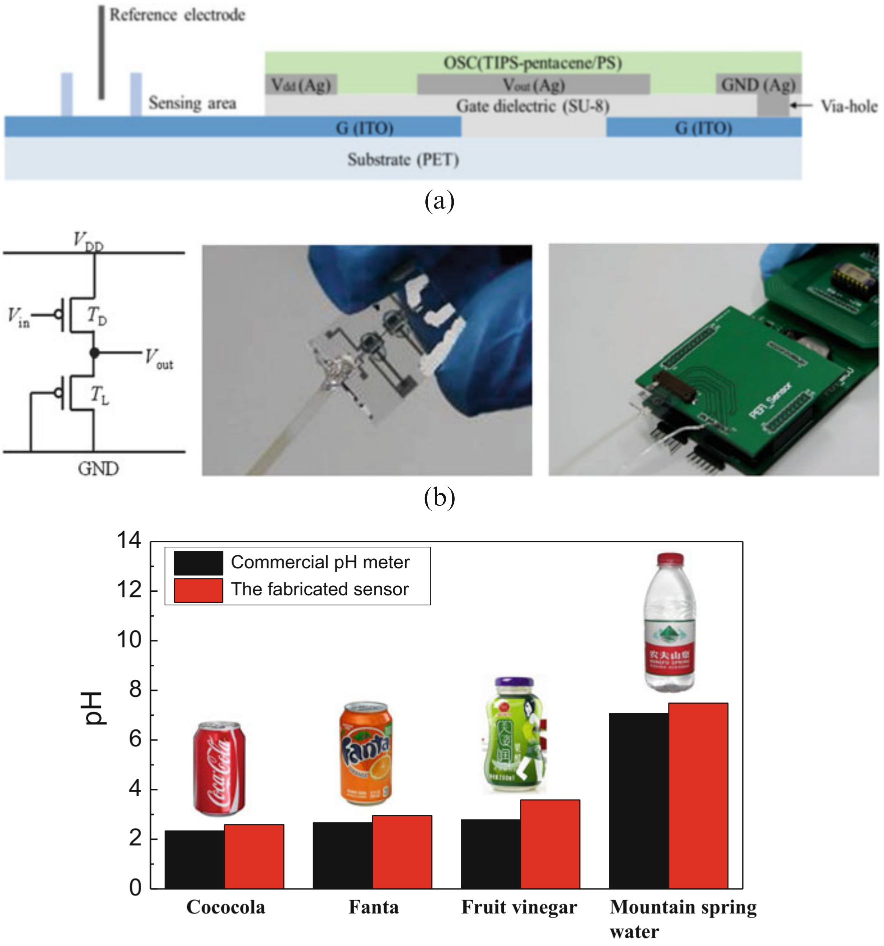


Fig. 7 The measured transfer (I_D-V_{GS}) and output (I_D-V_{DS}) electrical characteristics, and the extracted mobility of the devices with channel width and length of 2,000 μm and 35 μm), respectively

and 35 μm), respectively. The device presents a saturated mobility of about 0.8 cm^2/(V s), subthreshold swing (*SS*) of 100 mV/decade, threshold voltage (V_{th}) of about −0.5 V, ON/OFF ratio of about 10^4 and negligible hysteresis. The uniformity is also reasonably good considering in the university lab conditions. Moreover, the low voltage OTFT using a thick dielectric layer was also shown to be able to sustain high voltage operation [16]. Therefore, this device technology would be applicable to applications where a relatively high operation voltage.

5 Application Example for pH Sensing

A pH sensor tag based on the printed low voltage OTFT was fabricated, with the cross-sectional structure and the circuit schematic given in Fig. 8a, and the photo of the fabricated sample given in Fig. 8b. The via-holes in the inverters for making interconnection between electrodes at different layers was formed by drop-casting deionized water onto the dielectric layer with a micro-syringe before



(a)

(b)

(c)

Fig. 8 (**a**) The cross-sectional structure of the fabricated OTFT based pH sensor tag. (**b**) The circuit schematic, the photo of the fabricated pH sensor sample and the read-out and signal processing circuit board. (**c**) The preliminary measurement results for four different drinks compared with the results measured by a commercial pH meter the cross-linking process. Before testing conductive silver paste was drop-casted on all the electrodes' end pads for robust electrical interconnect [17]. Indium tin oxide (ITO) is used as the gate and can react with. During a chemical reaction in solution, an electrochemical potential is established by the reactants and this potential is influenced by the concentration of corresponding ions. The gate potential change established by chemical reaction of ITO with hydrogen ions (H^+) in the tested solution is converted to an output voltage signal (V_{out}), detected by the read-out circuit board as shown in Fig. 8b. The detected voltage signal was sent to a smart phone through near field communication (NFC). The pH values can then be calculated through a calibration algorithm run in the smart phone. Figure 8c shows the preliminary measurement results for four different drinks compared with the results measured by a commercial pH meter, indicating that the OTFT based pH sensor tag would be useful for quick quality checking for drinks or food. This simple application example also demonstrates that the fully printable low voltage OTFT could be a promising technology to implement transducer circuits for low cost and disposable sensor tags. For different sensing (bio, chemical, temperature, pressure, ...), an OTFT transducer circuit can be designed to convert the sensed signal to an output voltage to be detected by the same read-out hardware.

6 Summary

A hybrid integration concept composed a printed transistor transducer circuit and a common read-out and signal processing hardware or chip is proposed for future sensor development. For different types of sensing, a transducer circuit can be designed based on printed transistor technology to convert the sensed signal to a standard voltage output. Different sensing receptor materials and the antenna can also be easily integrated with the transducer circuit with printing processes. A fully printable low voltage OTFT technology is introduced for this integration architecture, which addresses the issues of relying on ultra-thin or high-k gate dielectric layer for low voltage by reducing the interface trap density of states. A simple application example for pH sensor tag is demonstrated, and this technology would also be applicable to other types of sensors. For the future development, the performance of the printed OTFTs can be continuously improved by adopting new materials. Since the OTFT has capabilities of driving, control and signal transmission for various functional printed electronics systems, the developed technology would also provide a promising platform for developing general low cost and disposable multi-function integrated systems, as described in Fig. 9.

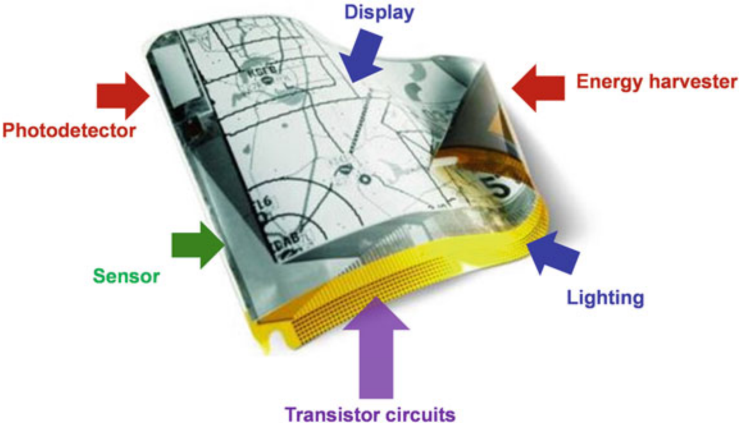


Fig. 9 Illustration of the multi-function integrated system based on printed electronics technologies

References

1. Perera C, Zaslavsky A, Christen P, Georgakopoulos D. Sensing as a service model for smart cities supported by internet of things. Trans Emerg Telecommun Technol. 2013;25:81–93.
2. Bogue R. Towards the trillion sensors market. Sens Rev. 2014;34:137–42.
3. Li J, et al. A stable solution-processed polymer semiconductor with record high-mobility for printed transistors. Sci Rep. 2012;2:754.
4. Xu X, Feng L, He S, Jin Y, Guo X. Solution-processed zinc oxide thin-film transistors with a low-temperature polymer passivation layer. IEEE Electron Device Lett. 2012;33(10):1420–2.
5. Shulaker MM, et al. Carbon nanotube computer. Nature. 2013;501:526–30.
6. Vladu MI. "Green" electronics: biodegradable and biocompatible materials and devise for sustainable future. Chem Soc Rev. 2014;43:588–610.
7. Arias AC, MacKenzie JD, McCulloch I, Rivnay J, Salleo A. Materials and applications for large area electronics: solution-based approaches. Chem Rev. 2010;110:3–24.
8. Wöbkenberg PH, et al. Low-voltage organic transistors based on solution processed semiconductors and self-assembled monolayer gate dielectrics. Appl Phys Lett. 2008;93:013303.
9. Li J, Sun Z, Yan F. Solution processable low-voltage organic thin film transistors with high-k relaxor ferroelectric polymer as gate insulator. Adv Mater. 2012;24:88–93.
10. Hong SH, et al. Electrolyte-gated transistors for organic and printed electronics. Adv Mater. 2013;13:1822–46.
11. Veres J, Ogier SD, Leeming SW, Cupertino DC, Khaffaf SM. Low-k insulators as the choice of dielectrics in organic field-effect transistors. Adv Funct Mater. 2003;13(3):199–204.
12. Feng L, Tang W, Xu X, Cui Q, Guo X. Ultralow-voltage solution-processed organic transistors with small gate dielectric capacitance. IEEE Electron Device Lett. 2013;34(1):129–31.
13. Feng L, Xu X, Guo X. Structure-dependent contact barrier effects in bottom-contact organic thin-film transistors. IEEE Trans Electron Devices. 2012;59(12):3382–8.
14. Tang W, Feng L, Zhao J, Cui Q, Chen S, Guo X. Inkjet printed fine silver electrodes for all-solution-processed low-voltage organic thin film transistors. J Mater Chem C. 2014;2(11):1995–2000.

15. Tang W, Feng L, Jiang C, Yao G, Zhao J, Cui Q, Guo X. Controlling the surface wettability of the polymer dielectric for improved resolution of inkjet-printed electrodes and patterned channel regions in low-voltage solution-processed organic thin film transistors. J Mater Chem C. 2014;2(28):5553–8.
16. Feng L, Zhao J, Tang W, Xu X, Guo X. Solution processed organic thin-film transistors with hybrid low/high voltage operation. IEEE J Disp Technol. 2014;10(11):971–4.
17. Feng L, Tang W, Zhao J, Cui Q, Jiang C, Guo X. All-solution-processed low-voltage organic thin-film transistor inverter on plastic substrate. IEEE Trans Electron Devices. 2014;61(4):1175–80.

Part II
Imaging, Photography, and Video Analytics

The Three-Dimensional Evolution of Hyperspectral Imaging

Min H. Kim

Abstract Hyperspectral imaging has become more accessible nowadays as an image-based acquisition tool for physically-meaningful measurements. This technology is now evolving from classical 2D imaging to 3D imaging, allowing us to measure physically-meaningful reflectance on 3D solid objects. This chapter provides a brief overview on the foundations of hyperspectral imaging and introduces advanced applications of hyperspectral 3D imaging. This chapter first surveys the fundamentals of optics and calibration processes of hyperspectral imaging and then studies two typical designs of hyperspectral imaging. In addition to this introduction, this chapter briefly looks over the state-of-the-art applications of hyperspectral 3D imaging to measure hyperspectral intrinsic properties of surfaces on 3D solid objects.

1 Introduction

Hyperspectral imaging captures the visible and invisible spectral power distributions of a surface as a function of wavelength and provides wavelength-dependent reflectances or radiometric spectral power distributions for each pixel in the form of an image. Hyperspectral imaging has been practiced for physically meaningful measurements of the surface properties as an image, so-called imaging spectroscopy. It has been more accessible nowadays thanks to various imaging technologies such as a liquid crystal tunable filter and a pushbroom camera. In contrast to commodity trichromatic cameras, such hyperspectral imagers yield multiple channels of hyperspectral radiance or reflectance as a 2D image stack of wavelength-dependent spectral power distributions. Recently, these techniques have been practiced more often in computer graphics, computer vision, military, remote sensing, cultural heritage, etc.

M.H. Kim (✉)
KAIST, Computer Science Department, 291 Daehak-ro, Yuseong-gu, Daejeon, Korea
e-mail: minhkim@kaist.ac.kr

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_4

2 Hyperspectral Imaging

For a few recent decades, many hyperspectral imagers and applications have been developed. These systems can be classified as either bandpass- or dispersion-based imaging, depending on the optical design of the systems. This section surveys the fundamental designs of these systems.

2.1 Bandpass Filter-Based Imaging Spectroscopy

2.1.1 Bandpass Filter-Based Systems

Bandpass-based imaging systems contain a set of narrow-bandpass filters on a wheel or a liquid crystal tunable filter [1, 22, 31, 33, 40] as shown in Fig. 1a. This unit discriminates the incoming spectrum in a certain resolution determined by the filter specifications. In general, a monochromatic solid-state detector then captures the wavelength-dependent spectral energy. The spatial resolution of such systems is determined by the specification of the monochromatic image sensor. The spectral resolution varies according to the filter specification. The bandwidth of typical pigment-based bandpass filters is wider than ∼15 nm with a Gaussian transmission distribution in general. The bandwidth of each filter is non-uniform across the transmittance range.

Rapantzikos and Balas [32] developed a bandpass-based hyperspectral imaging system that produces 34 channels in a spectral range of 360—1,150 nm.

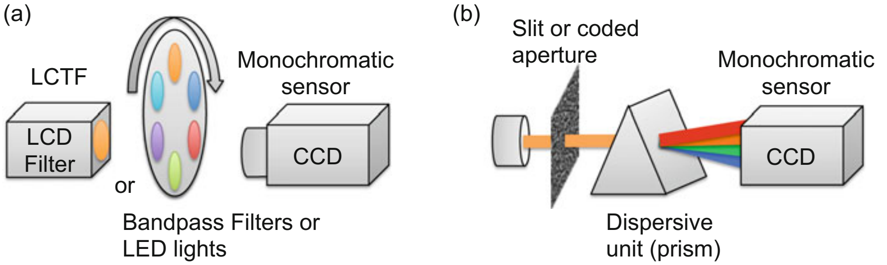


Fig. 1 (**a**) shows the bandpass-based spectroscopy with a set of narrow-bandpass filters or a liquid-crystal tunable filter. Such a bandpass filter unit and a solid-state image sensor are coupled to capture wavelength-dependent spectral energy as a two-dimensional image. The approach is commonly used for investigating static objects. (**b**) presents a simple example based on dispersion. Such a system includes a dispersive unit, for instance, a prism or a diffraction grating that discriminates the wavelength-dependent spectral energy. The detector records the dispersion effect. This design is commonly implemented with a slit in a so-called pushbroom spectral imager. In particular, pushbroom imaging systems are commonly used in remote sensing applications. Image courtesy of © 2012 ACM Transactions on Graphics [18]

Brauers et al. [5] presented a numerical approach to correct geometric distortions while capturing multispectral images. Their system is configured with seven bandpass filters with 40 nm bandwidth.

2.1.2 Liquid Crystal Tunable Filter-Based Systems

A liquid crystal tunable filter (LCTF) is also commonly used for hyperspectral imaging as the filter can be electronically controlled to change its spectral transmittance. This filter is free of physical vibration, compared with the motorized wheel of bandpass filters. LCTFs can provide a spectral resolution on the order of several nanometers with a narrow bandwidth such as ∼7 nm.

Hardeberg et al. [12] measured the spectral reflectance with a hyperspectral imager equipped with an LCTF, yielding a 17-channel hyperspectral image in visible spectrum. Attas et al. [1] presented a two-dimensional multispectral imager with an LCTF. This system captures near-infrared spectra with a bandwidth of 10 nm. Recently, Lee and Kim [22] introduced a hyperspectral imaging system to produce 101 spectral channels from the visible to the near-infrared spectrum. Their system is designed in a two-way imaging structure that includes a polarization-based broadband beam splitter and two LCTFs of approximately ∼7 nm bandwidth. The two-way structure allows us to capture the visible and near-infrared spectra simultaneously to reduce the total acquisition time in half. See Fig. 2 for a photograph of the system and its optical design.

2.2 Dispersion-Based Imaging Spectroscopy

Dispersion-based imaging systems include a dispersive unit, such as a prism or a diffraction grating. This dispersive unit discriminates spectral wavelength and a monochromatic solid-state array captures the wavelength-dependent spectral energy. Dispersion-based systems can be categorized to two different types: pushbroom-based and snapshot-based ones.

2.2.1 Pushbroom-Based Systems

Pushbroom-based systems have recently become popular and are used in many applications of remote sensing such as air- and space-borne scanning systems. These systems measure wavelength-dependent spectral power distributions by scanning a solid-state sensor integrated with a dispersive unit. The spectral resolution of the dispersion-based imager is determined by the refractive index of the dispersive unit and the spatial resolution of the coupled solid-state detector, i.e., the spectral resolution is determined by the number of pixels within the spectral dispersion. Owing to this structure, such systems can provide a high spectral resolution along

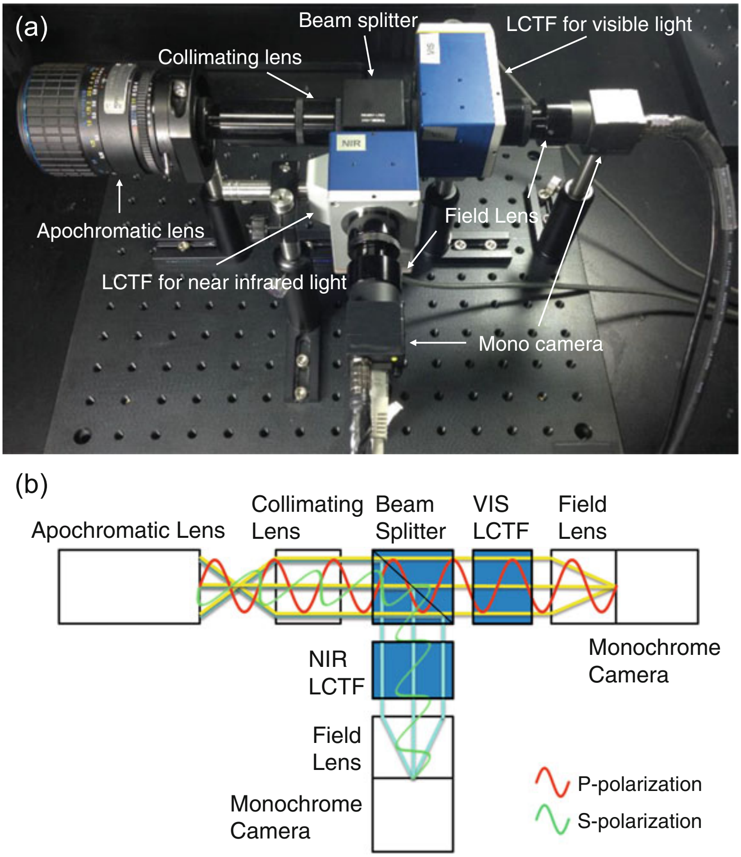


Fig. 2 A photograph of a bandpass-based hyperspectral imager and its optical design, introduced by Lee and Kim [22]. (**a**) presents a snapshot of the hyperspectral imager. The schematic diagram (**b**) shows the optical path inside the imaging system, which includes an apochromatic lens, a collimating lens, a polarization-based beam splitter with a wide range of spectral transmittance, two LCTFs, field lenses, and cameras. The *blue boxes* indicate the polarization units. Image courtesy of © 2014 Springer Lecture Notes in Computer Science [22]

the particular dispersion direction. In addition to the high spectral resolution, this pushbroom design can be integrated easily into existing imaging systems by simply appending a dispersive unit and a slit that discriminates the incident light into a column. See Fig. 1b. However, many pushbroom camera systems have suffered from irregular spatial resolution as either the horizontal or vertical axis is scanned by the mechanical movement of the sensor unit. Therefore, Mouroulis et al. [26] introduced a frequency-based optimization method that allows us to reconstruct spatially uniform spectral information from pushbroom systems. Recently, Hoye and Fridman [15] presented a pushbroom camera system by physically attaching a set of light mixing chambers to the slit.

2.2.2 Snapshot-Based Systems

Although pushbroom-based imaging systems can provide a high resolution in terms both of spectrum and space, the target objects are limited to steady objects as the scanning unit requires to travel over the steady scene. Therefore, snapshot-based systems have been investigated as alternatives. Similar to the pushbroom-based systems, snapshot-based systems also include a dispersive unit like a diffraction grating or a prism, coupled with a coded-aperture mask. The key idea is that the spectral dispersion of the incident image is captured by a two-dimensional monochromatic sensor. However, the spectral dispersions of neighboring rays are projected and recorded with one dimensional overlaps. The individual spectral dispersion of each ray can be iteratively solved from the overlaps by accounting for known spatial constraints of the coded aperture.

Wagadarikar et al. [38, 39] introduced a single disperser architecture that can capture low-resolution multispectral images up to 30 frames per second; the Du et al. [9] system utilizes a prism instead of a diffracting grating. Although this system allows us to capture multispectral videos with a relatively narrow bandwidth (to 2 nm), the spatial resolution and the frame-per-second value of the snapshot-based system are sacrificed in proportion. Kawakami et al. [17] append an additional high-resolution trichromatic camera on the snapshot-based imager to estimate high-resolution multispectral information, assuming low-frequency nature of reflectance of general objects. Kittle et al. [21] introduce an approach to enhance spatial resolution by adding more multi-frame translations of the coded aperture as input. Habel et al. [11] proposed an advanced imager in terms of spectral resolution. The imager is formed with relatively cheap apparatuses while providing up to 4.89 nm spectral resolution (54 spectral bands); its spatial resolution is limited to 120 × 120 pixels. Recently, Kim et al. [18] introduced a hyperspectral imager as a part of hyperspectral 3D imaging system. The system can cover a wide range of spectra from near ultraviolet to infrared wavelengths. This system captures hyperspectral reflectance from 369 to 1,003 nm at 12 nm spectral resolution.

2.2.3 Resolution of Dispersion-Based Systems

The spectral resolution of the bandpass-based imagers is coarser than that of snapshot-based imagers. However, the snapshot-based imagers require a long processing time and suffer from computational artifacts. Although the snapshot-based systems can provide a higher spectral resolution than the bandpass-based imagers, the spatial resolution is insufficient to capture high-frequency surface patterns. Pushbroom-based systems can provide higher-resolution hyperspectral images than the snapshot-based ones. However, as mentioned, the vertical or horizontal resolution of the pushbroom cameras is lower than that of commodity cameras [30]. The acquisition time is also significantly longer than with the snapshot-based imagers and is not practical for general imaging applications. For instance, a pushbroom camera takes an hour to capture a shot in the visible spectral range, albeit at a lower spatial resolution along the mechanically moving axis. The spectral imager presented by Kim et al. [18] takes the advantages of both bandpass- and snapshot-based imaging designs to achieve a higher spatial resolution without sacrificing spectral resolution.

2.3 Calibration of a Hyperspectral Imager

A naïve hyperspectral imaging device returns signals proportional to the incident wavelength-dependent energy. In order to capture physically-meaningful measurements of the radiometric spectral power distributions, the radiometric and geometric properties of the system need to be calibrated.

2.3.1 Radiometric Calibration

The sensor response of a bandpass-based hyperspectral imaging system can be described as a linear product of the quantum efficiency at each wavelength Q_λ of a monochromatic sensor, the transmittance efficiency T_λ through the optical path, and the transmittance functions of the bandpass filters F_λ [18, 22]. Let the camera response f_λ of each bandpass filter be a linear function:

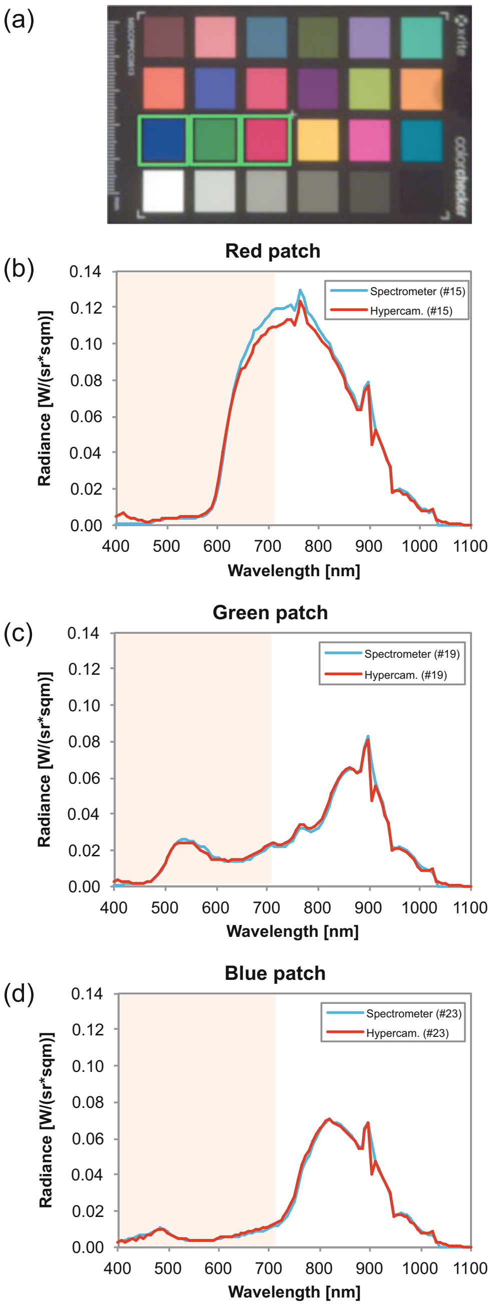
$$f_\lambda = Q_\lambda T_\lambda F_\lambda L_\lambda,$$

where L_λ is the radiance that enters the hyperspectral imaging system. In order to convert the raw signal levels to the incident radiance, a linear transformation C_λ can be determined as an inverse camera response function in order to describe $(Q_\lambda T_\lambda F_\lambda)^{-1}$. A radiometric calibration process can be conducted as follows. Suppose we measured a set of radiance values of training colors of an X-rite ColorChecker and a Spectralon (calibrated to 99 %) under two halogen lights. We then can determine a linear mapping function C_λ of the raw signals that correspond to the incident radiance. The multiplication of f_λ and C_λ yields the physically-meaningful radiance L_λ. Figure 3 demonstrates examples of radiometric measurements on a color target using an LCTF-based hyperspectral imager [22].

2.3.2 Geometric Calibration

An incident ray in an hyperspectral imaging system refracts slightly differently depending on its wavelength. This physical property of a transmissive material at each wavelength is quantified as the refractive index. One of the typical problems in a hyperspectral imaging system is that this refraction effect results in forming images of different wavelengths in different sizes on the image plane. In order to

Fig. 3 (**a**) is an example of a hyperspectral image, captured by Lee and Kim [22]. This hyperspectral image consists of 101 spectral channels from the visible to the near-infrared spectrum. (**b**)–(**d**) present the spectral power distribution of the captured radiance of the red, green and blue patches. The radiometric accuracy of the hyperspectral imager is compared with reference measurements by a hyperspectral spectroradiometer (OceanOptics USB 2000). Image courtesy of © 2014 Springer Lecture Notes in Computer Science [22]



calibrate this geometric mismatch through the optical path, geometric homographies need to be determined by solving an affine transformation for warping the formed images of each wavelength.

Lee and Kim [22] proposed a geometric calibration method for a hyperspectral imager, where a standard checkerboard for camera calibration is captured, and the image coordinates of the corners are corrected through image warping each wavelength. The measurements of the corner points allow us to determine an affine transform per each spectral channel A_λ to calibrate geometric distortion per wavelength. Once the homographies of each wavelength are solved, each affine transform can be applied for warping each spectral channel L_λ to the reference image (e.g., at 554 nm in their system [22]), yielding the hyperspectral radiance L'_λ along the wavelength axis as follows:

$$L'_\lambda = A_\lambda L_\lambda.$$

This geometric calibration converts the hyperspectral images to images consistent in the size across different wavelengths.

2.3.3 Color Transformation

Once a hyperspectral radiance image is captured, the radiometric information can be recorded as a set of two-dimensional image layers. Kim et al. [20] proposed to use a multi-layer image file format, the OpenEXR format [24], to store a hyperspectral image.

In contrast to the large number of hyperspectral channels, ordinary computer displays still have trichromatic RGB color channels. Therefore, in order to display such multi-channel color information, we have to project the multi-channels to three primary colors with respect to visible spectrum. In order to render plausible colors on an RGB display, Kim et al. [20] proposed the following steps. First, the spectral layers L'_λ are projected to the tristimulus values using the CIE color matching functions M_{XYZ} of 2-degree observation [8]. The tristimulus values in CIEXYZ are transformed to the sRGB color values C_{RGB} using the standard sRGB transform M_{sRGB} [29] and then applied for either the gray-world white balancing algorithm [6] or the manual white balancing by manually determining the reference white in the scene to simulate the chromatic adaptation in the human visual system:

$$C_{RGB} = M_{sRGB} M_{XYZ} L'_\lambda.$$

Finally, these calibrated color images C_{RGB} are displayed via gamma correction ($\gamma = 2.2$). Figure 3a shows the converted color examples of a standard ColorChecker target from 101 hyperspectral channels, accompanied with radiometric measurements.

3 Hyperspectral Three-Dimensional Imaging

Reconstructing a 3D object model from multiple overlapping geometry scans has been an active area in the past decade in computer graphics [4]. In general, there are two different types of 3D scanning systems: typically a triangulation-based system for small objects and a time-of-flight system for large scale objects such as a building. Such 3D scanning systems have been coupled to a color imager and lights to capture surface color information as texture. Camera calibration and registration between the camera system and the three-dimensional scanner are necessary to investigate the interrelationship between the spectral and geometric information. Bernardini and Rushmeier [4] surveyed and summarized the general 3D scanning pipeline that has been employed by many research projects and commercial systems. Recently, Holroyd et al. [14] presented a two-way 3D imaging system that allows us to extract both three-dimensional shapes and reflectance functions from the same set of image data. The system achieves a great accuracy in registration of the shape and reflectance and also explores the directional changes of material appearance. However, the spectral resolution of the system is limited to the trichromatic RGB channels.

Recently a straightforward approach to hyperspectral 3D imaging has been introduced by swapping out the standard RGB camera used in current 3D scanning systems and replacing it with a two-dimensional hyperspectral imager, as shown in Sect. 2. As one of the seminal approaches, Mansouri et al. [25] attempted to integrate a two-dimensional multispectral imager to a three-dimensional range scanning system. Similar to Brauers et al. [5], a set of seven bandpass filters is employed and accompanied with an LCD projector. This projector illuminates the surface of the target object to measure the topology of the 3D surface. This system captures a hyperspectral image and maps it to a scanned surface as a texture map. This seminal system is limited to capturing a flat surface only.

3.1 3D Imaging Spectroscopy

Kim et al. [18] introduced a 3D imaging spectroscopy (3DIS) system by integrating 2D imaging spectroscopy and 3D scanning, which is the first complete hyperspectral 3D imaging system to yield complete 3D scanning models. This enables measuring physically-meaningful 3D hyperspectral patterns of arbitrarily-shaped solid objects with high accuracy. In particular, they proposed a modification on a dispersion-based hyperspectral imaging design [39] to achieve high enough spatial and spectral resolution to build a 3D hyperspectral pattern from the captured 2D hyperspectral images. Figure 4 shows an overview of the hyperspectral 3D imaging system.

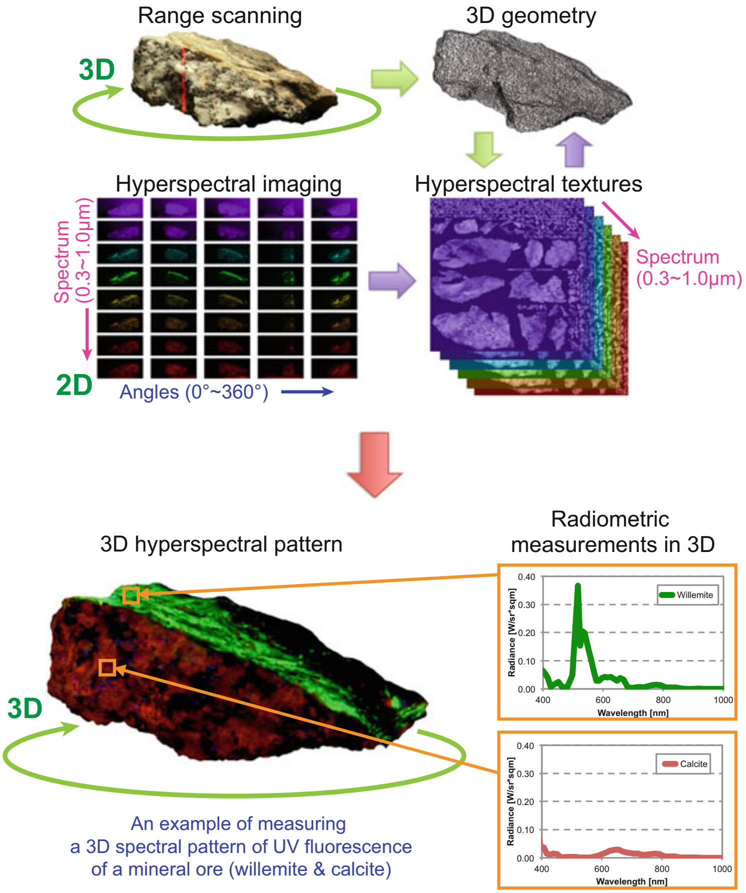


Fig. 4 An overview of the hyperspectral 3D imaging system built by Kim et al. [18] for measuring 3D hyperspectral patterns on 3D solid objects. This 3D imaging spectroscopy system, so-called 3DIS, measures 3D geometry and hyperspectral radiance simultaneously. Piecewise geometries and radiances are reconstructed into a 3D hyperspectral pattern. This 3DIS system is used to acquire physically-meaningful 3D hyperspectral patterns of various wavelengths for scientific research. Image courtesy of © 2012 ACM Transactions on Graphics [18]

3.1.1 The Dispersion-Based Hyperspectral Imager

The design of the hyperspectral imager [18] originated from the ground of the snapshot-based design (described in Sect. 2.2.2). The authors couple dispersive prism optics and a coded aperture mask to resolve spatio-spectral information for radiometric sampling. They then solve an under-determined system by solving sparsity-constrained optimization problems.

Different from bandpass filter-based systems, their imaging system measures continuous hyperspectral patterns from NUV-A (359 nm) to NIR (1 μm). In order

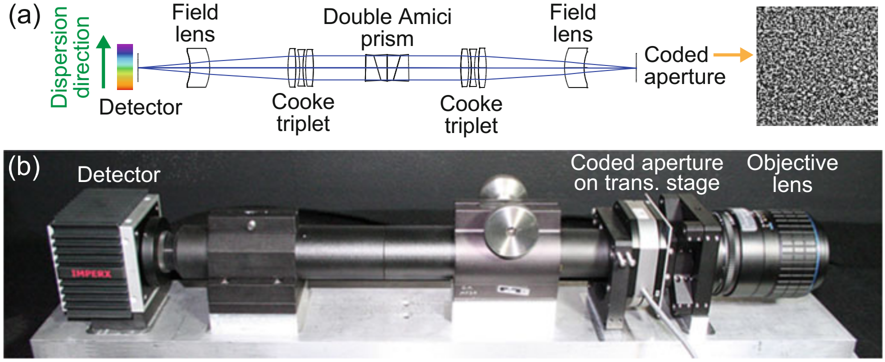


Fig. 5 (**a**) represents the optical paths from the coded aperture (*right*) to the detector (*left*) in the 3DIS system [18]. A snapshot of the coded aperture with a monochromatic light (bandwidth: 10 nm from 560 to 570 nm) is inserted on the *right*. (**b**) shows a photograph of the 2D imaging unit in the 3DIS system. Image courtesy of © 2012 ACM Transactions on Graphics [18]

to increase the efficiency of the UV spectrum, this system was built with specialized optics, such as fused silica (FS) and calcium fluoride (CaF_2). These optical components enable this system to exceed the spectral range of traditional imaging systems, where UV transmittance below 400 nm decreases rapidly due to absorption effects. In contrast to the IR band, the UV band is challenging for imaging due to the inherent transmittance characteristics of the optical substrate of glass components.

In the hyperspectral imager, a random-pattern coded aperture is lithographically etched on a quartz substrate. A piezoelectric translation stage modulates the aperture, the aperture code is then directly relayed onto the monochromatic imaging sensor. The system includes relay lenses and a double Amici prism to disperse the incoming rays. The light sources used in this system are a Xenon light bulb and a UV fluorescence light for measuring UV fluorescence. See Fig. 5 for the optical design of the hyperspectral imaging unit.

3.1.2 3D System Integration

The 2D hyperspectral imager is integrated into a 3D imaging pipeline in order to measure not only 3D shape but also reflectance of 3D solid objects. Their 3D scanning pipeline is based on a classic 3D imaging workflow as described in [10]. The hyperspectral imager, a laser range scanner and a Xenon light source are mounted together on a standard optical table, located in a darkroom. The positions of the imager system, light source and turntable axis are calibrated in terms of the laser scanner coordinate system, using standard calibration targets.

Figure 6 presents the design of the hyperspectral 3D imaging system. A laser projector shines a thin sheet of light onto the object. The laser sensor detects, on each scan line, the reflected laser light to produce a depth map. A Xenon light

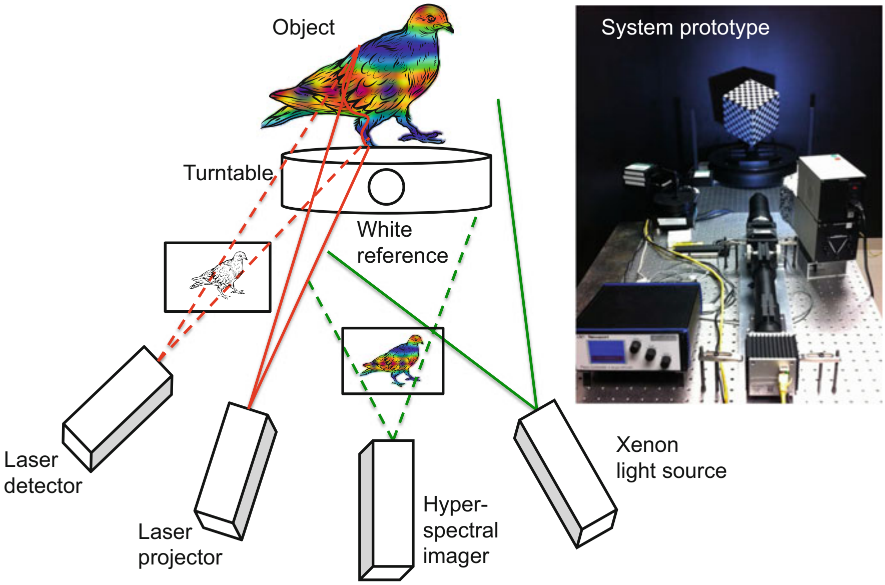


Fig. 6 Principal design of the 3DIS system [18]. Inset: a photograph of the prototype system. Image courtesy of © 2012 ACM Transactions on Graphics [18]

source (or a UV fluorescent light) illuminates the surface with a broad spectrum. The hyperspectral imager measures reflected radiance to compute a reflectance map with respect to the reference white.

3.1.3 Measuring 3D Hyperspectral Patterns

The 3DIS system is demonstrated with various practical and scientific applications for measuring hyperspectral patterns on 3D solid objects. The imaging system was used for non-destructive measurements of 3D reflectance and fluorescence patterns of biological organisms, minerals and an archaeological artifact in the collaboration with Yale Peabody Museum of Natural History. Figures 7 and 8 demonstrate examples of scanning stuffed-bird specimens.

3.2 Hyperspectral Photometric Stereo

3.2.1 Photometric Stereo

Photometric stereo is a 3D imaging technique that has been commonly performed for capturing the shape of 3D solid objects in computer vision for more than three

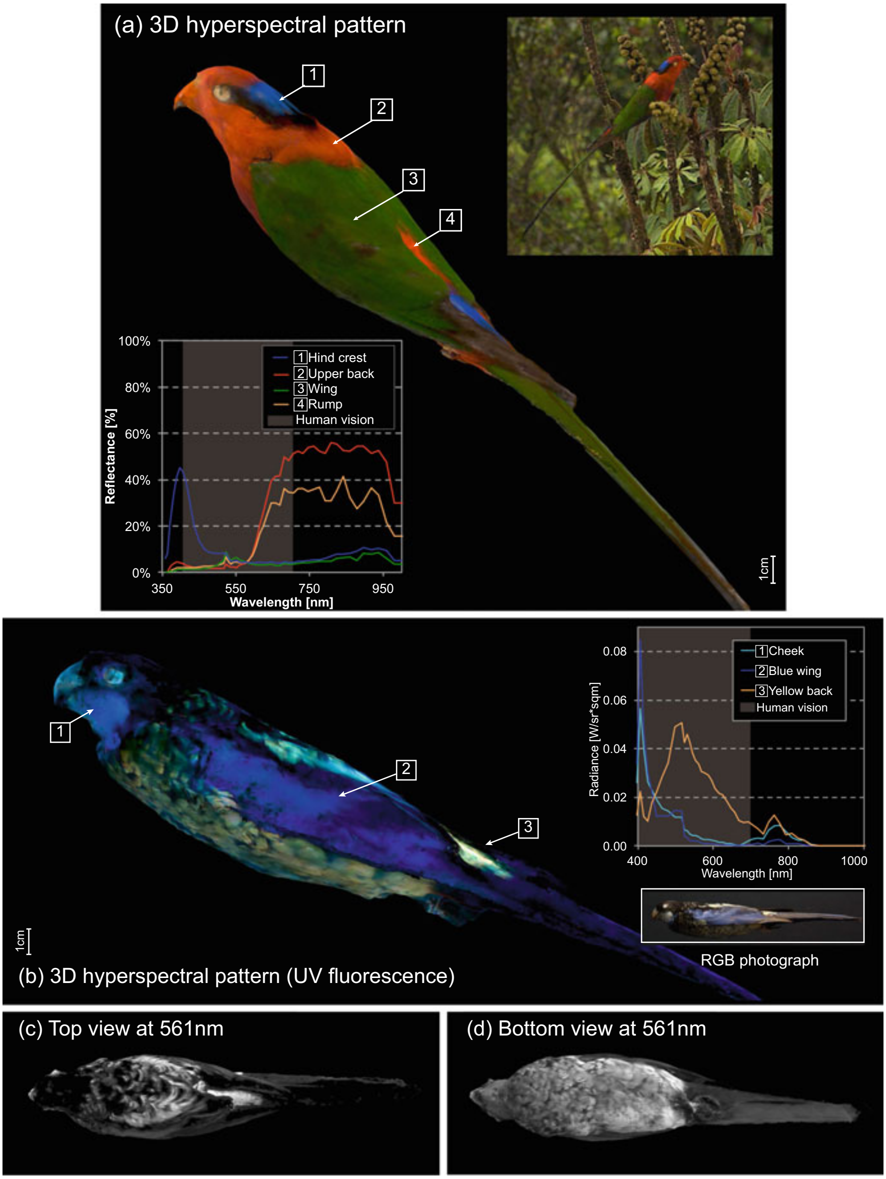


Fig. 7 Examples of 3D hyperspectral patterns captured by the 3DIS system [18], followed by 3D spectrum analysis. (**a**) and (**b**) are 3D models of stuffed birds, detailed with the surface reflectance and fluorescence readings. (**a**) is scanned under a Xenon light to capture the hyperspectral material appearance of the specimen. Inset (*top*): simulation rendering of the bird in its natural environment. (**b**) is scanned under a UV light to reveal the UV-induced visible-fluorescence of the feather. (**c**) and (**d**) are the photon catches of specific wavelengths. Inset (*bottom*): diffuse reflectance readings (359 nm–1 μm) on the 3D hyperspectral patterns as compared to human vision. Image courtesy of © 2012 ACM Transactions on Graphics [18]

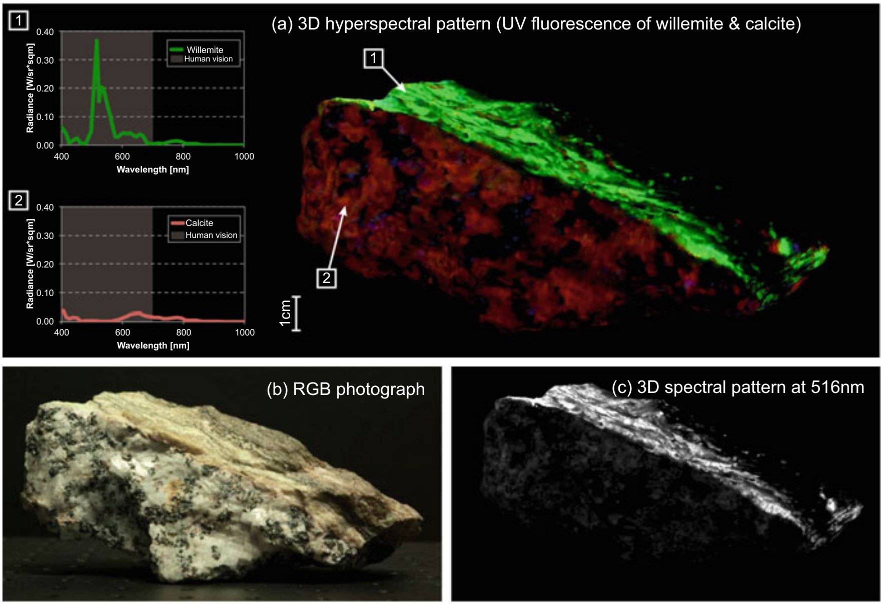


Fig. 8 An example of scanning an ore that includes willemite, calcite and magnetite, captured by the 3DIS system [18]. (**a**) is a physically-based rendering result of the 3D hyperspectral pattern of UV fluorescence with associated spectral readings. Inset (*top*): NUV-induced fluorescent radiance of willemite. Inset (*bottom*): fluorescent radiance of calcite. An ultra-violet spectrum (260–390 nm) illuminates the object, and the emitted fluorescent radiance (excluding reflected UV) is measured and rendered in 3D. (**b**) shows the appearance of the ore under white light captured in a photograph. (**c**) is a rendering of the 3D spectral pattern at 516 nm, where the willemite presents a spectral peak of fluorescence. Image courtesy of © 2012 ACM Transactions on Graphics [18]

decades. Photometric stereo estimates surface normal vectors over the surface of the 3D solid objects, yielding a normal map, the surface topology description of the 3D shape orientation. Photometric stereo captures the shading information over the surface by varying the position of point light sources. This allows us to estimate the normal vectors from shading, measured as pixel intensities by a monochromatic camera [3]. In contrast, the classical binocular stereo estimates the depth information from the parallax disparity, caused by placing the two cameras distant away in a base line. One of the virtues of photometric stereo is to produce a high-resolution normal map from a relatively simple setup, which includes a camera and multiple light sources. However, photometric stereo and hyperspectral imaging have still been rarely combined and practiced together in the computer vision and hyperspectral imaging communities.

3.2.2 Combining Hyperspectral Imaging with Photometric Stereo

In photometric stereo, many optical phenomena occur as obstacles, such as indirect illumination, specular reflection and self shadows, degrading the accuracy of the shape measured by photometric stereo. Much research and development have been conducted with focus on reconstructing surface normals from Lambertian and non-Lambertian reflections by removing self shadows and specular reflections in photometric stereo [2, 3, 7, 13, 34, 36, 41]. However, removing the indirect illumination effect in photometric stereo [23, 28] has been less discussed relatively.

Interreflection is an optical phenomenon that occurs over a concave surface. When one of the sides is illuminated, the reflected light illuminates the neighboring side in the concave shape, where two points over the surface face each other. Most photometric stereo methods are designed with the general assumption that the surface has Lambertian reflection and the geometric proxy of the object is a convex shape, which does not suffer from interreflection. However, objects in the real world are built with a mixture of convex and concave shapes. Removing interreflection in photometric stereo is a traditional chicken-and-egg problem as we need to account for interreflection without knowing geometry. This is a typical problem when we capture a 3D surface geometry with concave shape using photometric stereo. See Fig. 9c for examples.

Nam and Kim [27] proposed a hyperspectral (a.k.a. multispectral, which means acquiring visible spectral information in hyperspectral imaging) photometric stereo method in order to remove interreflection on diffuse materials of the concave surface while capturing a 3D shape with photometric stereo. They estimate the amount of interreflection on a monochromatic diffuse surface using the reflectance information of visible spectrum. The method is integrated into a typical photometric stereo system by simply substituting the RGB camera with a multispectral imager as the method does not rely on additional structured or colored lights. Figure 9 shows the schematic overview of their hyperspectral photometric stereo method.

3.2.3 Removing Interreflection

Removing interreflection is challenging as the effect is integrated in a light path, where a ray of light is emitted from a light source, travels through a medium like air, reflects on the surface of an object, and comes into the camera. The multiple-bounded light is affected by the surface albedo of the reflecting surface. The multiple-bounded light becomes a new light source, so-called radiosity, and the reflection is added at each point. Most of energy that comes into the camera is either directly from the light source or one bounded light from the object surface, which is the inner product of the light and surface reflectance. The portion of second and higher bounded light varies from scene to scene. Some objects or scenes are more susceptible to indirect illumination. In this case, the radiance measured by a camera is affected not only by the illumination and the surface albedo, but also by the reflectance of the surrounding surfaces.

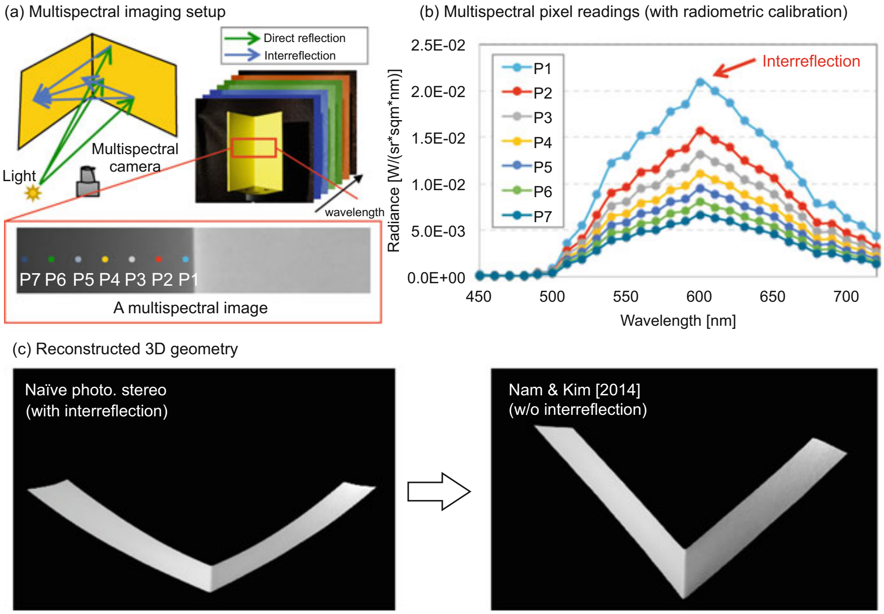


Fig. 9 Schematic diagram of the multispectral photometric stereo method proposed by Nam and Kim [27]. This method allows to remove interreflection over the monochromatic diffuse surfaces. (**a**) presents the imaging setup and an example of the captured multispectral image of an L-shaped object in 90°. As shown in the closeup view, the inner faces of the object present interreflection along the direct reflection. (**b**) shows the spectral power distribution measured by the multispectral imager over the seven points in the closeup view. (**c**) compares the surfaces reconstructed by ordinary photometric stereo (*upper*) and their proposed method (*lower*). The reconstructed geometry using the proposed method is much closer to the physical shape (L-shaped in 90°) of the captured object, compare the geometry obtained by a naïve photometric stereo method. Image courtesy of © 2014 IEEE Computer Graphics and Applications [27]

The key insight of this hyperspectral photometric stereo method is that a hyperspectral camera can capture spectrum-dependent albedo with many channels. They therefore model multiple bounces of wavelength-dependent interreflection as a polynomial function and optimize the interreflection effect through hyperspectral reflectance analysis. This allows us to separate interreflection over diffuse surfaces from measured radiance. Their hyperspectral photometric stereo does not rely on multiplexing spectral lights like [35, 37] so that the proposed method [27] is capable of acquiring any arbitrary shape and illumination without the help of structured light and colored light.

3.2.4 Measuring a Shape with Hyperspectral Imaging

Nam and Kim [27] demonstrated several benefits of the hyperspectral (a.k.a. multispectral) photometric stereo method such as removing interreflection and reconstructing the 3D shapes of objects with high accuracy.

Figure 10 compares the performance variation of their method according to the number of input spectral channels. They scan a concave-shaped soap, which can maximize the interreflection. As soap has specular reflection, polarizing filters are attached in front of light sources in order to prevent the specular reflection coming into the sensor directly. This figure compares three different 3D photometric stereo results to the ground truth obtained by a 3D laser scanner (NextEngine). (b) shows the reconstruction results of the normals and 3D geometry with the naïve photometric stereo approach (without removing interreflection). The reconstructed geometry is flattened compared to the ground truth. (c) and (d) present the normals and 3D models using the proposed method with two different cameras: an RGB camera and a hyperspectral imager. In (c), although the reconstructed geometry is still somewhat flattened, there is an improvement in terms of sharpness on the edges. (d) shows the results of the hyperspectral photometric stereo system using 29 channels. The proposed method yields a virtually identical geometry to the ground truth. Using a sufficient number of channels, they can acquire high-frequency details of the object surface, yielding high-fidelity normals and 3D shapes.

Figure 11 shows the results of scanning a human face by the hyperspectral photometric stereo method. Facial features, such as a nose, a mouth and eyes, make the overall shape of the face a mixture of concave and convex areas. In particular, the area between two eyes, the area between a nose and a cheek and the area between two lips are vulnerable to the interreflection. In addition, the rough skin also produces the interreflection in a micro scale. (a) and (b) compare the results of the naïve photometric stereo and those of the hyperspectral photometric stereo method. The reconstructed geometry in (a) shows a flattened facial shape and a smooth skin. In contrast, in (b), the concavity, as well as the convexity, of the face are well preserved and the skin keeps the roughness of the human skin. (c) and (d) show the close-ups of (a) and (b), respectively. Note that, tiny features in the skin, such as acne, are blurred in (c) and only recovered by the proposed method in (d).

3.3 Rendering Three-Dimensional Hyperspectral Data

The data structure of such 3D hyperspectral data is significantly different from that of ordinary imaging data. The numeric values in the hyperspectral images are calibrated as radiance (unit: W/sr/sqm) so that the values describe the radiometric power per wavelength. These values are not only represented as floating numbers but the spectral image also includes more than three channels such as red, green and blue. Ordinary image files such as an integer-based RGB image format therefore are not suitable for storing hyperspectral image data. In addition to the number

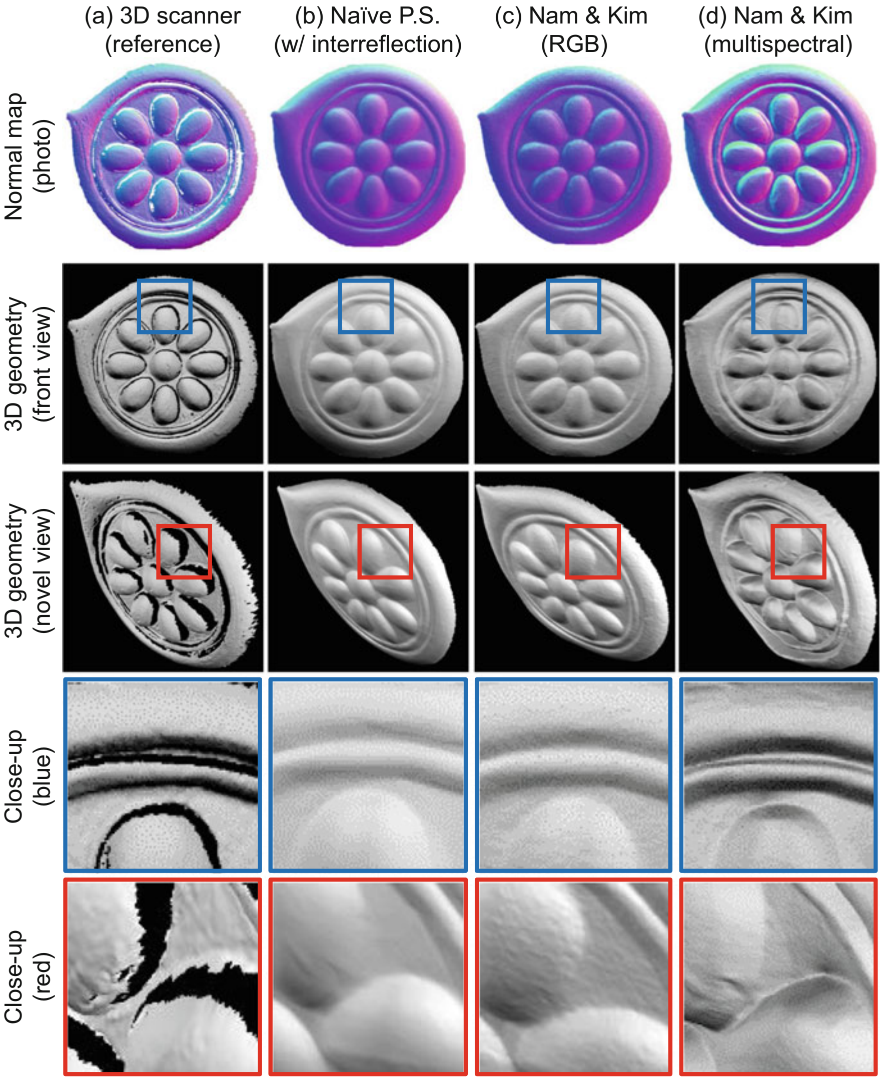


Fig. 10 Nam and Kim [27] compared the reconstruction 3D models depending on the number of input spectral channels. (**a**) shows the ground truth obtained by a 3D laser scanner (NextEngine). (**b**) is the result of naïve photometric stereo without removing interreflection. (**c**) is the result of their hyperspectral method applying it to an RGB camera using three spectral channels. (**d**) is the result of the proposed hyperspectral photometric stereo using 29 visible channels. The reconstructed shape is virtually identical to the ground truth. Image courtesy of © 2014 IEEE Computer Graphics and Applications [27]

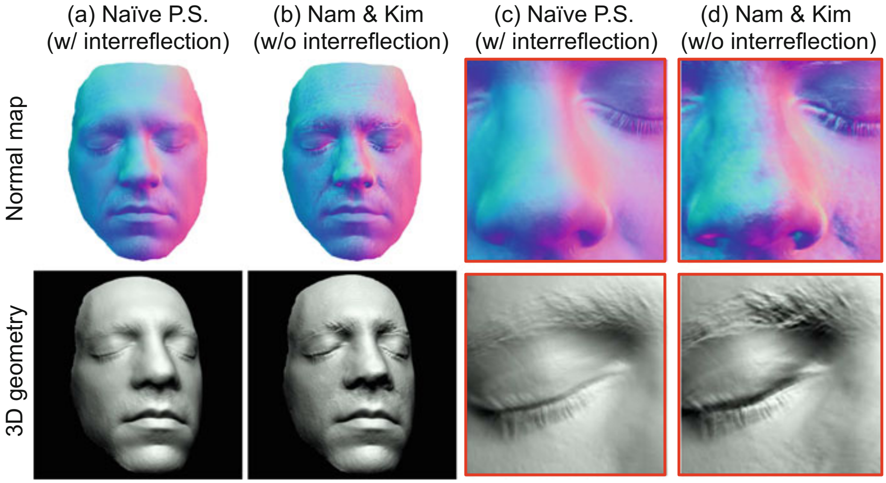
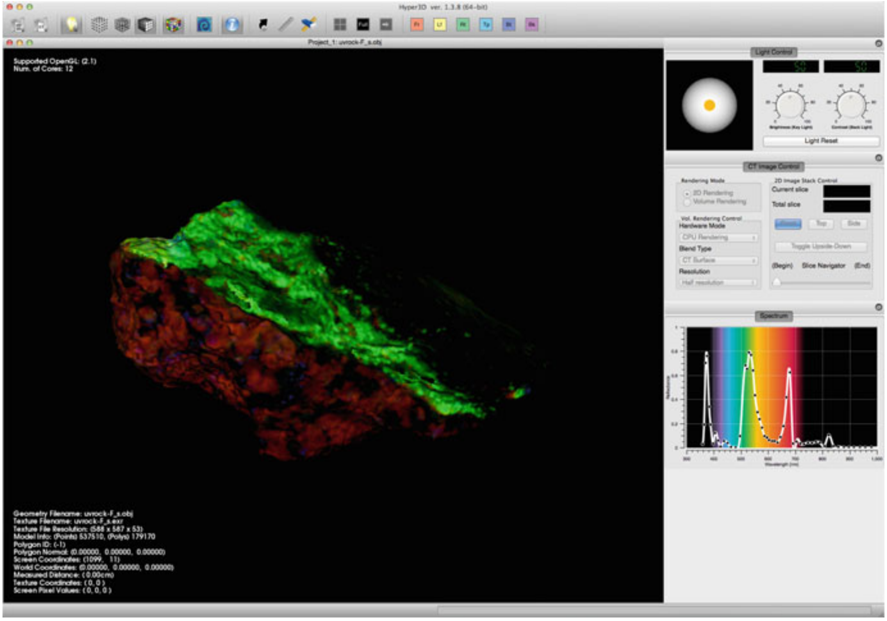


Fig. 11 Results of scanning a human face, acquired by the hyperspectral photometric stereo method [27]. (**a**) shows the normal map and the reconstructed 3D geometry using naïve photometric stereo. (**b**) shows the results from the hyperspectral photometric stereo. (**b**) presents the concavity and the convexity of the facial features better than (**a**). In addition, the roughness of the human skin is well recovered in (**b**). (**c**) and (**d**) show the close-ups of (**a**) and (**b**), respectively. Note that, tiny features in the skin, e.g., acne, are only recovered by the proposed method in (**d**). Image courtesy of © 2014 IEEE Computer Graphics and Applications [27]

difference of multiple channels, an extra process to transform the radiometric colors to trichromatic RGB signals (described in Sect. 2.3.3) is necessary to visualize such hyperspectral data on a trichromatic display. We can utilize a multi-layer image file format, such as the OpenEXR format [24] to store multispectral reflectance.

Kim et al. [19, 20] introduced a hyperspectral image software tool, which enables visualizing two- and three-dimensional hyperspectral image data. In addition to visualizing such hyperspectral data, this open-source software tool handles various types of 2D and 3D image data (also preserving user-generated metadata and annotations [20]), including ordinary 3D scanning models and medical computed tomography imaging data.

Figure 12 shows a screenshot of visualizing a 3D hyperspectral model, scanned by the hyperspectral 3D imaging system [18]. The left-hand-side column presents the main window that renders the 3D hyperspectral measurements. By navigating the mouse pointer over the object's surface, the multispectral spectrum is scientifically visualized at the bottom of the right column. The right-hand-side column shows the light controls for the 3D model/volume rendering visualization (key/fill light), CT stack data navigator and volume rendering options. The widget at the bottom is for radiance or reflectance measurements over the 3D surface.



```
Supported OpenGL: (2.1)
Num. of Cores: 12
```

```
Geometry Filename: uvrock-F_s.obj
Texture Filename: uvrock-F_s.exr
Texture File Resolution: (588 x 587 x 53)
Model Info: (Points) 537510, (Polys) 179170
Polygon ID: (-1)
Polygon Normal: (0.00000, 0.00000, 0.00000)
Screen Coordinates: (1099, 11)
World Coordinates: (0.00000, 0.00000, 0.00000)
Measured Distance: ( 0.00cm)
Texture Coordinates: ( 0, 0 )
Screen Pixel Values: ( 0, 0, 0 )
```

Fig. 12 A screenshot of a software, so-called Hyper3D [16], developed by Kim et al. [19, 20]. This software supports three different types of imaging data: a three-dimensional surface geometry model, a two-dimensional stack visualization of CT imaging, and the color visualization of a hyperspectral image. This is an example that shows the visualization of a hyperspectral 3D object (the spectral plot widget on the *right bottom* illustrates the spectral reading of the *green region*). Image courtesy of © 2014 The International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST) [19]

4 Conclusions

This chapter briefly surveys the foundations of hyperspectral imaging, bandpass filter-based and dispersion-based imaging spectroscopy, followed by the radiometric and geometric calibration processes. This chapter also provides an overview of the state-of-the-art applications of hyperspectral 3D imaging, which is a combination of hyperspectral imaging and 3D imaging, in order to measure hyperspectral intrinsic surface properties on 3D solid objects. One is a 3D imaging system to measure physically-meaningful 3D hyperspectral patterns of arbitrarily-shaped solid objects [18]. The other is a photometric stereo method with a hyperspectral imager to remove indirect illumination from reflection and to reconstruct high-fidelity surface normals and geometry exclusively from direct illumination with a high accuracy [27]. More hyperspectral 3D imaging applications are expected to come in the near future.

Acknowledgements This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Attas M, Cloutis E, Collins C, Goltz D, Majzels C, Mansfield JR, Mantsch HH. Near-infrared spectroscopic imaging in art conservation: investigation of drawing constituents. J Cult Herit. 2003;4(2):127–36.
2. Barsky S, Petrou M. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. IEEE Trans Pattern Anal Mach Intell. 2003;25(10):1239–52.
3. Basri R, Jacobs DW, Kemelmacher I. Photometric stereo with general, unknown lighting. Int J Comput Vis. 2007;72(3):239–57.
4. Bernardini F, Rushmeier H. The 3D model acquisition pipeline. Comput Graph Forum. 2002;21(2):149.
5. Brauers J, Schulte N, Aach T. Modeling and compensation of geometric distortions of multispectral cameras with optical bandpass filter wheels. In: 15th European signal processing conference, vol. 15; 2007. p. 1902–6.
6. Buchsbaum G. A spatial processor model for object colour perception. J Franklin Inst. 1980;310(1):1–26.
7. Chandraker M, Agarwal S, Kriegman D. Shadowcuts: photometric stereo with shadows. In: IEEE conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07). Piscataway: IEEE, p. 1–8.
8. CIE. Colorimetry. CIE Pub. 15.2, Commission Internationale de l'Eclairage (CIE), Vienna; 1986.
9. Du H, Tong X, Cao X, Lin S. A prism-based system for multispectral video acquisition. In: Proceedings of the International Conference on Computer Vision (ICCV), Piscataway: IEEE, 2009. p. 175–82.
10. Farouk M, Rifai IE, Tayar SE, Shishiny HE, Hosny M, Rayes ME, Gomes J, Giordano F, Rushmeier HE, Bernardini F, Magerlein K. Scanning and processing 3D objects for web display. In: Proceedings of the international conference on 3D Digital Imaging and Modeling (3DIM); 2003. p. 310–7.
11. Habel R, Kudenov M, Wimmer M. Practical spectral photography. In: Computer graphics forum, vol. 31. Wiley-Blackwell: Hoboken, 2012. p. 449–58.
12. Hardeberg JY, Schmitt F, Brettel H. Multispectral color image capture using a liquid crystal tunable filter. Opt Eng. 2002;41(10):2532–48.
13. Hernández C, Vogiatzis G, Cipolla R. Shadows in three-source photometric stereo. In: Computer vision-ECCV 2008. Springer; 2008. p. 290–303.
14. Holroyd M, Lawrence J, Zickler T. A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance. ACM Trans Graph (Proc SIGGRAPH 2010). Los Angeles, United States, 2010;29(3):1–12. Article no. 99.
15. Hoye G, Fridman A. Mixel camera—a new push-broom camera concept for high spatial resolution keystone-free hyperspectral imaging. Opt Exp. 2013;21(9):11,057–77.
16. Hyper3D. An open-source project in SourceForge.net. 2012. http://sourceforge.net/projects/hyper3d/.
17. Kawakami R, Wright J, Tai YW, Matsushita Y, Ben-Ezra M, Ikeuchi K. High-resolution hyperspectral imaging via matrix factorization. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2011. p. 2329–36.
18. Kim MH, Harvey TA, Kittle DS, Rushmeier H, Dorsey J, Prum RO, Brady DJ. 3D imaging spectroscopy for measuring hyperspectral patterns on solid objects. ACM Trans Graph (Proc SIGGRAPH 2014). 2012;31(4):38:1–11.

19. Kim MH, Rushmeier H, ffrench J, Passeri I. Developing open-source software for art conservators. In: The international symposium on virtual reality, archaeology and intelligent cultural heritage, eurographics association. Brighton, England; 2012. p. 97–104.
20. Kim MH, Rushmeier H, ffrench J, Passeri I, Tidmarsh D. Hyper3D: 3D graphics software for examining cultural artifacts. ACM J Comput Cult Herit 2014;7(3):1:1–19.
21. Kittle D, Choi K, Wagadarikar A, Brady DJ. Multiframe image estimation for coded aperture snapshot spectral imagers. Appl Opt 2010;49(36):6824–33.
22. Lee H, Kim MH. Building a two-way hyperspectral imaging system with liquid crystal tunable filters. In: Springer LNCS 8509 (Proceedings of ICISP 2014). Normandy: Springer; 2014. p. 26–34.
23. Liao M, Huang X, Yang R. Interreflection removal for photometric stereo by using spectrum-dependent albedo. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR); 2011. p. 689–96.
24. Lucas Digital Ltd. OpenEXR. 2009. http://www.openexr.com/.
25. Mansouri A, Lathuiliere A, Marzani FS, Voisin Y, Gouton P. Toward a 3d multispectral scanner: an application to multimedia. MultiMedia, IEEE. 2007;14(1):40–7.
26. Mouroulis P, Green RO, Chrien TG. Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information. Appl Opt. 2000;39(13):2210–20.
27. Nam G, Kim MH. Multispectral photometric stereo for acquiring high-fidelity surface normals. IEEE Comput Graph Appl. 2014;34(6):57–68.
28. Nayar SK, Ikeuchi K, Kanade T. Shape from interreflections. Int J Comput Vis (IJCV). 1991;6(3):173–95.
29. Nielsen M, Stokes M. The creation of the sRGB ICC profile. In: Proceedings of the color imaging conference IS&T; 1998. p. 253–7.
30. Qin J. Hyperspectral imaging instruments. In: Sun DW, editor. Hyperspectral imaging for food quality analysis and control. Elsevier: Amsterdam, 2010. p. 129–75.
31. Rapantzikos K, Balas C. Hyperspectral imaging: potential in non-destructive analysis of palimpsests. In: Proceedings of the International Conference on Image Processing (ICIP), vol. 2; 2005. p. 618–21.
32. Rapantzikos K, Balas C. Hyperspectral imaging: potential in non-destructive analysis of palimpsests. In: IEEE International Conference on Image Processing, 2005 (ICIP 2005), vol. 2. Piscataway: IEEE, 2005. p. II–618.
33. Sugiura H, Kuno T, Watanabe N, Matoba N, Hayashi J, Miyata Y. Development of a multispectral camera system. Proc SPIE. 2000;3965:331–9.
34. Sun J, Smith M, Smith L, Midha S, Bamber J. Object surface recovery using a multi-light photometric stereo technique for non-lambertian surfaces subject to shadows and specularities. Image Vis Comput. 2007;25(7):1050–7.
35. Takatani T, Matsushita Y, Lin S, Mukaigawa Y, Yagi Y. Enhanced photometric stereo with multispectral images. In: International conference on Machine Vision Applications (MVA). IAPR: Kyoto, 2013.
36. Verbiest F, Van Gool L. Photometric stereo with coherent outlier handling and confidence estimation. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR); 2008. p. 1–8.
37. Vogiatzis G, Hernández C. Self-calibrated, multi-spectral photometric stereo for 3d face capture. Int J Comput Vis (IJCV). 2012;97(1):91–103.
38. Wagadarikar A, John R, Willett R, Brady DJ. Single disperser design for coded aperture snapshot spectral imaging. Appl Opt. 2008;47(10):B44–51.
39. Wagadarikar AA, Pitsianis NP, Sun X, Brady DJ. Video rate spectral imaging using a coded aperture snapshot spectral imager. Opt Exp. 2009;17(8):6368–88.
40. Ware G, Chabries D, Christiansen R, Brady J, Martin C. Multispectral analysis of ancient Maya pigments: implications for the Naj Tunich Corpus. In: Proceedings of the IEEE geoscience and remote sensing symposium, vol. 6; 2000. p. 2489–91.
41. Wu TP, Tang KL, Tang CK, Wong TT. Dense photometric stereo: a markov random field approach. IEEE Trans Pattern Anal Mach Intell. 2006;28(11):1830–46.

Computational Photography Using Programmable Aperture

Hajime Nagahara and Rin-ichiro Taniguchi

Abstract Since 1960s, aperture patterns have been studied extensively and a variety of coded apertures have been proposed for various applications, including extended depth of field, defocus deblurring, depth from defocus, light field acquisition, etc. Researches have shown that optimal aperture patterns can be quite different due to different applications, imaging conditions, or scene contents. In addition, many coded aperture techniques require aperture patterns to be temporally changed during capturing. As a result, it is often necessary to have a *programmable aperture camera* whose aperture pattern can be dynamically changed as needed in order to capture more useful information. In this paper, we propose a programmable aperture camera using a Liquid Crystal on Silicon (LCoS) device. This design affords a high brightness contrast and high resolution aperture with a relatively low light loss, and enables one change the pattern at a reasonably high frame rate. We build a prototype camera and evaluate its features and drawbacks comprehensively by experiments. We also demonstrate four coded aperture applications in defocus deblurring, depth from defocus, light field acquisition, and motion deblurring.

1 Introduction

In the past decades, coded aperture techniques have been studied extensively in optics, computer vision and computer graphics, and a variety of coded aperture techniques have been proposed for various applications. The optimal aperture patterns can be quite different from one application to another. For defocus deblurring, coded apertures are optimized to be broad-band in the Fourier domain [1, 2]. For depth from defocus, coded apertures are optimized to have more zero-crossing frequencies [3, 4]. For multiplexing light field acquisition, an optimal set of aperture patterns are solved for the best signal-to-noise ratio (SNR) after de-multiplexing [5]. Aperture can also be coded in the temporal dimension for motion deblurring [6]. Coded

H. Nagahara (✉) • R. Taniguchi
Kyushu University, 744 Motooka Nishiku, Fukuoka, 819-0001, Japan
e-mail: nagahara@ait.kyushu-u.ac.jp; rin@ait.kyushu-u.ac.jp

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_5

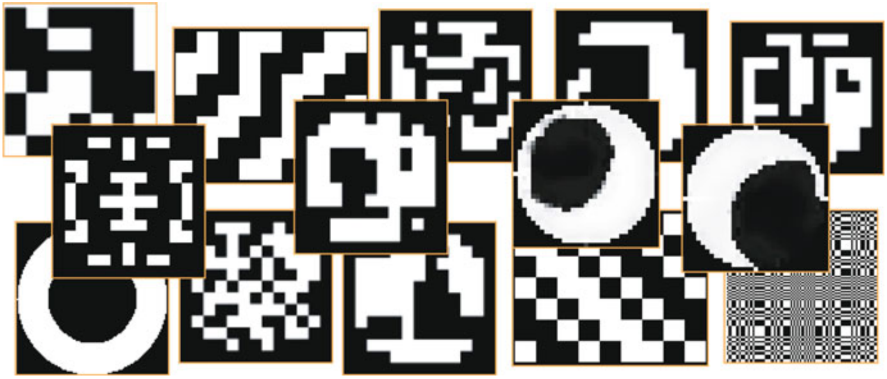


Fig. 1 A variety of coded aperture patterns proposed for various applications

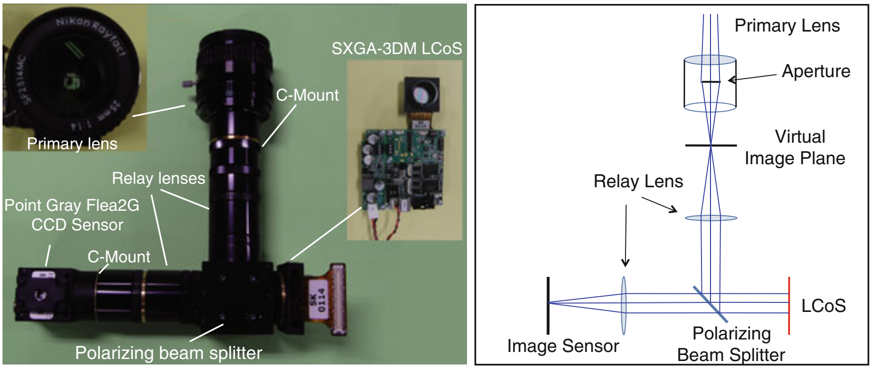
aperture methods have also been used in many other applications, including lensless imaging [7, 8], natural matting [9], etc. Figure 1 shows a collection of some coded apertures that were proposed in the past.

There are many situations where the aperture pattern should be dynamically updated as needed. First, from the aspect of information capturing, ideally aperture pattern should be adaptive to scene contents. For example, the pattern should be optimized for defocus deblurring if the scene has a large depth, and it should be optimized for motion deblurring if the scene has many objects in motion. Secondly, aperture pattern should be adaptive to the specific application purpose. For example, people have shown that a coded aperture optimized for defocus deblurring is often a bad choice for depth from defocus [4]; and multiplexing light field acquisition technique requires different aperture codings for different target angular resolutions. Thirdly, the pattern should be adaptive to the imaging condition. For example, the optimal aperture pattern for defocus deblurring is different at different image noise levels as shown in Zhou's work [2]. In addition, some coded aperture techniques need to capture multiple images with different aperture patterns (e.g., coded exposure [6], coded aperture pair [4] and multiplexed light field acquisition [5]). In all these situations, people need a *programmable aperture camera* whose aperture pattern can be updated at a reasonable speed.

In literature, people has used transmissive liquid crystal displays (LCD) to control aperture patterns [5, 8]. However, the LCD implementation has severe drawbacks. The electronic elements on LCD pixels occlude lights and lead to a low light efficiency. These occluders also cause strong and complicated defocus and diffraction artifacts. These artifacts can be very strong and eliminate the benefits of aperture codings. Consider the popular applications of coded aperture (e.g., defocus deblurring, depth from defocus), we argue that a good programmable aperture is necessary to have the following features:

1. Easy mount. For different applications or scenes, people may use different lenses and sensors. Therefore, it is important to build a programmable aperture that can be easily mounted to different lenses and sensors.
2. High light efficiency. The loss of light leads to decreased SNR. As shown in some papers [2, 10], a high light efficiency is the key to achieve high performance in defocus deblurring, depth from defocus, multiplexing light field acquisition, etc.
3. Reasonable frame rate. Some coded aperture techniques capture multiple images of a scene using different aperture patterns [4, 5]. For dynamic scenes, these techniques require multiple images to be captured within a reasonable short time in order to reduce motion blur, and at the same time, the aperture pattern must also be updated at the same frame rate and be synchronized with the sensor exposure.
4. High brightness contrast. Most optimized aperture patterns in literature have high brightness contrast—many of them are binary patterns. We may fail to display optimized patterns without a high brightness contrast.

To meet these requirements, we proposed a programmable aperture camera [11–13] by using a Liquid Crystal on Silicon (LCoS) device as shown in Fig. 2. LCoS is a reflective liquid crystal device that has a high fill factor (92 %) and high reflectivity(60 %). Compared with transmissive LCD, an LCoS device usually suffers much less from light loss and diffraction. Figure 2 shows the structure of our proposed programmable aperture camera. The use of LCoS device in our prototype camera enables us to dynamically change aperture patterns as needed at a high resolution (1,280 × 1,024 pixels), a high frame rate (5 kHz maximum), and a high brightness contrast. By using the relay optics, we can mount any C-Mount or Nikon F-Mount lens to our programmable aperture camera. Remarkably, our implementation used only off-the-shelf elements and people may reproduce or even improve the design for their own applications.



(a) Our prototype programmable aperture camera (b) The optical diagram of the prototype camera

Fig. 2 Programmable aperture camera using an LCoS device. (**a**) Our prototype LCoS programmable aperture camera. In the *left-top corner* is the Nikon F/1.4 25 mm C-mount lens that is used in our experiments. On the *right* is an LCoS device. (**b**) The optical diagram of the proposed LCoS programmable aperture camera

A detailed description and analysis to our proposed system will be given in Sect. 3. The features and limitations of the present prototype camera are evaluated via experiments in Sect. 4. The proposed coded aperture camera can be a platform to implement many coded aperture techniques. As examples, in Sect. 5, we demonstrate the use of our prototype camera in four applications: defocus deblurring [1, 2], depth from defocus [4], light field acquisition [5] and motion deblurring [13].

2 Related Work

Coded aperture technique was first introduced in the field of high energy astronomy in 1960s as a novel way of improving SNR for lensless imaging of x-ray and γ-ray sources [14]. It is also in the 1960s that researchers in optics began developing unconventional apertures to capture high frequencies with less attenuation. In the following decades, many different aperture patterns were proposed (e.g., apodizations [15–18], and MURA [7]).

Coded aperture research resurfaces in computer vision and graphics in recent years. People optimize coded aperture patterns to be broad-band in the Fourier domain in order that more information can be preserved during defocus for the later deblurring [1, 2]. Levin et al. [3] optimizes a single coded aperture to have more zero-crossings in the Fourier domain so that the depth information can be better encoded in a defocused image. Zhou et al. [4] show that by using the optimized coded aperture pair, they will be able to simultaneously recover a high quality focused image and a high quality depth map from a pair of defocused images. In the work [5], Liang et al. proposed to take a bunch of images using different coded aperture patterns in order to capture the light field.

Coded apertures have also been used for many other applications. Zomet and Nayar propose a lensless imaging technique by using an LCD aperture [8]. Raskar et al. use a coded flutter shutter aperture for motion deblurring [6].

Coded aperture camera can be implemented in several ways. One popular coded aperture implementation is to disassemble the lens and insert a mask, which can be made of a printed film or even a cut paper board [1–3]. The major disadvantages of this method are that one has to disassemble the lens, and that the pattern cannot be easily changed once the mask is inserted. Note that most commercial lenses cannot be easily disassembled without serious damages. People have also used some mechanical ways to modify apertures. Aggarwal and Ahuja propose to split the aperture by using a half mirror for high dynamic range imaging [19]. Green et al. build a complicated mechanical system and relay optics to split a circular aperture into three parts of different shapes [20]. Dowski et al. proposed wave-front coding [21] which places special optical element, called cubic phase plate, on an aperture position for extended depth from field by deblurring.

To dynamically change aperture patterns during capturing, people have proposed to use transmissive liquid crystal display (LCD) devices as in the works [5, 8]. One problem with the LCD implementation is that the electronic elements sit in the

LCD pixels not only block a significant portion of incoming light but also cause significant diffractions. Some custom LCDs are designed to have a higher light efficiency. However, these LCDs usually either have much low resolution (e.g., 5×5 pixels in Liang's implementation [5]) or are prohibitively expensive. We propose to use a reflective liquid crystal on silicon (LCoS) device [22], which has much higher light efficiency and suffers less from diffraction. LCoS has been used before in computer vision for high dynamic range imaging [23]. Another similar device that could be used to modulate apertures is the digital micro-mirror device (DMD). Nayar and Branzoi use a DMD device to control the irradiance to each sensor pixel for various applications, including high dynamic range and feature detection [24]. However, each DMD pixel only has two states and therefore DMD devices can only be used to implement binary patterns. More importantly, existing DMD is restricted mirror angles to ± 12 degrees only (DMD cannot take 0 degree as a mirror angle). If we apply DMD to an aperture in same fashion with our LCoS implementation, the aperture is slanted to an optical axis and coding result is optically different from regular coded aperture because of the restriction.

3 Optical Design and Implementation

We propose to implement a programmable aperture camera by using a liquid crystal on silicon (LCoS) device as an aperture. LCoS is a reflective micro-display technique typically used in projection televisions. An LCoS device can change the polarization direction of rays that are reflected by each pixel. Compared with the typical transmissive LCD technique, it usually produces higher brightness contrast and higher resolution. Furthermore, LCoS suffers much less from light loss and diffraction than LCD does. This is because the electronic components sitting on each pixel of LCD device block lights and cause significant diffraction, and on the contrary, an LCoS device has all the electronic components behind the reflective surface and therefore provides much higher fill factors.

One of our major design goals is to make the primary lens separable from the programmable aperture in order that people can directly attach any compatible lenses without disassembling the lens. To achieve this, we propose to integrate an LCoS device into relay optics.

As shown in Fig. 2, our proposed system consists of a primary lens, two relay lenses, one polarizing beam splitter, an LCoS device, and an image sensor. Only off-the-shelf elements are used in our prototype camera implementation. We choose a Forth dimension display SXGA-3DM LCoS micro-display. Table 1 shows the specifications of this LCoS device. We use two aspherical lenses of 100 mm and 125 mm focal lengths (Edmund Optics, part #447641 and #47642) for each lens of the relay optics. The compound focal length of the lens is 55 mm. We choose a cubic polarizing beam splitter (Edmund Optics, part #49002), and a Point Grey Grasshopper GRAS-14S5C-C camera (2/3″ CCD, 1,384 × 1,036 pixels at 25 fps).

Table 1 Specification of LCoS device

Resolution	1,280 × 1,024 pixels
Reflective depth	8 bits
Pixel fill factor	>92 %
Reflectivity	60 %
Contrast ratio	400:1
Physical dimension	17.43 × 13.95 mm
Switching pattern	40 μs

The camera shutter is synchronized with the LCoS device by using an output trigger (25 Hz[1]) of the LCoS driver.

People have a plenty of freedom in choosing primary lenses for this system. The primary lens and the image sensor are attached to the optics via the standard C-mount. Therefore, a variety of C-mount cameras and lenses can be directly used with this prototype system. SLR lenses (e.g., Nikon F-mount lenses) can also be used via a proper lens adopter. In our experiments, we use a Nikon Rayfact 25 mm F/1.4 C-mount lens.

We can see from Fig. 2b that an incoming light from a scene is first collected by the primary lens and focused at the virtual image plane. A cone of light from each pixel of the virtual image plane is then forwarded by the first relay lens to the polarizing beam splitter. The beam splitter separates the light into S-polarized and P-polarized (linear polarizations perpendicular to each other) lights by reflection and transmission, respectively. The reflected S-polarized light is further reflected by LCoS. The LCoS device can rotate the polarization direction at every pixel by arbitrary degrees. For example, if the pixel on LCoS is set to 255 (8 bit depth), the polarization of the light is rotated by 90° and becomes P-polarized, and then the light will pass through the splitter and reach to the sensor. If the pixel on LCoS is set to 0, the polarization will not be changed by LCoS and the reflected light will be blocked by the splitter.

Consider the LCoS device as a mirror, the diagram in Fig. 2b can be easily shown equivalent to that in Fig. 3. The proposed optics can be better understood from Fig. 3. The virtual image plane is placed f_p from the aperture position of primary lens. The distances from the plane to first relay lens, the lens to LCoS, the LCoS to the second relay lens, and the lens to sensor device are f_r for realizing relay optics. Here, f_p is the focal length of the primary lens and f_r is that of relay lens ($f_p =$ 25 mm, $f_r =$55 mm in our prototype). As a result, the sensor plane is conjugate to the virtual image plane. The LCoS device is relatively smaller than other stops in this optical system and works as the aperture stop.

[1] Note that the LCoS can be modulated at 5 kHz maximum. We use 25 Hz in order that it can be synchronized with the sensor.

4 Optical Analysis and Experimental Evaluation

We have analyzed and evaluated the prototype camera for estimating the specification or current limitations.

4.1 Effective F-Number

Since the LCoS device works as the aperture stop in the proposed system, F-number ($f/\#$) of the primary lens is no longer the effective $f/\#$ of the camera. The actual $f/\#$ of the system is decided by focal length of the relay lens f_r and physical size of LCoS. For a circular aperture, $f/\#$ is usually defined as the ratio of focal length to the aperture diameter. For the rectangle nature of the LCoS, we use $2\sqrt{uv/\pi}$ as the diameter, where (u, v) is the dimension of LCoS. Therefore have:

$$f/\# = \frac{2}{f_r}\sqrt{\frac{uv}{\pi}}. \tag{1}$$

According to Eq. (1), the effective $f/\#$ of the prototype can be computed as $f/2.96$, while the $f/\#$ of the primary lens is $f/1.4$.

4.2 Field of View

Figure 3 shows that the relay system copies the virtual image to sensor plane by a magnification ratio of 1 : 1. Therefore, the field of view (FOV) of the proposed camera is the same as if the sensor were placed at the virtual image plane. The FOV

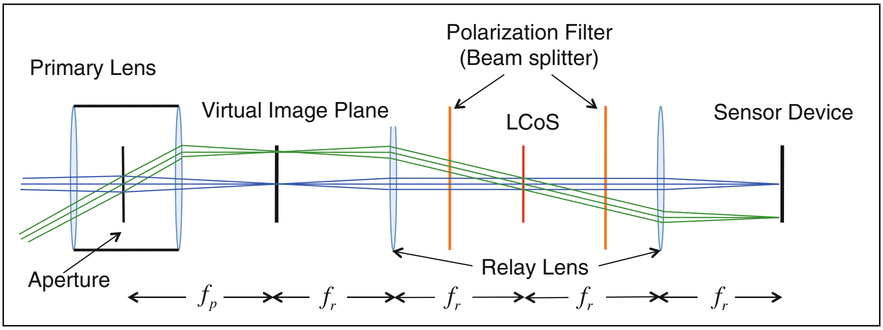


Fig. 3 An equivalent optical diagram to that in Fig. 2b. The virtual image plane and the sensor plane are conjugated by the relay lens. The LCoS is the aperture stop of the system

can be estimated by using the sensor size and the effective focal length of the primary lens:

$$FOV \approx 2\arctan\frac{d}{2f_p}, \tag{2}$$

where d is a diagonal size of the sensor and f_p is the effective focal length of the primary lens.

Our prototype camera uses a 25 mm lens and therefore the camera FOV can be computed as 24.8° according to Eq. (2). Of course, we can change the FOV by using a primary lens with a different focal length.

4.3 Light Efficiency

Light efficiency is one of the most important index in a coded aperture camera. Ideally, the light efficiency of our prototype camera is calculated by:

$$27.6\,\% = 50\,\%(\text{polarization}^{2}) \times 92\,\%(\text{fill factor}) \times 60\,\%(\text{reflectivity}). \tag{3}$$

We notice that many other optical elements in the camera (e.g., a beam splitter, two relay lenses, and an LCoS device) may also attenuate the intensity of captured images. To measure the light efficiency accurately, we captured two images of a uniformly white plane. One image was captured using our prototype camera, and another image was captured without the LCoS aperture (the same sensor and the same lens with $f/\#$ set to 2.8). The ratio of the averaged brightness of these two captured images is computed as 37.85:229.4, which indicates the light efficiency of the system. The light efficiency of our system is about 16.5 %.

The theoretical light efficiency of a transmissive LCD[3] can also be calculated using a similar formula:

$$7.4\,\% = 50\,\%(\text{polarization}^{4}) \times 55\,\%(\text{fill factor}) \times 27\,\%(\text{transmittance}). \tag{4}$$

[2] A polarized beam splitter splits incoming lights based on their polarizations. Although the light interacts with the splitter twice, the light efficiency of beam splitter is still 50 %. This is because 100 % light will pass through the splitter at the second interaction when its polarization is aligned to that of the splitter.

[3] Note that the fill factor or transmittance of the LCD can be slightly different due to different implementations (e.g., physical sizes and resolutions). We assume a typical LCD with a similar physical size and resolution to the LCoS used in our implementation.

[4] An transmissive LCD also needs polarizers which are placed on both side of the LCD panel. The LCD panel just twist a ray polarization as same as LCoS. The difference between LCoS and LCD is reflecting on or transmitting though liquid crystal material for changing the polarization.

The light efficiency of our LCoS implementation is at least three times higher than that of the LCD implementation.

4.4 Vignetting

From the two images captured with and without the LCoS aperture, we can compute the vignetting curves of the prototype camera and a normal camera. The horizontal vignetting curves of our prototype camera and a normal camera are shown in Fig. 4 in red and blue solid lines, respectively. The corresponding dashed lines show the vertical vignetting curves. Figure 4 shows intensity attenuation by vignetting especially on horizontal direction. This is caused by additional relay lens optics. However, once we estimate the effect as shown in Fig. 4 as a look up table, we can easily calibrate it.

4.5 Transmission Fidelity

Another important quality index of a coded aperture implementation is the transmission fidelity—the consistence between the actual transmittance of coded aperture and the input intensity of the LCoS device. To evaluate the transmission fidelity, we captured images of uniformly white plane using circular apertures of different

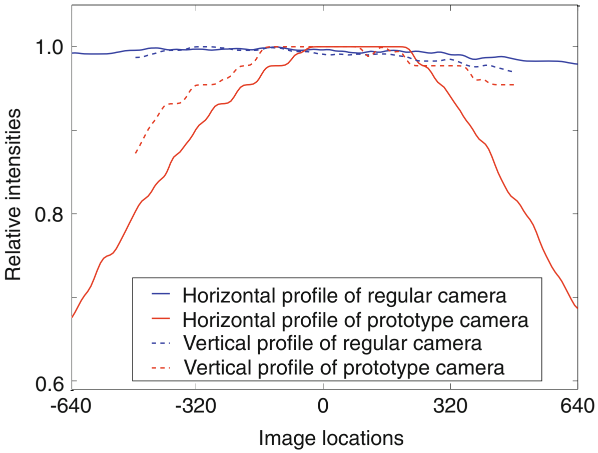


Fig. 4 Vignetting profiles. The *red and blue solid lines* indicate the horizontal vignetting curves of the prototype camera and a regular camera, respectively. The *dashed lines* indicate their vertical vignetting profiles

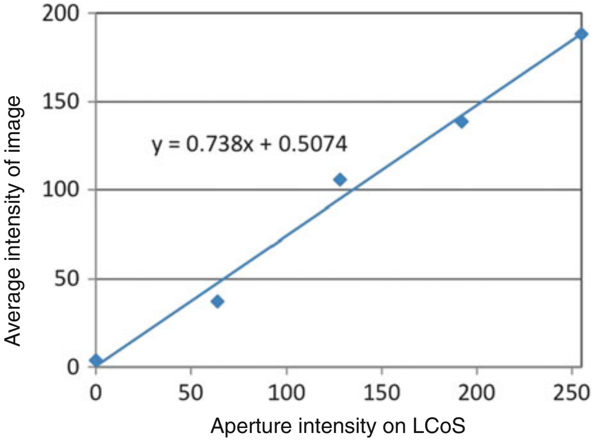


Fig. 5 The aperture transmittance is linear to the LCoS intensity

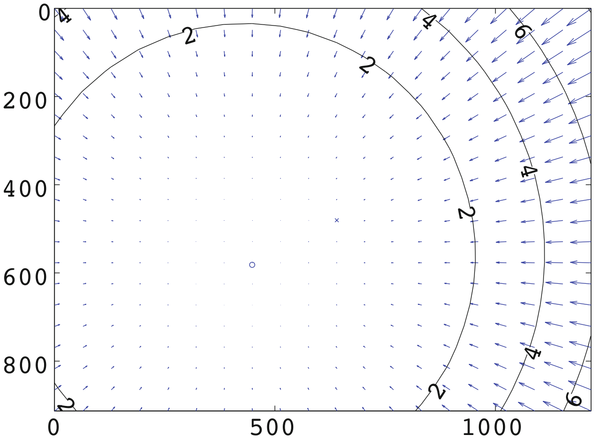


Fig. 6 Geometric distortion due to the use of simple optics

intensities. Figure 5 shows the average intensity of captured images with respect to the input intensities of the circular aperture (implemented using LCoS device). The linear regression result also shows in the plot and the average residual was 0.9942. This plot confirms that the aperture intensity is linear to the actual light transmittance rate.

4.6 *Distortion*

Another problem that has been caused by the use of simple lenses in the relay optics is image distortion. The geometric distortion is calibrated by using the Matlab camera calibration toolbox as shown in Fig. 6. The circle indicates a center of distortion and the arrows represent displacements of the pixel introduced by the

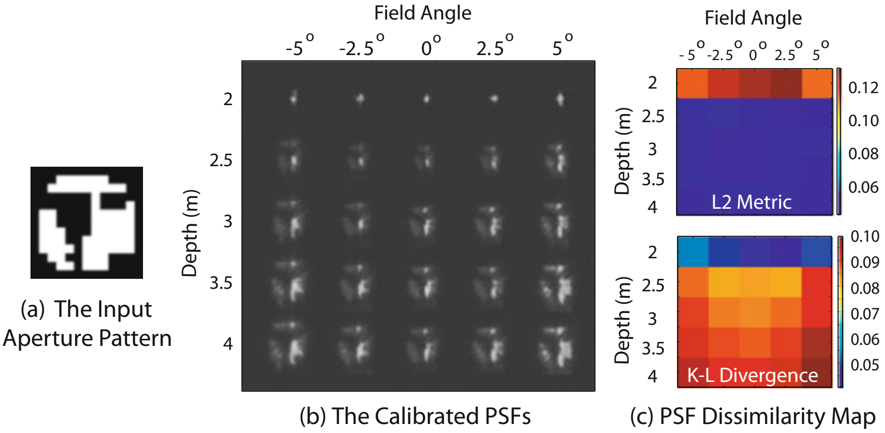


Fig. 7 Evaluating the PSFs of the prototype camera. (**a**) The coded aperture pattern used in the evaluation. This pattern is picked without specific intentions. (**b**) The calibrated PSFs at five depths ranging from 2 to 4 m, and five field angles (image location in viewing angle) ranging from −5 to 5°. We can see that the scale of PSFs varies with both depth and field angle (due to field curvature), while the shape of PSFs appears similar. (**c**) The shape dissimilarity between the input pattern and each PSF is computed according to two metrics: L_2 distance at the *top*, and K-L divergence at the *bottom* (as used in the work [25])

lens distortion. These calibrated camera parameters will be used to compensate the geometric distortions in the captured images.

4.7 PSF Evaluation

Lens aberration and diffraction may distort the actual PSFs. To assess the PSF quality of the prototype camera, we display a coded aperture and then calibrate the camera PSFs at five depths and five different view angles. Without specific intentions, we use the aperture pattern as shown in Fig. 7a in this evaluation. Figure 7b shows how PSFs varies with depth and field angle (image location in viewing angle). The PSFs were captured by the prototype camera with the aperture pattern as shown in Fig. 7a. We set a point light source on various positions in a scene so that we obtain the designated field angles and depths. The captured PSFs were normalized after cropping from an image, then were arranged them into Fig. 7b. We can see that the scale of PSF is related to the field angle. This is because the use of simple lenses in the relay optics leads to a field curvature.

We can see that the shapes of most PSFs are still very similar. To measure the similarity between these PSFs and the input aperture pattern, we normalize the scale of each PSF and compute its L_2 distance to the input pattern. A distance map is shown in the top of Fig. 7c. We can see that according to the L_2 distance, the PSF shape deviation decreases as the blur size increases. It is known that L_2 distance is

Table 2 Specification of the prototype camera

Image resolution	1,384 × 1,036 pixels
Frame rate	25 fps
Minimum F-number	2.96
FOV(diagonal)	24.8° (25 mm Nikon C-mount)
Light transmittance	16.5 %

not a good metric to measure the PSF similarities in defocus deblurring. To measure the dissimilarity between two PSFs, we use the Wiener reconstruction error when an image is blurred with one PSF and then deconvolved with another PSF. This reconstruction error turns out to be a variant of K-L divergence as shown in the work [25]. We plot this dissimilarity map in the bottom of Fig. 7c. We can see that all the dissimilarity values are small and decrease as the blur size decrease.

The specifications of the prototype programmable aperture camera are shown in Table 2 as a summary.

5 Evaluation by Applications

5.1 Programmable Aperture for Defocus Deblurring

Another important limit of most existing coded aperture implementations is that the actual shape of the produced PSF often deviates from the input pattern due to lens aberration and diffraction. Note that the effects of lens aberration and diffraction can be quite different in different lenses. For the complexity of the modern lenses, it is difficult to take these effects into account during pattern optimization. The effects of these imperfections on the optimality of the apertures are often overlooked in the literature.

With a programmable aperture camera, we will be able to evaluate the input aperture pattern by analyzing the captured images, and then improve the aperture patterns dynamically for a better performance. In this experiment, we apply this idea to the coded aperture technique for defocus deblurring.

Zhou and Nayar [2] propose a comprehensive criterion of aperture evaluation for defocus deblurring, which takes image noise level, the prior structure of natural images, and deblurring algorithm into account. They have also shown that the optimality of an aperture pattern can be different at different noise levels and scene settings. For a PSF k, its score at a noise level σ is measured as:

$$R(K|\sigma) = \varSigma \frac{\sigma^2}{|K|^2 + \sigma^2/|F_0|^2}, \tag{5}$$

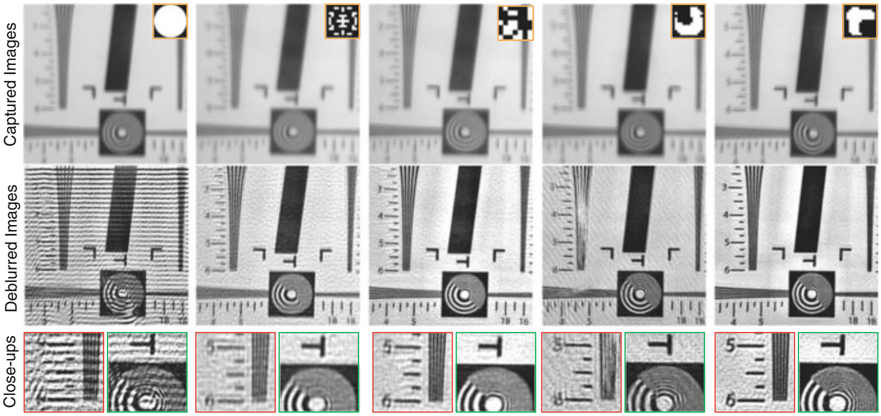
where K is the Fourier transform of the PSF k, and F_0 is the Fourier transform of the ground truth focused image. This definition can be re-arranged as

$$R(K|\sigma) = \Sigma \frac{\sigma^2 \cdot |F_0|^2}{|K|^2 \cdot |F_0|^2 + \sigma^2} \approx \Sigma \frac{\sigma^2 \cdot A}{|F|^2 + \sigma^2} \propto \Sigma \frac{A}{|F|^2 + \sigma^2}, \tag{6}$$

where A is the average power spectrum of natural images as defined in the work [2], and F is the Fourier transform of the captured image. Therefore, given a captured defocused image F, Eq. (6) can be used to directly predict the quality of deblurring without calibrating the PSF and actually performing deblurring, while all the effects of aberrations and diffraction have been taken into account. Obviously, for the best deblurring quality, we should choose the aperture pattern which yields the lowest R value. The detailed discussion of the aperture selection and theoretical background are in Zhou's paper [2].

In our experiment, we capture a set of defocused images of an IEEE resolution chart (shown in the first row of Fig. 8) by using the aperture patterns shown in Fig. 1. We compute the R value from each captured image and find that the lowest R value is achieved by using the pattern shown in Fig. 8e. This indicates that this pattern is the best among all these candidate patterns in the present imaging condition and scene settings.

Note that this prediction is made directly from the observed defocused images without PSF calibration or deblurring. The computation only involves few basic arithmetic operations and one Fourier transform, and therefore can be done at real time. For comparison, the second row of Fig. 8 shows the deblurring results of several different aperture patterns. These results confirm that the pattern in (e) is the best for defocus deblurring in this particular image condition.



(a) Circular Pattern (b) Levin et al. (c) Veeraraghavan et al. (d) Zhou&Nayar σ=0.02 (e) Zhou&Nayar σ=0.03

Fig. 8 Pattern selection for defocus deblurring by using the programmable aperture camera. We capture a set of defocused images of an IEEE resolution chart using the patterns shown in Fig. 1, and evaluate their qualities using Eq. (6). The pattern shown in Column (e) is found to be the best according to our proposed criterion. To verify this prediction, we calibrate the PSFs in all the captured images, do deblurring, and show deblurring results (*the second and third rows*). We can see that the deblurring result in Column (e) is the best, which is consistent with the prediction

5.2 Programmable Aperture for Depth from Defocus

We tested the coded aperture technique for depth from defocus (DFD) proposed by Zhou et al. [4]. It is well know that PSF, which has a lot of zero-crossings in a frequency domain, is preferable for good depth discrimination. Conversely, defocus deblurring requires broadband PSF. Hence, they proposed to use a coded aperture pair for having contradictive properties; one of the pair has zero-crossings, but both of them realize broadband. The programmable aperture camera can be easy to get images with the two different coded apertures.

If we assumed that f_0 is a latent all in focused image, images f_i captured by coded apertures k_i are modeled by:

$$f_i = f_0 \otimes k_i^{d^*} + \eta_i, i = 1, 2, \tag{7}$$

where d^* is an actual scene depth and $k_i^{d^*}$ indicate a scaled version of blurred kernel k_i corresponding to the depth d^*. η is an additive image noise. Equation (7) can be expressed in frequency domain as:

$$F_i = F_0 \cdot K_i^{d^*} + \zeta_i,\ i = 1, 2, \tag{8}$$

where F_i, F_0, $K_i^{d^*}$ and ζ are Fourier transform of f_i, f_0, $k_i^{d^*}$ and η, respectively. Our objective is to find the depth $\hat{d}$ and deblurred image $\hat{F}_0$. We define a objective function for solving this problem as:

$$W^{(d)} = \sum_{i=1,2} |IFFT(\widehat{F_0}^{(d)} * K_i^{\hat{d}} - F_i)|, \tag{9}$$

$$\widehat{F_0} = \frac{F_1 \cdot \bar{K_1}^{\hat{d}} + F_2 \cdot \bar{K_2}^{\hat{d}}}{|K_1^{\hat{d}}|^2 + |K_2^{\hat{d}}|^2 + |C|^2},$$

where $IFFT$ is the 2D inverse Fourier transform. $\bar{K}$ is the complex conjugate of K and $|K|^2 = K \cdot \bar{K}$. C is a reciprocal of signal to noise ratio of the captured images. You can find the detailed descriptions or derivations is in [4]. By minimizing $W^{(d)}(x, y)$ for each pixel, we can obtain the depth map U as:

$$U(x, y) = \arg\min_{d \in D} W^{(d)}(x, y). \tag{10}$$

Also we can find the all in focus image I from the estimated depth map as:

$$I(x, y) = \hat{F}_0^{(U_{x,y})}(x, y). \tag{11}$$

We carried out an experiment of DFD application by using this algorithm. We captured a scene which has three object and backdrop with different depths (a red

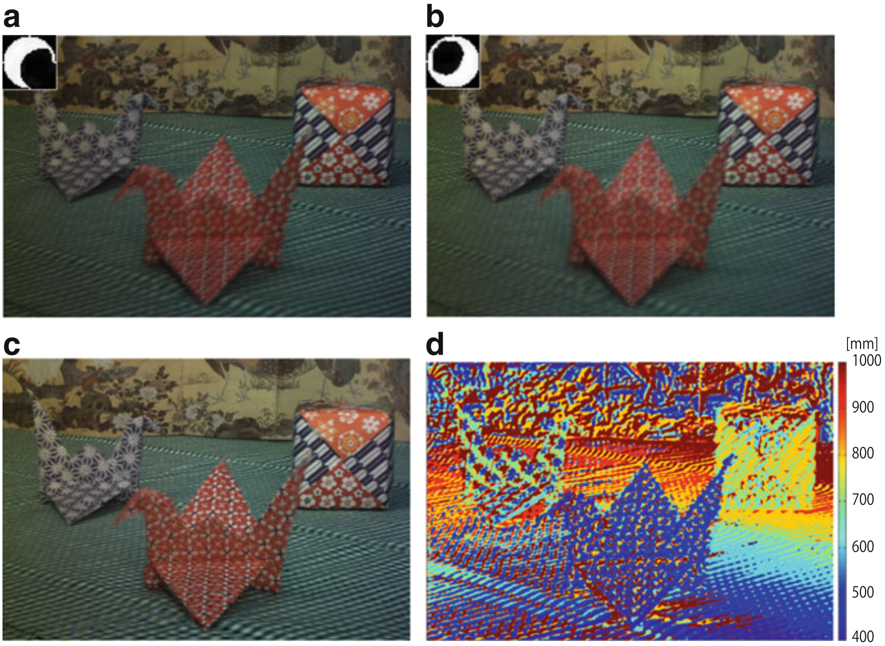


Fig. 9 Depth from defocus by using coded aperture pair. Input images (**a**), (**b**) captured by coded apertures indicated on the insets. The deblurred image (**c**) and depth map (**d**) are estimated by the proposed method. (**a**) Input image f_1. (**b**) Input image f_2. (**c**) Deblurred image. (**d**) Estimated depth map

crane was on 500 mm, a blue crane and cube were on 700 mm and a backdrop was on 1,000 mm from the camera) as shown in Fig. 9. Figure 9a,b shows the input images captured by a coded aperture pair. The insets on top left of the images indicate the shapes of coded apertures when the images were captured. Figure 9c,d shows the estimated results of the deblurred image and the depth map. Figure 10 shows a similar experimental result by using conventional circular apertures with different radii for comparison. Figure 10d shows the circular DFD cannot estimate a scene depth. Notice that we did not use any post-processing, such as BP or Graph Cut etc. and the both depth maps were raw pixel-wise estimations described by Eq. (10), so that we can easy to understand the coding difference. As a result, there are a lot of artifacts, ringing edge and remaining blur, in the deblurred image as shown in Fig. 10c, since the image were deblurred by different size of kernels. On the other hand, you can see that in-focused radiance of the scene and the scene depths were correctly recovered as shown in Fig. 9c,d. We confirmed that prototype camera can be applicable for DFD application and the coded aperture pair has an advantage to the conventional one. In the paper [4], they replaced the SLR lenses with different aperture masks for taking input images of DFD. The programmable aperture camera is feasible for the DFD application, since the camera easy and quickly changes the patterns at video rate.

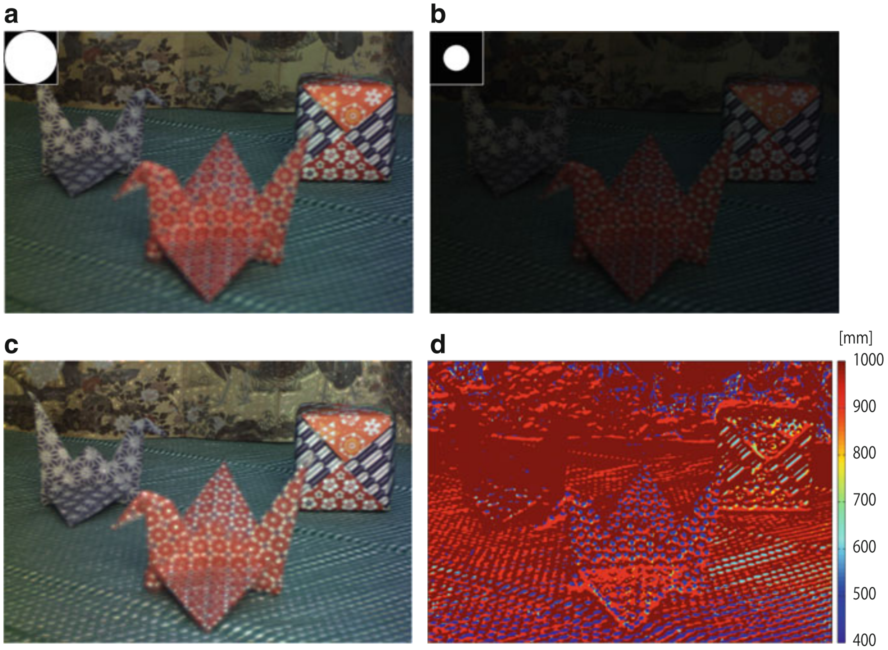


Fig. 10 DFD by a conventional circular aperture pair. (**a**) Input image f_1. (**b**) Input image f_2. (**c**) Deblurred image. (**d**) Estimated depth map

5.3 Programmable Aperture for Light Field Acquisition

We finally use our prototype programmable aperture camera to re-implement the multiplexing light field acquisition method, which is first proposed by Liang et al. [5] A 4D light field is often represented as $l(u, v, x, y)$ [26], where (u, v) is the coordinates on the aperture plane and (x, y) is the coordinates in the image plane.

For a light field acquisition technique using coded aperture, the spatial resolution in the (x, y) space is simply determined by the sensor resolution and the angular resolution in the (u, v) space is determined the resolution of coded apertures. Bando et al. [9] use a 2×2 color coded aperture to capture light fields and then use the information to do layer estimation and matting. Liang et al. [5] propose a multiplexing technique to capture light fields up to 7×7 angular resolution. For any $m \times n$ angular resolution light field acquisition, the multiplexing method requires $m \times n$ images captured using $m \times n$ different coded apertures.

With our prototype programmable aperture camera, it is easy to capture light fields with various angular resolutions. We use S-matrix for the multiplexing coding (see [27] for a deep discussion on the multiplexing coding). Figure 11(top) shows

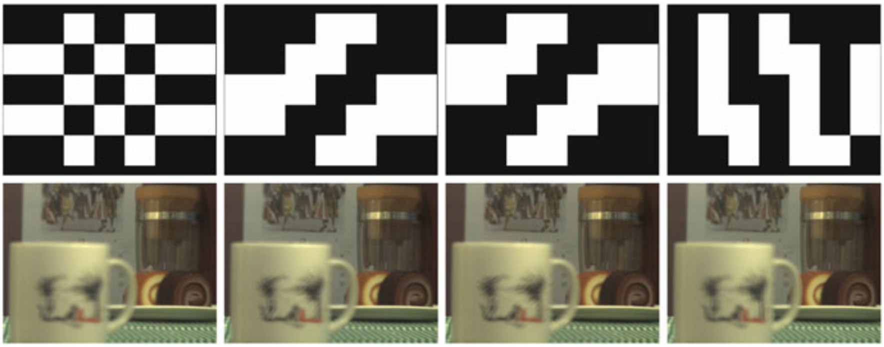


Fig. 11 Four multiplexing aperture codings and the corresponding captured images. *Upper row* shows four of the 31 aperture patterns that we generate from an S-Matrix. *Bottom row* shows the four corresponding captured images

four of the 31 aperture patterns[5] that we generate from an S-Matrix. Since the aperture pattern of the prototype camera can be updated at a video frame rate (25 fps), it only takes 1.2 s to capture all of the images. If we could increase the camera frame rate further or lower the aperture resolution, the programmable aperture camera could be able to capture light fields of moving objects.

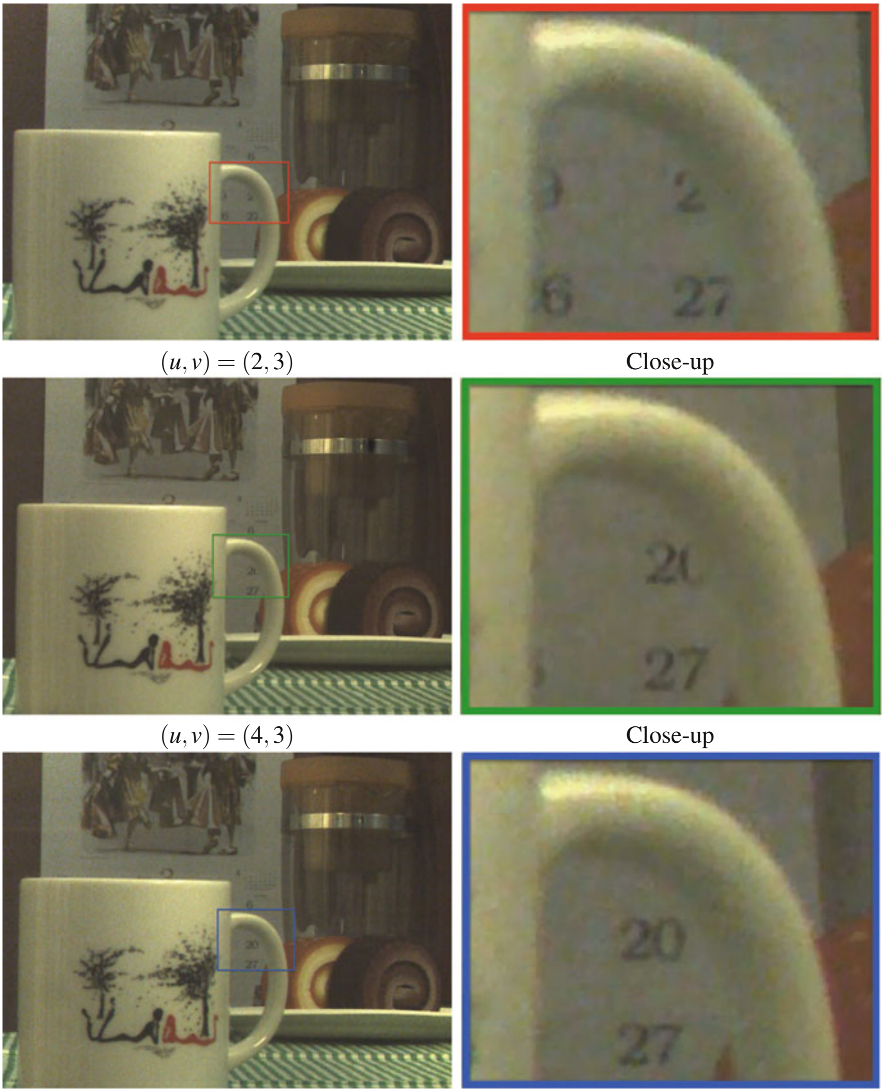
From the 31 captured images, we recover the light field of resolution 1,280 × 960 × 31 (7 × 5 (u,v) resolution excluding the four corners). Figure 12 shows the images for different viewpoints (u, v) and their close-ups. From the close-ups, we can see the disparities of the text clearly. With the recovered light field, people will be able to do further post-processing including depth estimation and refocusing as shown in literatures [5, 9].

5.4 Programmable Aperture for Motion Deblurring

Motion blur in an image is the result of either camera shake or object motion in a scene. When an object moves or the camera shakes during an exposure, the captured image contains blur caused by these motions, since the obtained image is a superimposition of the different positions of the objects at the different times. we propose to use a programmable aperture for motion deblurring [13].

In a conventional photograph, objects moving at different speeds cause varying degrees of motion blur with different shapes and lengths. To remove such motion blur, we must estimate the speed of each object. To address this problem, Levin et al. [28] realized motion-invariant photography that makes the PSF invariant to motion through the use of parabolic camera motion during an exposure. Because of this

[5] This is because the code length of S-matrix must be $2^n - 1$.



$(u,v) = (2,3)$ Close-up

$(u,v) = (4,3)$ Close-up

$(u,v) = (6,3)$ Close-up

Fig. 12 The reconstructed 4D light field. Images from three different view points (u, v) are generated from the reconstructed 4D light field, and their close-ups are shown in their *right*. From the close-up images, we can see the disparities that the numbers, 20 and 27, on the backside calender is moving left w.r.t. the frontal mag handle position

invariance, we can remove all blur for all moving objects by deconvolution using a single PSF. In Levin's work [28], the obtained PSF that is invariant to motion is expressed as:

$$\phi(x) = \frac{\lambda(x)}{2T\sqrt{s_i^2 - 2a_i x}}, \tag{12}$$

$$\lambda(x) = \begin{cases} 2, & \frac{s_i^2}{2a_i} \le x < s_i T - a_i T^2, \\ 1, & s_i T - a_i T^2 < x < s_i T + a_i T^2, \\ 0, & \text{otherwise,} \end{cases}$$

assuming that the image has acceleration a_i derived from camera motion and object velocity s_i derived from object motion. Both a_i and s_i are described in the image space. Here x is the position in the image and $2T$ is the exposure time. These assumptions are expressed as:

$$x(t) = s_i t + \frac{a_i t^2}{2}. \tag{13}$$

For detailed derivations of Eqs. (12) and (13) the reader is referred to [28]. We extend the concept of motion-invariant photography to be realized by a coded aperture.

The center of the camera aperture is the center of the projection in the projective geometry, and a projective change can thus be realized by motion of this aperture position. We realize the virtual camera motion needed for motion-invariant photography by temporally changing the aperture patterns. Figure 13 shows the projective geometry of the programmable aperture camera. For simplicity, Fig. 13 shows the geometry on the $X - Z$ space, where a point in the scene is $P(X, Z)$, the aperture is located in the plane $Z = 0$, and a point on the image space $(Z = -Z_p)$ is $p(x)$. The lens focal distance is denoted as f. The distance Z_p between the lens plane and focal point Q can then be expressed as

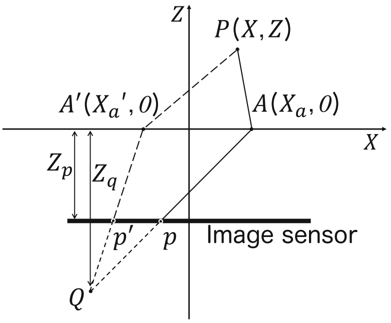


Fig. 13 Virtual camera motion model for a programmable aperture

$$\frac{1}{f} = \frac{1}{Z} + \frac{1}{Z_q}. \tag{14}$$

By setting the position of the pinhole aperture as $A(X_a, 0)$, as shown in Fig. 13, the ray radiating from P goes towards the focal point Q through lens refraction via A and is projected onto point p on the image. The relation of the projection of projective point p can be modeled as

$$x(t) = \alpha X(t) - \beta X_a(t), \tag{15}$$

where

$$\alpha = \frac{-Z_p}{Z},$$
$$\beta = Z_p \left(\frac{1}{f} - \frac{1}{Z} - \frac{1}{Z_p} \right).$$

From Eqs. (13) and (15), we obtain the aperture motion as

$$X_a(t) = -\frac{a_a}{2\beta} t^2 + \frac{\alpha X(t) - s_i t}{\beta} \tag{16}$$

Equation (16) shows that aperture motion is parabolic motion. We realizes the parabolic motion of apeture by sequential patterns of pinholes placed at different positions.

Since our proposed method is inspired from the camera motion in Levin's work [28], the limitation that the object motion must have constant acceleration and one-dimensional horizontal motion corresponding to the camera motion is also the same as in that work.

We carried out some real experiments as shown in Figs. 14, 15, 16 and 17. We used toy trains as moving objects. We set a backdrop for the background, the far side railroad track, the near side railroad and a miniature car at a distance of 505 mm, 500 mm, 495 mm and 490 mm, respectively, so that these objects appeared in the DOF.

Since we set the camera exposure time to 45 ms, motion blur appeared with a length of about 45 pixels(the far side train) and 63 pixels(the near side train) in a normal photograph. (The length was obtained from Fig. 14.) We coded the motion blur using this acceleration. Figure 15 shows the image captured by our motion-invariant photography. In this figure, both the moving trains and the static background are blurred. We used the measured PSF, which was captured in advance, for deconvolution. (For deconvolution, we used BM3D deconvolution proposed by Dabov et al. [29].) Figure 16 shows the restoration result obtained by deconvolution of the captured image (Fig. 15) with the single measured PSF. This figure shows that we can reduce motion blur through deconvolution since the edge of the image is sharper than that captured by normal photography (Fig. 14). In addition, we reduced

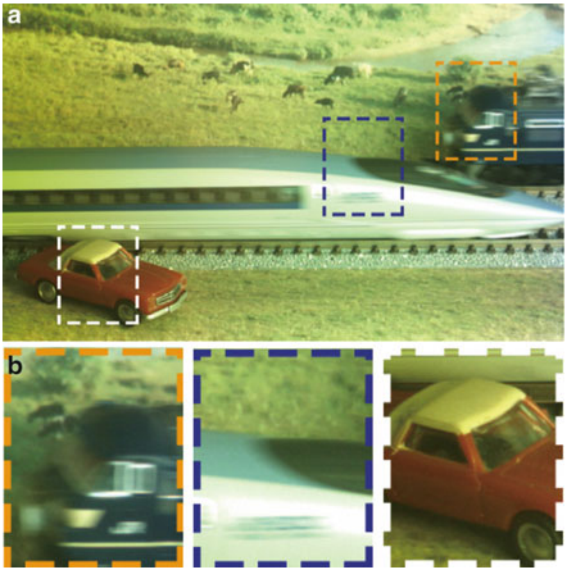


Fig. 14 Image captured by normal photography. (**a**) Entire image. (**b**) Magnified images (*left and center*: moving object, *right*: static object)

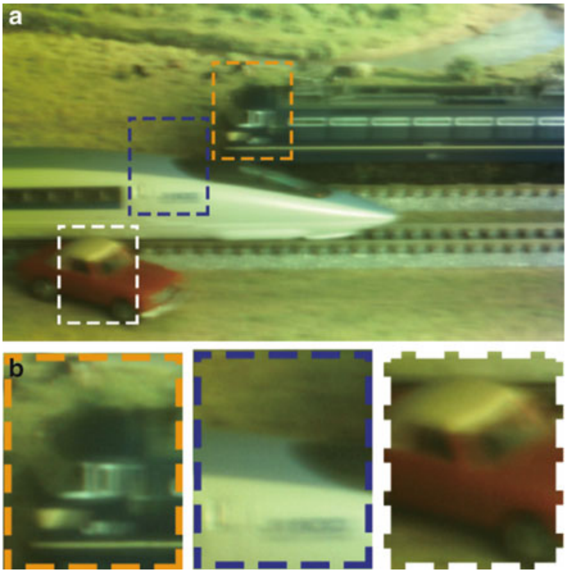


Fig. 15 Blurred image recorded by motion-invariant photography. (**a**) Entire image. (**b**) Magnified images (*left and center*: moving object, *right*: static object)

blur of the static background equally without motion estimation or segmentation. We also compared our result with a simple short exposure photograph. Figure 17 depicts the image obtained through short exposure imaging with the exposure time set to 0.679 ms so as to ignore all motion in the scene. In this case we used F2.8, which is the maximum radius setting of the lens. It can be seen that the short exposure results in a noisy image with loss of gradation, because the amount of

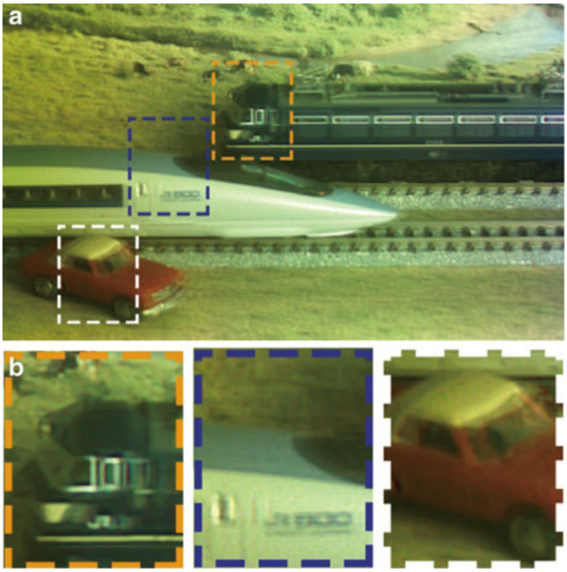


Fig. 16 Deconvolved image. (**a**) Entire image. (**b**) Magnified images (*left and center*: moving object, *right*: static object)

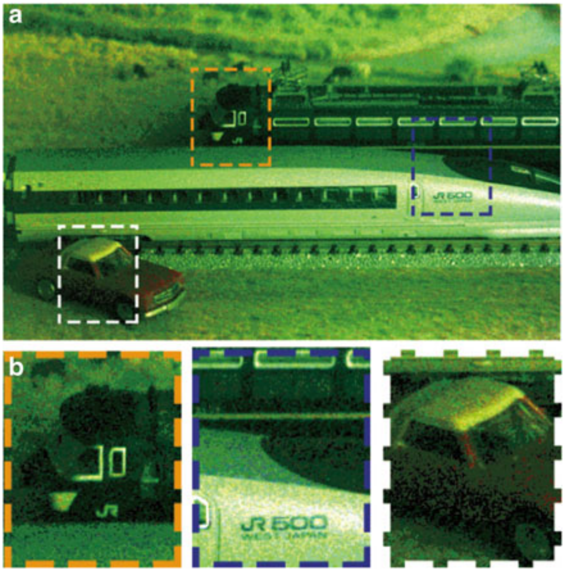


Fig. 17 Image captured using short exposure. (**a**) Entire image. (**b**) Magnified images (*left and center*: moving object, *right*: static object)

light is reduced and noise is emphasized by the intensity adjustment. Our coded and deconvolved result yields a better image, which is both sharper than the blurred image and brighter than the short exposure image.

6 Conclusion and Perspectives

In this paper, we propose to build a programmable aperture camera using an LCoS device which enables us to implement aperture patterns of high brightness contrast, light efficient and resolution at a video frame rate. Another important feature of this design is that any C-Mount or F-Mount lenses can be easily attached to the proposed camera without being disassembled. These features make our design applicable to a variety of coded aperture techniques. We demonstrate the use of our proposed programmable aperture camera in four applications: defocus deblurring, coded aperture pair for DFD, multiplexing light field acquisition, and motion deblurring.

Acknowledgements This work was supported by KAKENHI 21680018 (Grant-in-Aid for Young Scientists (A)), KAKENHI 20650023 (Grant-in-Aid for challenging Exploratory Research), Microsoft Core8 program.

References

1. Veeraraghavan A, Raskar R, Agrawal A, Mohan A, Tumblin J. Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Trans Graph. 2007;26:Article No. 69
2. Zhou C, Nayar S. What are good apertures for defocus deblurring? In: International conference of computational photography, San Francisco; 2009.
3. Levin A, Fergus R, Durand F, Freeman W. Image and depth from a conventional camera with a coded aperture. Proc ACM SIGGRAPH. 2007;26:70.
4. Zhou C, Lin S, Nayar S (2009) Coded aperture pairs for depth from defocus. In: Proc. international conference on computer vision, Kyoto; 2009.
5. Liang CK, Lin TH, Wong BY, Liu C, Chen H. Programmable aperture photography: Multiplexed light field acquisition. ACM Trans Graph. 2008;27:Article No. 55.
6. Raskar R, Agrawal A, Tumblin J. Coded exposure photography: motion deblurring using fluttered shutter. ACM Trans Graph. 2006;25:795–804.
7. Gottesman S, Fenimore E. New family of binary arrays for coded aperture imaging. Appl Opt. 1989;28:4344–52.
8. Zomet A, Nayar S. Lensless imaging with a controllable aperture. In: Proceedings of the computer vision and pattern recognition; 2006. pp. 339–46.
9. Bando Y, Chen B, Nishita. Extracting depth and matte using a color-filtered aperture. ACM Trans Graph. 2008;27:134.
10. Hasinoff S, Kutulakos K, Durand F, Freeman W. Time-constrained photography. In: Proc. international conference on computer vision; 2009. pp. 1–8.
11. Nagahara H, Zhou C, Watanabe T, Ishiguro H, Nayar S. Programmable aperture camera using LCoS. In: Proceedings of the European conference on computer vision. No. LNCS6316, vol. 2; 2010. pp. 337–50.
12. Nagahara H, Zhou C, Watanabe T, Ishiguro H, Nayar S. Programmable aperture camera using LCoS. IPSJ Trans Comput Vis Appl. 2012;4:1–11.
13. Sonoda T, Nagahara H, Taniguchi R. Motion-invariant coding using a programmable aperture camera. ISPJ Trans Comput Vis Appl. 2014;6:25–33.
14. Caroli E, Stephen J, Cocco G, Natalucci L, Spizzichino A. Coded aperture imaging in X-and gamma-ray astronomy. Space Sci Rev. 1987;45:349–403.
15. Welford W. Use of annular apertures to increase focal depth. J Opt Soc Am A. 1960;50:749–53.

16. Mino M, Okano Y. Improvement in the OTF of a defocused optical system through the use of shaded apertures. Appl Opt. 1971;10:2219–25.
17. Varamit C, Indebetouw G. Imaging properties of defocused partitioned pupils. J Opt Soc Am A. 1985;2:799–802.
18. Ojeda-Castafieda J, Andres P, Diaz A. Annular apodizers for low sensitivity to defocus and to spherical aberration. Opt Lett. 1986;11:487–9.
19. Aggarwal M, Ahuja N. Split aperture imaging for high dynamic range. Int J Comput Vis. 2004;58:7–17.
20. Green P, Sun W, Matusik W, Durand F. Multi-aperture photography. Proc. ACM SIGGRAPH. 2007;26:Article No.68.
21. Dowski ER, Cathey WT. Extended Depth of Field through Wave-front coding. Appl Opt. 1995;34(11):1859–66.
22. Wikipedia. Liquid crystal on silicon. (http://en.wikipedia.org/wiki/Liquid_crystal_on_silicon).
23. Mannami H, Sagawa R, Mukaigawa Y, Echigo T, Yagi Y. High dynamic range camera using reflective liquid crystal. In: Proceedings of the international conference on computer vision; 2007. pp. 14–20.
24. Nayar SK, Branzoi V, Boult T. Programmable imaging: Towards a flexible camera. Int J Comput Vis. 2006;70:7–22.
25. Nagahara H, Kuthirummal S, Zhou C, Nayar S. Flexible depth of field photography. In: Proceedings of the European conference on computer vision, vol. 3; 2008.
26. Levoy M, Hanrahan P. Light field rendering. In: Proceedings of the ACM SIGGRAPH; 1996. pp. 31–42.
27. Schechner Y, Nayar S, Belhumeur P. A theory of multiplexed illumination. In: Proceedings of the international conference on computer vision, vol. 2; 2003. pp. 808–15.
28. Levin A, Sand P, Cho PS, Durand F, Freeman WT. Motion-invariant photography. ACM Trans Graph. 2008;27(3):71.
29. Dabov K, Foi A, Egiazarian K. Image restoration by sparse 3D transform-domain collaborative filtering. In: Proc. SPIE; 2008.

Exploratory Visual Analytics for Winter Road Management Using Statistically Preprocessed Probe-Car Data

Yuzuru Tanaka, Hajime Imura, and Jonas Sjöbergh

Abstract Social CPSs (Cyber-Physical Systems) denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure systems. They utilize both cyber data stored in databases and physical data coming from sensor networks in the target physical world for the analysis and optimized control of urban infrastructure systems such as traffic, energy, and water services. This paper focuses on the winter road management in Sapporo where we have the world biggest annual snow fall among the cities with more than 1 million populations. For monitoring the road conditions over the whole city, the use of probe-car data without violating personal data protection is fundamental. This paper first shows that probe car data statistically preprocessed over road links for an urban-scale area still allow us to visualize the dynamic change of the traffic flow in terms of the divergence and flow vector field. These give us sufficient information about the dynamic change of hotspots of traffic, main traffic streams, and route selection preference. The paper also shows more complex and advanced analyses of such data, especially for better winter road management in Sapporo. We extend the well-known multiple coordinated views framework for exploratory visual analytics to multiple coordinated views and analyses by integrating analysis tools with their result visualization views into the same environment. These newly added views may also coordinate with others, and allow users to directly select clusters or mined patterns calculated at runtime to further quantify the underlying database view. Exploratory visual analytics with such an environment enables us to detect road links for effective pinpoint snow removal.

Y. Tanaka (✉)
Graduate School of Information Science and Technology, Hokkaido University,
Kita 13 Nishi 8, Kita-Ku, Sapporo, Hokkaido, Japan
e-mail: tanaka@meme.hokudai.ac.jp

H. Imura • J. Sjöbergh
Meme Media Laboratory, Hokkaido University, Kita 13 Nishi 8, Kita-Ku,
Sapporo, Hokkaido, Japan
e-mail: hajime@meme.hokudai.ac.jp; js@meme.hokudai.ac.jp

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_6

1 Introduction

Advances of sensor devices and their networking technologies have enabled the real-time monitoring of physical systems ranging from small-size healthcare devices to large-scale plant systems, or even to urban-scale infrastructure service systems. Social CPSs (Cyber-Physical Systems) denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure systems. A CPS originally denotes an integration of a control system and a computer system that utilizes both cyber data stored in files or databases in the computer system and physical data coming from sensor networks in the target physical system for the sand boxing of embedded medical devices and the optimized operation of large plant systems. A social CPS also utilizes both cyber data stored databases and physical data coming from sensor networks in the target physical world for the analysis and optimized control of urban infrastructure systems such as traffic, energy, and water services. This paper focuses on the winter road management in Sapporo, where we have the world biggest annual snowfall among the cities with more than 1 million populations. Since the targets of Social CPSs are necessarily complex systems of systems, they need to deal with a variety of heterogeneous cyber and physical, real-time and retrospective big data including, for example, probe-car data, weather data, snow removal records, and traffic accident records.

One of the problems we are facing these days in big data R&Ds may be a big gap between the core technology R&Ds and the application R&Ds especially for complex target systems. Through the involvement in a Japanese government-initiative project from 2012 to 2016 on social cyber-physical systems for optimizing social system services such as the snow plowing and removing in Sapporo City, we have been facing difficulties filling the gap between varieties of available data analysis methods and the goal of finding optimized resource schedulings for snow plowing and removing.

For the analysis of dynamically changing traffic and road conditions in an urban-scale area, probe-car data may play the most important role. Inherently they are real time data, and have the potential to cover urban-scale areas. They can tell us not only about dynamically changing traffic and road conditions, but also about people's dynamically changing mobility demands and/or activities.

With the increasing use of car navigation systems networked to service centers, urban-scale or even nationwide scale probe car data are accumulated at service providers and ready for advanced meaningful analyses. However, their advanced utilization is currently facing two major obstacles. First, they are accumulated in different silos belonging to different service providers, and are not open for any analyses by third parties. Second, there are privacy concerns since each trajectory obtained from probe-car data may reveal lots of personal information about its driver, such as his or her home location, office location, frequently visited places like malls, and visits to some acquaintances. The second obstacle is often one of the main reasons for service providers not to open those data silos to public use of their probe-car data. The service providers are generally too cautious in providing their probe-car data to third parties even for nonprofit utilization. On the other hand, probe-car data can inherently give us lots of information about traffic and road conditions, route selection preference, travelling time, and daily or seasonal dynamic change of population mobility in urban-scale, or even nationwide-scale areas. Their utilization is fundamental in the analysis and optimization of the sustainable and safe management of urban infrastructure services such as energy supply and transportation. IT-based management systems for this are sometimes called social cyber-physical systems. IBM and Microsoft, for example, call such systems smarter city systems and urban computing systems, respectively.

There may be two known practical solutions to ensure the protection of personal information when using probe car data. One is to remove the head and tail of each trajectory for some constant time interval or for some constant distance to hide the origin and destination. Such modification can also further fragment the remaining part of each trajectory into segments of some constant time interval or of some constant distance. Another approach is to provide only statistically processed probe car data including for example the average speed, the maximum speed, and the number of cars in each direction of each road link during every time interval of a fixed length, e.g., 5 min. However, it has not been well discussed what kind of statistical data enables what kind of analysis. We need mathematical research on this subject. This paper will first show that we can calculate the divergence and the flow-vector field of the traffic only using the dynamic change of the average speed and the number of cars in each direction of each road link during each time interval of a fixed length. Using statistically preprocessed probe-car data of occupied taxi cars, we can visualize the temporal change of the hotspots for taking taxis and getting-off taxis during a day. We can also observe the temporal change of route preference depending on the traffic jams caused by heavy snowfalls through flow-vector field visualization of the traffic.

In the latter half of this paper, we will deal with more complex analyses for better winter road management in Sapporo. Probe-car data are fundamental to monitor and estimate the urban-scale dynamic change of traffic and road conditions of all the road links, and also to monitor the snow removal operations. We can use probe data from private cars and taxis for the former purpose, and from snow plowing and removing vehicles for the second purpose. Sapporo has 1.9 million citizens and an average annual snowfall of 6 m. It spends more than 150 million US dollars every winter just for snow plowing and removing. Our preparatory study on the influence of snowfall and snow removal operations on traffic revealed that the effect is not uniform even among consecutive road links along the same main route in the central city area. Because of this heterogeneity, we believe that macro analysis applied to the whole urban area cannot give us any meaningful result. One of the possible solutions may be exploratory visual analytics that enables us to freely repeat micro analyses, consisting of hypothesis making and leading to improvisational data segmentation and hypothesis checking through improvisational data analysis and visualization. This paper introduces the Geospatial Digital Dashboard as an open system for such exploratory visual analytics. This system exploits a coordinated multiple views framework, which is a well-known framework for exploratory visual analytics, and extends this framework to integrate analysis tools with their result visualization views as additional coordinated views. During the last 1 year and a half of our project, we have focused on the detection of those directed road links which seem to have caused severe traffic jams in nearby road links. This paper shows that exploratory repetition of micro analyses using the Geospatial Digital Dashboard enables us to detect such hotspot road links.

2 Related Research

Personal information might be handled in many different ways in the probe vehicle system [17]. Even if each trajectory data is anonymized there are many possibilities to identify the same vehicle among a large number of different trajectories through behavioral analysis. ISO/TC204/WG16 published international standard about personal data protection in probe vehicle system as ISO 24100 Intelligent transport systems—Basic principles for personal data protection in probe vehicle information services in 2009 [7]. Even if data cannot identify an individual directly, if it can do so indirectly it should be regarded as personal data, and as a target of protection. While the fragmentation of trajectory data into anonymized pieces of data may protect the privacy, trajectories cannot be reconstructed anymore, i.e., the data lose traceability. Some solutions are proposed ranging from practical ones either to trim both the head and the tail of each trajectory or to provide only statistical data for each road link, to more theoretical ones [1, 6, 13, 21]. The first one [1] protects privacy by shifting trajectory points in space that are already close to each other in time. Clusters of k trajectories are enforced to be close to each other so that they fall in the same area of uncertainty. In the second one [6], privacy is preserved by removing some points such that uncertainty between consecutive points is increased to avoid identification. Both of these assume the allowance to bring some uncertainty into trajectory data. Work in [21] limits the probability of disclosing the tail of the trajectories given the head of the trajectories. The last one [13] extends the concept of k-anonymity to trajectory data.

This paper focuses on one of the practical approaches: giving only statistically preprocessed data. While this lacks traceability, we show that we can still calculate the divergence rate and flow vector field of the traffic to analyze the hotspots and the dynamic changes of the traffic flow and the route preference.

There are many systems for visualization and exploration of data that use coordinated multiple views [16]. However, to the best of our knowledge, there are no other systems that provide an interactive visualization environment with coordinated multiple views and advanced analysis tools. We believe that allowing advanced analysis components in such an environment to create new complex data at runtime, and making such components coordinated with other views like any other visualization view components, are useful extensions of the coordinated multiple views framework. We have not seen any systems that allow interaction with visualization results that then affect the input to such analysis components, and that allow further visualization of and interaction with the results, where changes in any coordinated views affect all the other coordinated views and analyses.

There are systems that use graphical interfaces such as a flow chart style interface to set up what types of preprocessing, analysis, data mining, and visualization etc. should be done. Changes in the flow chart lead to updated visualizations, but interaction with the actual visualization results is very limited or not possible. Examples of such systems include RapidMiner [12] and DEVise [11].

There are systems that use coordinated multiple views of data and allow interaction with the visualization results. Selecting data in one view updates the other views to show only this selection, or clicking on one item in one view shows details about it in another view, etc. Creating new complex data in analysis components is, however, not possible. Examples of such systems include: SpotFire [2] Tioga-2 [3] (now TiogaDataSplash), and Snap-Together Visualization [14].

One example system quite similar to the Geospatial Digital Dashboard described in this paper is KNIME [4], which uses a graphical flow chart to set up processing and visualization of data. It allows adding new user built components and uses multiple linked views for visualization. Selecting subsets of data in one view highlights these in other views, but unlike our system, it does not trigger recalculation of related data mining results.

Another system with many of the features of Geospatial Digital Dashboard is Orange [5]. It sets up data flows in a flow chart, has both visualization and analysis components, allows user built components, and selections in a visualization component can trigger recalculation in data mining components. Unlike our system, two components cannot feed back into each other, so selections in one component can affect the other, but selections in the other component cannot be reflected back to the first.

There are also some precursors to our system. The VERD [18] system is based on IntelligentBox, a 3D version of Meme Media [19]. It visualizes relational databases and allows Web resources to be treated as relational schema. Unlike our system, visualizations are set up in a flow and interaction with visualization results only affects other visualizations in the downstream of the interaction point.

3 Divergence and Flow-Vector Field Analyses of Urban-Scale Traffic

Let us consider a small geographical area S_i. Let $P_i(t)$, $D_i(t)$, $G_i(t)$, $A_i(t)$ respectively denote the population, the divergence rate, the generation rate, and the absorption rate at time t of this area S_i. The population is the number of objects, e.g., cars in this case, in this area at time t. The divergence rate is the difference between the rate of the outbound flow from this area and the rate of the inbound flow to this area at time t. The generation rate and the absorption rate are respectively the increase rate of the newly generated objects and the decrease rate of the newly absorbed objects in this area at time t. For a sufficiently small time interval δ, the following holds:

$$P_i(t+\delta) - P_i(t) = (-D_i(t) + G_i(t) - A_i(t))\delta$$

There are two typical cases of mobility data. In case of mobile phone mobility data, we may assume that mobile phones are usually kept on, which implies that $G_i(t) = A_i(t) = 0$. Therefore, the following holds:

$$D_i(t)\delta = P_i(t) - P_i(t+\delta)$$

This implies that the divergence can be calculated simply from the time variation of the population during δ. In case of cars, we may assume that the flow is rather stable during a sufficiently small time interval, i.e., $P_i(t+\delta) \approx P_i(t)$. This implies the following:

$$D_i(t) \approx G_i(t) - A_i(t)$$

Let us consider the case of taxi cars. Suppose that the probe-car data are available with the distinction of empty cars and occupied cars. For the flow of occupied taxi cars, the rate $G_i(t)$ denotes the rate of taking a taxi in the area S_i, while the rate $A_i(t)$ denotes the rate of getting off a taxi in this area. The assumption $Pi(t+\delta) \approx P_i(t)$ means that the population in each area is stable and does not change during a sufficiently small time interval.

Suppose we have statistically processed probe data of occupied taxi cars showing the number of cars and the average speed in each road link at every 5 min. For a sufficiently small area S_i, every road link may cross the border of this area at one point, two points, or no point. The last case may have two cases, i.e., the whole road link may lie either inside or outside of this area. Figure 1 shows these four cases. It is obvious that the road links a, c, and d in this figure have no contribution to the divergence rate $D_i(t)$. The road links a and c have neither any outbound flow from this area nor any inbound flow to this area. The road link d has the same amounts of outbound flow and inbound flow to cancel each other. We only need to consider those road links such as b that cross the area border at only one point. Let XR_i denote the set of all the road links that cross the border of S_i, only once.

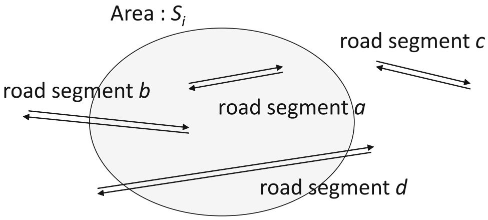


Fig. 1 Four different relationships between a road link and a small geographical area

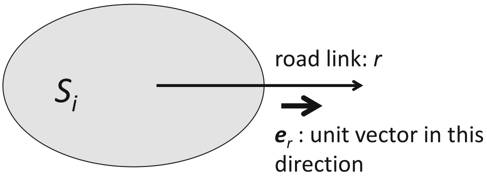


Fig. 2 Outbound directed road link that crosses the border of S_i only once

Figure 2 shows an outbound directed road link that crosses the border of S_i only once. Let the length of this road link r be L_r. Let the total number and the average speed of occupied taxi cars in this road link r at time t be respectively $N_r(t)$ and $V_r(t)$. Approximately we may conclude that cars in this area that are within the distance $v_r\delta$ from the border can go out of this area by time $t + \delta$. Approximately, we may assume that cars are uniformly distributed along this road link with the density of $N_r(t)/L_r$. Therefore, the total number of cars that will go out of this area along this outbound directed road link in the time interval δ can be approximated as $(N_r(t)/L_r)v_r(t)\delta$. Actually we have both inbound and outbound traffic on each road link, i.e., $(N_r^+(t), v_r^+(t))$ for inbound traffic, and $(N_r^-(t), v_r^-(t))$ for outbound traffic. If the road link r is a one way road, either of $N_r^+(t)$ or $N_r^-(t)$ always becomes zero. Therefore the following holds:

$$D_i(t)\delta = \Sigma_{r\in XR_i}(N_r^-(t)\cdot v_r^-(t) - N_r^+(t)\cdot v_r^+(t))\delta/L_r$$

From this equation, the divergence rate of the area S_i at time t can be approximately calculated as follows:

$$D_i(t) = \Sigma_{r\in XR_i}(N_r^-(t)\cdot v_r^-(t) - N_r^+(t)\cdot v_r^+(t))/L_r$$

Figure 3 shows the divergence rate of the occupied taxi traffic flow at 8:00 a.m. in the central area of Sapporo City, using gray scale. This was calculated for each cell in a mesh of 250 m by 250 m cells using the average speed and the number of occupied taxi cars in each road link at every 5 min. Around the center of this map is Sapporo Station. This station has two taxi stands, the north one and the south one. The visualized result shows that, in the morning, people get off taxis mainly at the north stand, and take taxis mainly at the south stand. Sapporo has its main business areas in the south of its main station. People like to take taxis at the south stand to go directly to their business. Since this causes the congestion around the south stand, people taking trains from the station like to get off taxis at the north stand.

Using the same data set, we can also calculate the traffic flow vector field. Let $\boldsymbol{F}_i(t)$ denote the traffic flow vector of a sufficiently small area S_i, and $\boldsymbol{e}_r$ denote a unit vector along a directed road link r which crosses the border of this area at least once. If the directed road link r is curved, we approximate the unit vector $\boldsymbol{e}_r$ to be in the direction from the origin to the destination of this directed road link. This time we need to consider the contribution of those directed road links that cross the border twice since the traffic along this directed road link may partially contribute a traffic

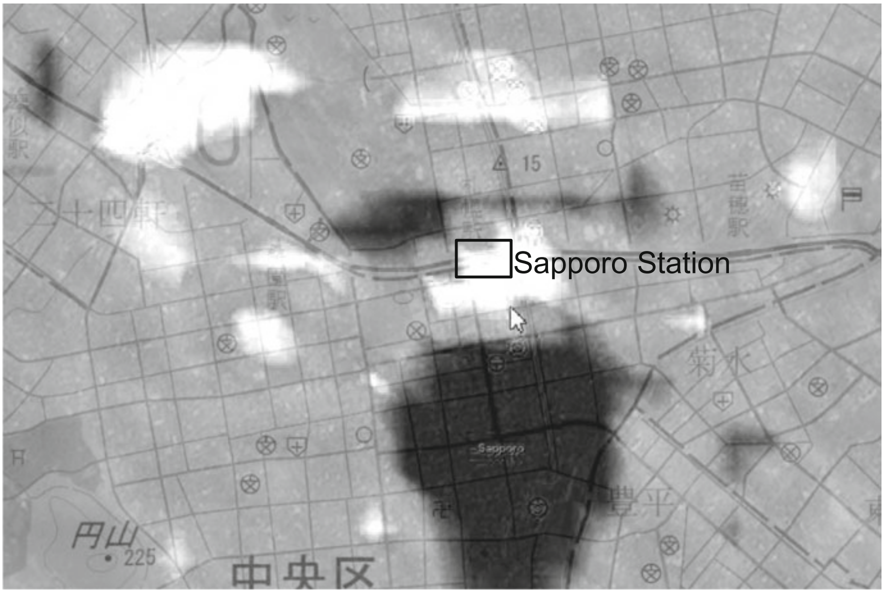


Fig. 3 A divergence rate of the occupied taxi traffic flow at 8:00 a.m.

flow passing through this area. The traffic flow vector $\boldsymbol{F}_{\boldsymbol{i}}\ (t)$ can be approximately calculated as follows:

$$\boldsymbol{F}_{\boldsymbol{i}}(t) = \Sigma_{r \in XR_i^*}(N_r(t)v_r(t)/L_r)\boldsymbol{e}_{\boldsymbol{r}}$$

Here we use XR_i^* instead of XR_i. The set XR_i^* is the set of all the directed road links that crosses the border of S_i. Figure 4a shows the flow vector field of the occupied taxi traffic at 19:00 in the central area of Sapporo City. Sapporo has a nightlife district called "Susukino" around the area 1 km to the south of Sapporo Station. Every evening, you can observe a big inbound flow of occupied taxis toward Susukino. Figure 4b, c show how route preference may change between a day in early December before any serious snowfall and a day in late December after heavy snowfalls. They show that snowfalls in late December cause traffic jams along the biggest southbound stream in early December, and shift the biggest southbound stream to the east.

These analyses show that statistically processed probe car data containing only the number of cars and the average speed in each road link will give us sufficient information about the divergence rate and the vector field of the traffic flow if the time interval is sufficiently small.

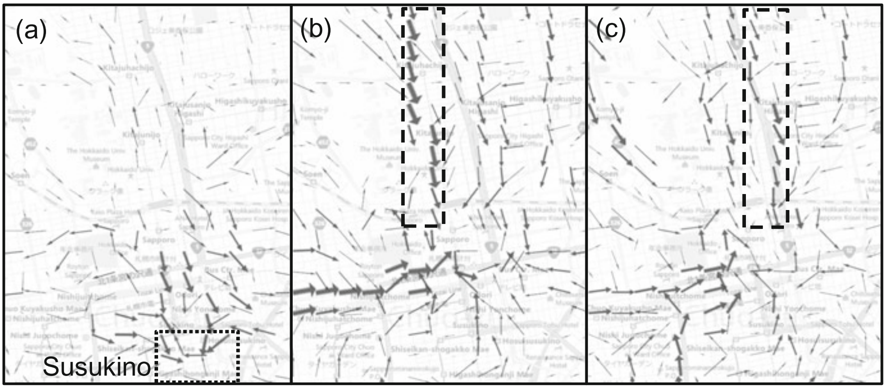


Fig. 4 Flow vector fields of occupied taxi traffic

4 Exploratory Visual Analytics for Pinpoint Snow Removal

Our preparatory study on the big data approach for efficient snow removal in Sapporo showed that macro analyses of probe car data for the whole central city area would not give any meaningful knowledge about the influence of snow fall, snow plowing and removing on traffics [20]. Using the average speed in each road link at every 5 min interval for 24 h, we can characterize each road link as a vector of 288 dimensions. Clustering analyses of road links represented by such vectors told us that even those links along the same route may fall in different clusters on a day with heavy snow fall. In summer time, they almost always fall into the same cluster. This means that the influence of snow on each road link may follow different models even on the same route. Therefore, we may think that influence of snow on traffic in the whole city cannot be modeled by a single monolithic model covering the whole city. It should be considered as a complex system of different models, each of which may be a simple monolithic model. Therefore, macro analyses applied to the whole central city area will not give any meaningful result. In order to obtain meaningful knowledge from such a complex system, we first need to find out a set of subsystems, i.e., subsets of road links in this case, each of which consists of objects, i.e., road links in our case, that can be modeled by the same monolithic model or the same type of models.

One possible solution to this problem may be to exploit an exploratory visual analytics approach. It is well known that exploratory visual analytics [8, 22–24] may use the "coordinated multiple views" visualization framework [16] as its basis. This framework provides more than one view for the visualization of the same database Δ from different aspects. Each view V_i may be a chart view, a map view, a graph representation view, or a calendar view, and shows the evaluation result $Q_i(\Delta)$ of some query Q_i associated with this view using its specific visualization scheme. Each view allows users to select a set of visualized objects by directly specifying each of them or enclosing some of them, which defines a new additional

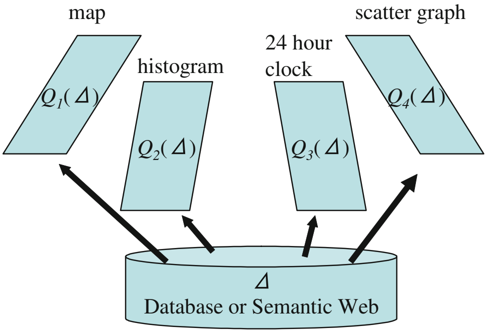


Fig. 5 A coordinated-multiple-views visualization framework

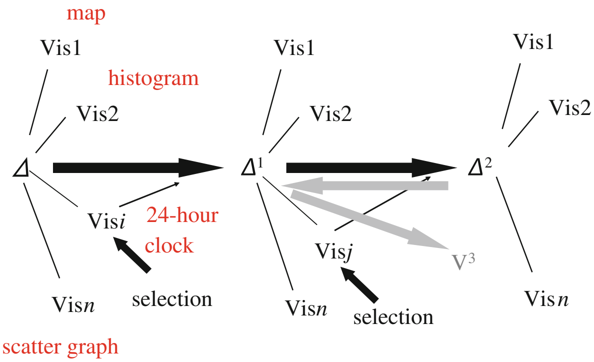


Fig. 6 A process of exploratory visual analytics

quantification condition C to quantify the objects stored in the underlying database Δ. This quantification defines a new database view Δ′ that is obtained by modifying the original view Δ with this additional quantification condition C. Each view V_j including V_i itself then immediately changes its visualization from $Q_j(\Delta)$ to $Q_j(\Delta')$, or just highlights the objects in $Q_j(\Delta')$ in the visualization of $Q_j(\Delta)$, depending on the user specification of its visualization mode. Multiple visualization views are mutually coordinated in this sense (Fig. 5).

In exploratory visual analytics with a coordinated-multiple-views visualization system, each user may start with the original database Δ, and repetitively try different selections of visual objects on different views for the exploration of different quantifications on database objects to find out a meaningful group of database objects. He or she may roll back the preceding visual object selection to try a different quantification through a different visualization view. Figure 6 schematically shows such a process of exploratory visual analytics.

Each view in coordinated-multiple-views visualization is, however, just a database visualization view. No analysis is actually applied in such exploratory visual analytics processes. In order to apply analysis tools to quantified sets of objects, we need to integrate these tools together with their analysis result visualizations into a coordinated multiple view visualization. Many researchers emphasized the importance of integrating various analyses and visualizations [9]. However, to the best of our knowledge, apart from some statistical chart tools to show histograms, correlations, or heatmaps [15], no other analysis tools such as clustering and frequent pattern mining tools have ever been integrated into coordinated multiple views visualizations to allow users to directly select a cluster or some of the mined frequent patterns for further quantifying database objects and for further analyzing those quantified database objects.

Exploratory visual analytics requires an extended coordinated multiple views visualization framework to which we can integrate any analysis tools so that their result visualizations can also be coordinated with other visualization views and analysis result views. It also requires an open library of analysis tools that can be integrated into the framework. These requirements made us to choose our Meme Media technologies [19] as enabling technologies to develop such a framework.

Figure 7 shows a coordinated multiple views visualization system using our Webtop Meme Media technology Webble World [10].

This system, called the Geospatial Digital Dashboard, was developed for exploratory visual analytics of social cyber-physical data related to the winter road management in Sapporo City. It can be set up with many different views. The setup in Fig. 4.2 has several views. The map view shows average taxi speed in each direction of each road link, and the distribution of tweets with geotags. It

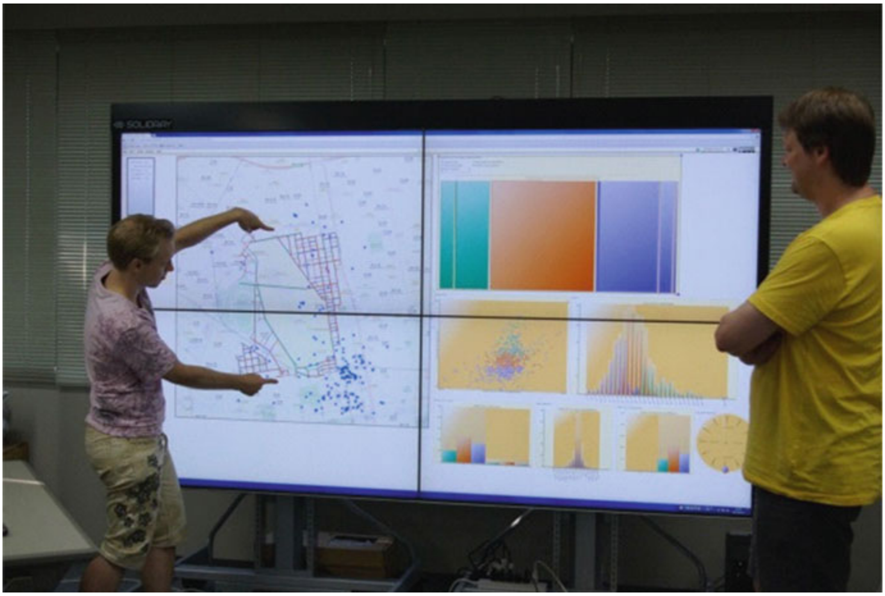


Fig. 7 The Geospatial Digital Dashboard as a coordinated multiple views visualization system

allows us to specify a rectangular area to select only the road links or tweets in this area. The clock view at the bottom right corner shows the population of taxis in each of the 24 h as the area of each circle around the circumference of the clock circle. It allows us to select time intervals for selecting statistical probe car data from only the specified time intervals. Other views include various kinds of charts such as correlation charts and histograms. You may also bring a calendar view with the weather data of each day into this environment. On such a calendar, you may specify only the days with heavy snowfall to focus on only data from heavy snowfall days.

In a coordinated multiple views framework, you may easily connect some analysis tool with its result visualization to any of these multiple views to apply this analysis to the set of objects visualized by this view. However, the visualized analysis result does not normally allow us to select one of the clusters or mined frequent patterns to further quantify those objects in the selected cluster or with the selected pattern. We need to integrate varieties of analysis tools and their result visualizations into our coordinated-multiple-views framework so that these analysis result views may also allow us to directly select some visualized objects, clusters, or mined patterns to further quantify the current database view. We call such an integrated framework a coordinated-multiple-analyses framework.

Let us first consider the integration of statistical analysis tools and their result visualizations into the coordinated-multiple-views framework. Any statistical analysis specifies the group-by attributes and, for each group of records, some aggregate function to calculate the aggregate value. The result can be represented as a relation Stat(GBattributes, Afunction), where the attribute GBattributes is a list of attributes specified as group-by attributes, and the attribute Afunction is a derived attribute whose value is obtained by applying the specified aggregate function such as average, count, minimum, maximum, and correlation to the set of values of the specified attribute in each group. This specified attribute is called the measure attribute. This relation can be visualized in various visualization schemes. Each visualization needs to provide a direct manipulation operation to quantify the GBattributes value and the Afunction value. This quantification further quantifies the values of the database attributes in GBattributes, which modifies the underlying database view from Δ to Δ'. Our taxi probe car data are stored in a relation Taxi(Date, Time, RoadLink, Speed, NumberOfCars, MaxSpeed). For the group-by attributes GBattribute=(Date, RoadLink) and the derived attribute Afunction=average(Speed × NumberOfCars), the attribute AFunction takes the value of the average taxi traffic flow in each day. If we quantify the average taxi traffic flow to be higher than a specified threshold, we will obtain, for each day, those road links satisfying this quantification. This quantification modifies the database view, and changes the other view visualizations and analysis visualizations in the coordinated-multiple-analyses system.

Let us now consider the integration of clustering tools and their result visualizations into the coordinated-multiple-views framework. The result of any clustering applied to objects identified by the values of some attribute A of the underlying database can be considered as a relation Cluster(A, ClusterID), where the values of A work as the object IDs of objects that are clustered, and ClusterID denotes

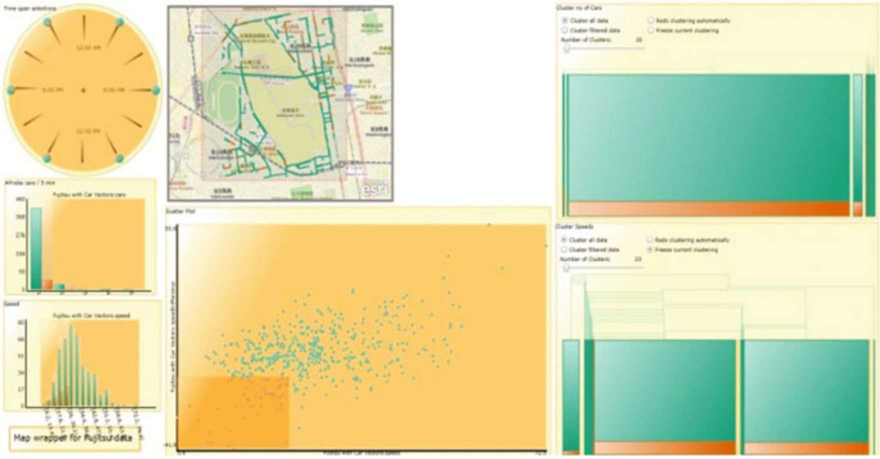
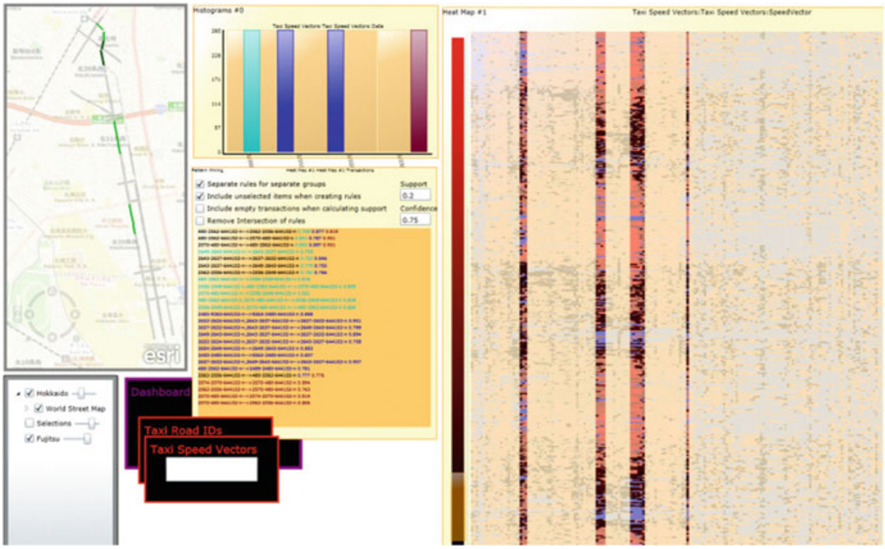


Fig. 8 Extended Geospatial Digital Dashboard with the integration of a clustering tool

the ID of each cluster. This relation Cluster(A, ClusterID) can be visualized in one of various visualization schemes. Each visualization needs to provide a direct manipulation operation for users to select some objects or some clusters. Such a direct selection corresponds to a quantification condition on the attribute A or ClusterID, which further quantifies the underlying database objects. Such a clustering tool with its result visualization can be easily implemented as a Webble, i.e., a visual component of the Webble World system. Figure 8 shows the extended Geospatial Digital Dashboard with the integration of a clustering tool. It has two clustering visualization views in its rightmost area. Each rectangle in each of them represents a cluster. Its size is proportional to the cluster size, i.e., the number of different A attribute values in the cluster. Each of these clustering result views also shows the phylogenetic tree of clusters over these clusters. In this example, road links are clustered in terms of the daily change of the number of taxis and the daily change of the average taxi speed in each road link. These values are available for each 5 min interval. Therefore, the changes of these values characterize each road link as two different vectors of 288 dimensions. For each of these two vector representations of each road link, we clustered the road links based on the similarity of their vector representations. In addition to the initial selection of road links on the map, the lower middle correlation chart showing the correlation between average speed of each road link in summer time and that in winter time is used to further select all the road links in the specified rectangular region. The selected portion of objects is highlighted in each cluster of the two clustering result. You may also choose one of these clusters to highlight only those road links in this cluster on the other visualization views.

Let us now consider the integration of frequent pattern mining tools and their result visualizations into the coordinated-multiple-views framework. Any frequent pattern mining result can be represented by two relations, Mining(Pattern, Supp


Histograms #0
Taxi Speed Vectors:Taxi Speed Vectors:Data
Pattern Mining
Heat Map #1:Heat Map #1:Transactions
Heat Map #1
Taxi Speed Vectors:Taxi Speed Vectors:SpeedVector
esri
Dashboard
Taxi Road IDs
Taxi Speed Vectors

☑ Separate rules for separate groups
☑ Include unselected items when creating rules
☐ Include empty transactions when calculating support
☐ Remove Intersection of rules
Support 0.2
Confidence 0.75

☑ Hokkaido
☑ World Street Map
☐ Selections
☑ Fujitsu

Fig. 9 Extended Geospatial Digital Dashboard with the integration of a frequent pattern mining tool

(, Conf)) and Include(A, Pattern). The first relation lists up each frequent pattern with its support index, and if necessary also with confidence index. The second relation tells which objects among those identified by the attribute value of A include each of the mined frequent patterns. The first relation can be visualized in various visualization schemes to list up mined frequent patterns together with their support and confidence index values. Each visualization needs to provide a direct manipulation operation for users to change the threshold values of support and confidence indices to list up only those patterns with their support and confidence indices higher than the thresholds, and further to directly select some frequent patterns in the list. Using the second relation, this selection of some patterns is converted to the corresponding quantification condition on the attribute A, which further quantifies the underlying database and changes the other coordinated visualization views. Figure 9 shows an extension of Geospatial Digital Dashboard with the integration of a frequent pattern mining tool as well as a clustering tool.

This figure shows a heatmap on the right side, and the result of association rule mining around the center. Each row of the heatmap represents a 5 min interval in a day. All the rows for 4 days are shown here. Each column represents a road link. The highlighted all the columns are those road links selected on the map view. Here we want to apply a micro analysis to the set of road links colored with light gray on the map. These are the road links along the busiest north-south route and its neighboring ones in the north ward in Sapporo. The intensity of each cell in this heatmap represents the average speed in each road link during each 5 min interval.

In general, such a heatmap can be defined for two nominal value attributes A1 and A2, and one numerical value attribute A3 that are arbitrarily selected out of the underlying database attributes. Its row and column represent the domains of A1 and A2 respectively, and each cell shows the numerical value of A3 as the color intensity. In this example, the attribute A1 is the Time attribute representing each 5 min interval, the attribute A2 is the RoadLinkID represented by a geo-location pair of two end points of each road link, and the attribute A3 is the AverageSpeed of taxis in this road link at this time interval.

On the left side of the heatmap, we have a color gradient bar showing how different intensities are mapped to different colors. This bar allows us to specify more than one intensity interval. If you specify k intervals, the heatmap webble generates k items for each of its columns, and will interpret each row as a transaction, and each column as k items. Each item in each transaction has a binary value.

We can apply item set mining and association rule mining to this set of transactions. In Fig. 9, we specified only one intensity interval to focus on the average speed lower than 10 km/h. We wanted to focus on traffic jams in road links. Figure 9 shows the highlighting of those road links with traffic jams in the heatmap, and only the result of applying the association rule mining to the set of transactions, i.e., all the 5 min intervals. This mining tool has two numerical entries for users to specify the threshold values of the support and the confidence. This association rule mining result view allows us to specify some of the mined patterns. This selection quantifies those transactions having one of these selected patterns in the heatmap. Unselected transactions, or rows, will be dimmed in the heatmap. The mining result view Webble also allows us to select each item appearing in each selected pattern. Since each item in this example is a road link, you may pick up a mined association rule to see how the traffic jam in some road links may propagate to other road links on the map view.

Figure 10 shows three analysis results obtained by the same environment as shown in Fig. 9. Here we used an item-set-mining analysis view instead of an association-rule-mining analysis view. The backmost one shows the analysis result for a day before a snowy day. The middle one shows the result for 2 days with very heavy snowfall. The front one shows the analysis result for the day after the snowfall and the immediate snow removal. These show that the traffic jam sections become seriously extended by the snowfall, and that the situation significantly improves after the snow removal.

These traffic jam expansions seem to be caused by areas with traffic jams already before the snowfall, when these areas that already have throughput problems become even worse from snow narrowing the street or making the roads slippery. These traffic jam areas require further analysis to examine if the problems are specific to winter or if there is a more general problem throughout the year. If some of them are specific to winter, they are usually caused by roadside snow piles locally stretching out into the street. Snow removal usually starts with snow plowing to pile snow on the sides of the roads. This is followed by snow removing, when snow is

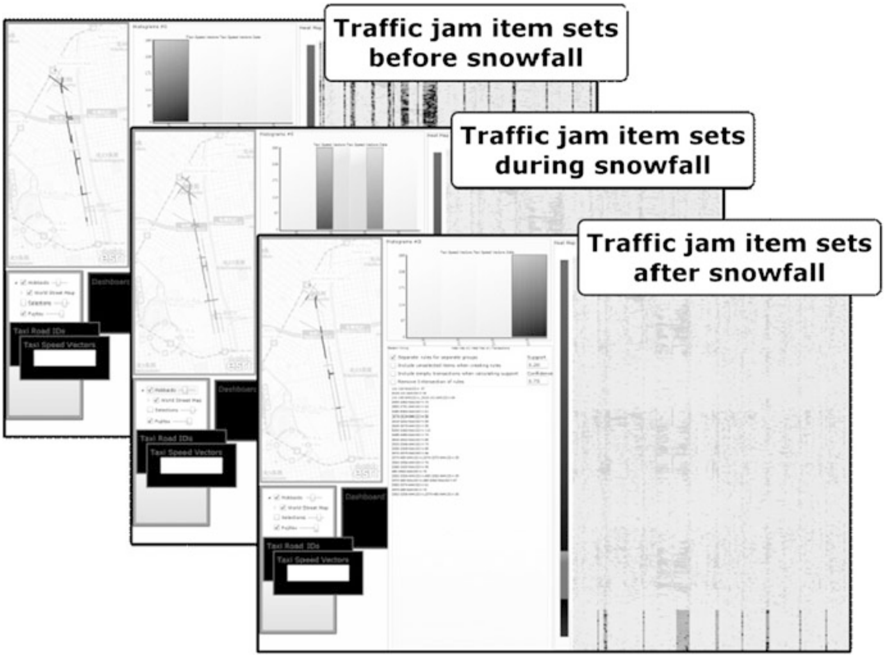

Traffic jam item sets before snowfall
Traffic jam item sets during snowfall
Traffic jam item sets after snowfall
Dashboard
Taxi Road IDs
Taxi Speed Vectors

Fig. 10 Item set mining results on traffic jams on a day before a big snowfall, the 2 days of snowfall, and the day after the snow and the snow removal

cut out of the roadside piles and removed (e.g. by truck) to recover road width. To keep the sidewalks and/or parking spaces clean, people further pile snow on top of the remaining snow piles, which often makes some part fall down into the road. Such snow piles locally stretching out into the streets are the main cause of winter specific local traffic jams. From experience we know that the locations with such local stretch out do not change. Winter-specific traffic jams exacerbated by snowfall are likely to have such local stretch out. Figure 11 shows 3D measurements of road side snow piles and road surface both around some road links with traffic jams that grew after snowfall, and around some road links along the same route but that had no traffic jam. The first one shows lots of local stretch-outs from snow piles, while the second one shows no such stretch out of them. We used a laser range scanner installed on a car to gather the data.

The 3D measurement data was used only as ground truth data to check the validity of our analysis and inference results. This kind of analysis of statistically preprocessed probe car data may give us information about locations of snow stretch outs that cause traffic jams, which may help snow removal operation centers to make decisions on pinpoint snow removal just before the expansion of traffic jams.

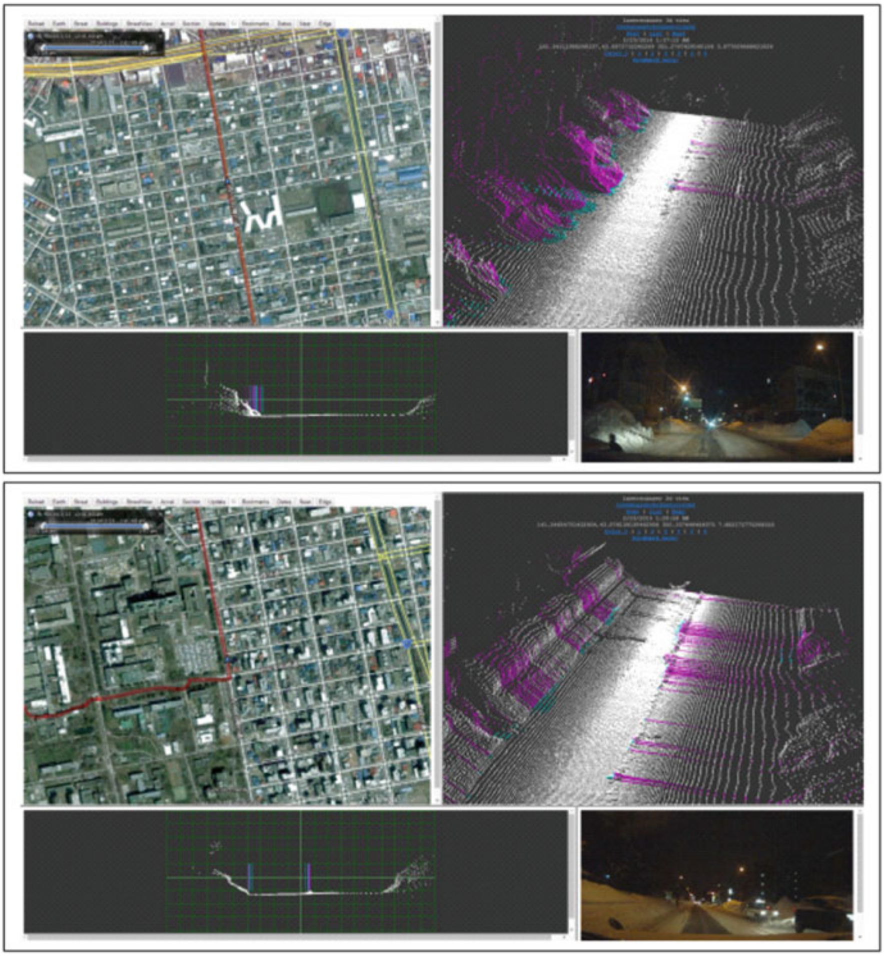


Fig. 11 3D measurement of road surface and road side snow piles both around a traffic jam area that grew during the snowfall, and around a no-traffic-jams area

5 Conclusions

This paper showed that advances of sensor devices and their networking technologies have enabled the real-time monitoring of physical systems ranging from small-size healthcare devices to large-scale plant systems, or even to urban-scale infrastructure service systems. It focused on social CPSs (Cyber-Physical Systems) that denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure services such as traffic, energy, and water services, especially on the winter road management in Sapporo, where we have the world biggest annual snowfall among the cities with more than 1 million populations. Since the targets of Social CPSs are necessarily complex systems of systems, they need to deal with a varier of heterogeneous cyber and physical, real-time and retrospective big data including, for example, probe-car data, weather data, snow removal records, and traffic accident records. For the monitoring and analysis of the traffic and road conditions of the whole city, the use of probe-car data may be the only practical solution today. Their advanced usage without violating personal data protection is fundamental.

This paper has shown that statistically preprocessed probe-car data of road links for an urban-scale area can give us lots of information about traffic flow, such as the temporal change of the traffic flow divergence and the traffic flow vector field. These will tell us the hotspots for taking taxis and getting off taxis, main traffic streams at every time, and dynamic change of route selection preference. Such data also allow us to do more advanced complex analyses. We have picked up the optimization of snow plowing and removing in Sapporo as our target of such analyses.

Since our preceding study showed that the influence of snow and/or snow removal on the traffic and road conditions is not homogeneous across even road links along the same main route in the central city area, we proposed the Geospatial Digital Dashboard system for exploratory visual analytics. This system uses the well-known multiple coordinated views framework for exploratory visual analytics, and extends it to integrate analysis tools with their result visualization views in the same environment so that these may also be coordinated with other visualization views. Users can quantify database views through direct selection of visual objects not only in coordinated views but also using analysis result views. Each quantification is immediately reflected not only in other views but also in all other linked analysis result views. Using the extended Geospatial Digital Dashboard, we have analyzed statistically preprocessed probe car data to detect road links with serious snow stretch-outs narrowing the effective road width. This may allow us to advise the snow removal center to conduct pinpoint snow removal to remove such snow stretch-outs, which may effectively prevent serious expansion of traffic jams, and reduce the total snow removal cost.

From the system architecture point of view, exploratory visual analysis in general requires the following features.

1. It should support the repetition of the hypothesis making through data segmentation of a specific set of data and the hypothesis checking through data analysis and visualization of the segmented data set. The analysis result may be also used for further data segmentation.
2. It should provide a large library of analysis and visualization tools open for the future extension. It should be easy to improvisationally wrap external tools and services into components and to register them into the library for their future reuse in the visual analytics environment through the improvisational federation of them with other tools and services.
3. It should allow users to improvisationally bring external data sources provided as web services into the visual analytics environment.

The first requirement made us propose a coordinated-multiple-analyses visualization framework as an extension of the coordinated-multiple-views visualization environment. We used the webtop meme media technology as the enabling technology for these frameworks.

Exploratory visual analytics requires various analysis tools and data sources. Its system should be open for the future integration of new analysis tools and new data sources to itself. Our framework is based on the webtop meme media system Webble World, which enables us to improvisationally wrap both varieties of tools developed in R, Octave, Python, and Ruby, and any analysis and/or data providing web services into webbles. These webbles can be registered into the open library to increase its variety. Users can improvisationally federate any of these wrapped tools and services to work together. For example, in the Geospatial Digital Dashboard System, the map view can also be used to visualize the geographical distribution of tweets. These tweets may be obtained from the twitter service. We can improvisationally wrap the twitter service into a webble, and improvisationally federate this webble with the map view webble to obtain such a visualization result.

These days we have a huge variety of related open data sources over the Web. They can be accessed through web services. It is important for us to be able to improvisationally federate these data sources with our visual analytics environment. We can also find out a huge variety of open tools and services for data analysis and visualization. The improvisational knowledge federation capability of Webble World will allow us to improvisationally wrap a large portion of them into webbles, and to reuse them in cooperation with other data sources, tools and services in our exploratory visual analytics environment.

References

1. Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th international conference on data engineering, ICDE ’08; 2008. pp. 376–85.
2. Ahlberg C. Spotfire: an information exploration environment. SIGMOD Rec. 1996;25(4):25–9.
3. Aiken A, Chen J, Stonebraker M, Woodruff A. Tioga-2: a direct manipulation database visualization environment. In: Proceedings of the 12th international conference on data engineering, New Orleans, February 26–March 1, 1996. pp. 208–17.
4. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B. Knime - the konstanz information miner: Version 2.0 and beyond. SIGKDD Explor Newsl. 2009;11(1):26–31.
5. Demšar J, Curk T, Erjavec A, Gorup V, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Štajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B. Orange: data mining toolbox in python. J Mach Learn Res. 2013;14(1):2349–53.
6. Hoh B, Gruteser M, Xiong H, Alrabady A. Preserving privacy in gps traces via uncertainty-aware path cloaking. In: Proceedings of the 14th ACM conference on computer and communications security, CCS ’07; 2007. pp. 161–71.
7. ISO. Intelligent transport systems—basic principles for personal data protection in probe vehicle information services. ISO 24100:2010, International Organization for Standardization, Geneva; 2010.

8. Keim D, Mansmann F, Stoffel A, Ziegler H. Visual analytics. In: Liu L, Özsu MA, editors. Encyclopedia of database systems. New York: Springer; 2009. pp. 3341–6.
9. Keim DA, Mansmann F, Thomas J. Visual analytics: how much visualization and how much analytics? SIGKDD Explor Newsl. 2010;11(2):5–8.
10. Kuwahara M, Tanaka Y. Webble world — a web-based knowledge federation framework for programmable and customizable meme media objects. In: IET international conference on frontier computing. theory, technologies and applications; 2010. pp. 372–7.
11. Livny M, Ramakrishnan R, Beyer K, Chen G, Donjerkovic D, Lawande S, Myllymaki J, Wenger K. Devise: integrated querying and visual exploration of large datasets. In: Proceedings of ACM international conference on management of data, SIGMOD '97; 1997. pp. 301–12.
12. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. Yale: rapid prototyping for complex data mining tasks. In: Proceeding of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06. New York: ACM; 2006. pp. 935–40.
13. Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: A generalization-based approach. Trans Data Privacy. 2009;2(1):47–75.
14. North C, Shneiderman B. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In: Proceedings of the working conference on advanced visual interfaces, AVI '00; 2000. pp. 128–35.
15. Perer A, Shneiderman B. Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. IEEE Comput Graph Appl. 2009;29(3):39–51.
16. Roberts JC. State of the art: coordinated & multiple views in exploratory visualization. In: Proceedings of the 5th international conference on coordinated and multiple views in exploratory visualization, CMV '07. Washington: IEEE Computer Society; 2007. pp. 61–71.
17. Sato M, Izumi M, Sunahara H, Uehara K, Murai J. Threat analysis and protection methods of personal information in vehicle probing system. In: Proceedings of the 3rd international conference on wireless and mobile communications; 2007. p. 58.
18. Sugibuchi T, Tanaka Y. Integrated visualization framework for relational databases and web resources. In: Intuitive human interfaces for organizing and accessing intellectual assets. Lecture notes in computer science, vol. 3359. Berlin Heidelberg: Springer; 2005. pp. 159–74.
19. Tanaka Y. Meme media and meme market architectures: knowledge media for editing, distributing, and managing intellectual resources. New York: Wiley; 2003.
20. Tanaka Y, Sjöbergh J, Moiseets P, Kuwahara M, Imura H, Yoshida T. Geospatial visual analytics of traffic and weather data for better winter road management. In: Cervone G, Lin J, Waters N, editors. Data mining for geoinformatics. New York: Springer; 2014. pp. 105–26.
21. Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories. In: Proceedings of the 9th international conference on mobile data management, MDM '08; 2008. pp. 65–72.
22. Thomas J, Kielman J. Challienges for visual analytics. Inf Visual 2009;8(4):309–14. doi:10.1057/ivs.2009.26. http://dx.doi.org/10.1057/ivs.2009.26.
23. Thomas JJ, Cook KA. Illuminating the path: the research and development agenda for visual analytics. National Visualization and Analytics Ctr (2005). http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0769523234.
24. Tukey JW. Exploratory data analysis. Reading: Addison-Wesley; 1977.

Part III
Gas and Odor Sensing

Novel Metal Oxide Gas Sensors for Mobile Devices

Ho Won Jang, You Rim Choi, and Yeon Hoo Kim

Abstract One of the top design priorities for semiconductor chemical sensors is developing simple, low cost, sensitive and reliable sensors to be built in handheld mobile devices. In this chapter, we discuss critical issues for the realization of miniaturized and integrated chemoresistive thin film sensors based on metal oxides and introduce notable recent achievements.

1 Semiconducting Metal Oxide Gas Sensors

Gas sensor is a subclass of chemical sensors. Gas sensor measures the concentration of gas in its vicinity. Gas sensor interacts with a gas to measure its concentration. Its wide variety of applications includes environmental monitoring, process control industries, fuel combustion monitoring for automobiles, boiler control, fire detection, alcohol breath tests, detecting harmful gases in mines and sewers, home safety and national security, gas leak detection for power stations and aerospace shuttles, and grading and condition monitoring of agricultural and marine products. With ever growing needs for a comfortable life, air quality monitoring of indoor and outdoor environments using gas sensors becomes more and more important.

There are various types of gas sensors such as optical, electrochemical, catalytic, surface acoustic wave, capacitive, and semiconductor gas sensors. Feature comparison of several main gas sensors are shown in Table 1. Optical gas sensors possess the largest market of worldwide gas sensors due to their excellent selectivity, stability, and sensitivity to various target gases. However, high cost and large volume limits the wide use of optical gas sensors. Electrochemical gas sensors can selectively detect both reducible gases such as oxygen, nitrogen oxides and chlorine at the cathode and oxidizable gases such as carbon monoxide, nitrogen dioxide, and hydrogen sulfide. They have linear output, low power requirements and good resolution with excellent repeatability and accuracy. While disadvantages of electrochemical gas sensors include narrow temperature range and short life time,

H.W. Jang (✉) • Y.R. Choi • Y.H. Kim
Department of Materials Science and Engineering, Seoul National University,
151-744 Seoul, South Korea
e-mail: hwang@snu.ac.kr

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_7

Table 1 Feature comparison of various gas sensors

Type of gas sensor	Sensitivity	Selectivity	Stability	Size	Cost
Optical	High	Very high	High	Large	Very high
Electrochemical	Moderate	High	Moderate	Large	High
Catalytic	Moderate	High	High	Moderate	High
Semiconductor	High	Low	Moderate	Small	Low

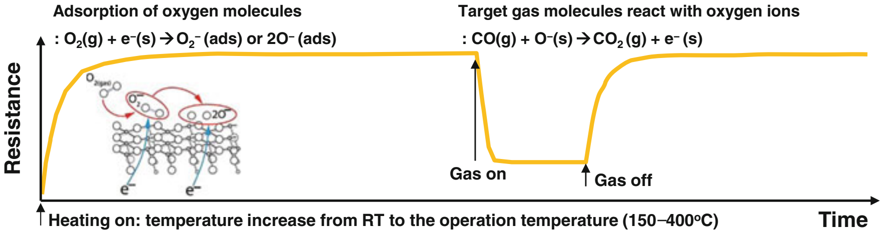


Fig. 1 Operation principle of semiconductor gas sensor

cross sensitivity with other gases, the large volume of electrochemical sensors is the main obstacle for their use in small personal gas detection devices. Catalytic gas sensors show excellent selectivity toward combustible gases. They are robust, simple to operate, and easy to install, calibrate, and use, but catalysts can become poisoned or inactive due to contamination. Their insensitivity toward many other gasses limits their applications to environmental monitoring. Semiconductor gas sensors based on metal oxides such as SnO_2, WO_3, TiO_2, In_2O_3, and ZnO are simple in operation, and easy to fabricate with integrated electronic circuits. Especially the small size and low cost are the most attractive features of semiconductor gas sensors over other types of gas sensors. However, poor selectivity toward a specific target gas and relatively low stability are challenges to be overcome for semiconductor metal oxide gas sensors. Nonetheless, the compatibility with conventional semiconductor manufacturing processes and the easy of making an array for detection of multiple target gases are the unparalleled advantages of semiconductor gas sensors for realizing smart sensor array mimicking the mammalian olfactory system frequently called electronic nose or E-nose, where a specific target gas is recognized by creating and processing a multidimensional pattern of many signals generated by a receptor array [1].

The operating principle of semiconductor gas sensors relies on chemical interactions between gas molecules and the surface of the material, as shown in Fig. 1. At elevated temperatures ranging from 150 °C to 400 °C, the chemical adsorption of oxygen species in air on the surface of the semiconductor material creates surface acceptor states (O_2^-, O^-, O^{2-}) that capture electrons near the surface and lead to the increase of the electron depletion region. When the semiconductor material is n-type, this leads to an increase in the resistance of the material. On exposure to reducing (oxidizing) gases, the quantity of the adsorbed oxygen ions decreases (increases), and thus the resistance of the material decreases (increases). For example, reducing CO molecules in air reacts with oxygen ions at the surface to form CO_2 and release electrons to the material, resulting in the decrease of the resistance. In contrast, oxidizing NO_2 molecules adsorb on the surface and are ionized with capturing electrons from the materials, resulting in the increase of the resistance. Such a change in the resistance upon exposure to a target gas defines the response of a semiconductor gas sensor that changes with the concentration of the target gas.

2 Challenges in Metal Oxide Gas Sensors for Mobile Applications

One of the top design priorities for semiconductor gas sensors is developing simple, low-cost, sensitive and reliable sensors to be built in handheld devices such as mobile phones and tablet computers [2]. Simplicity in operation, low cost, flexibility in production and small size make semiconductor metal oxide gas sensors based on metal oxides as a prime candidate over electrochemical, optical, acoustic and other types of gas sensors. Research efforts on semiconductor gas sensors for mobile applications are mainly focused on enhancing the sensitivity, selectivity, and stability of the sensors and reducing the power consumption of the sensors.

2.1 Improvement of Sensitivity

Generally, gas sensing properties of semiconductor metal oxide gas sensors are affected by three basic factors, namely, utility factor, transducer function, and receptor function [3]. The utility factor refers to the ability of inner oxide grains to access the target gas. The transducer function refers to the ability to convert the signal caused by chemical adsorption of the oxide surface into an electrical signal. Over the past decade, the use of metal oxide nanostructures with large surface-area-to-volume ratios such as nanoparticles, nanowires, nanorods, nanobelts, nanotubes, and hollow spheres has significantly enhanced both the utility factor and transducer function, leading to highly sensitive gas sensors. The receptor function refers to the ability of the oxide surface to interact with the target gas. This function could be largely modified to induce a considerable change in sensitivity and selectivity when additives such as noble metals are loaded on the oxide surface.

Examples of controlling the utility factor, transducer function, and receptor function for enhancing gas sensing properties of semiconductor gas sensors are shown in Fig. 2. Using monolayer close-packed polystyrene microspheres as a sacrificial template, a TiO_2 thin film based on a network of ordered hollow hemispheres was formed by room-temperature sputtering deposition and subsequent calcination at 550 °C. A thin film gas sensor based on the TiO_2 hollow hemispheres exhibited a

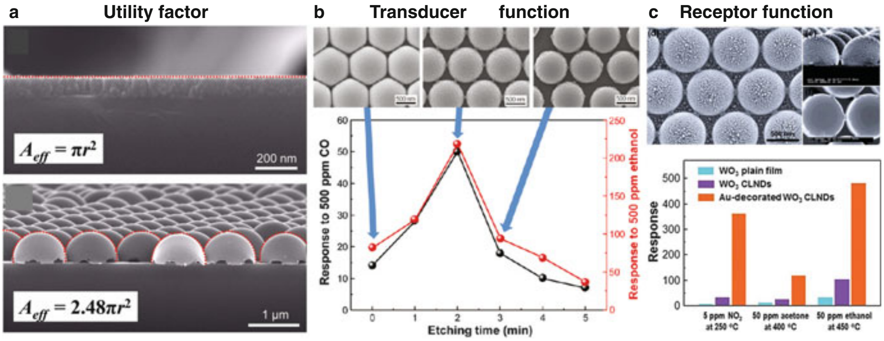


Fig. 2 (**a**) Effective surface area for the adsorption of gas molecules for a flat TiO_2 film and TiO_2 hollow hemispheres (Reproduced from Moon et al. [4]), (**b**) Responses of embossed TiO_2 Responses of embossed TiO_2 films to 500 ppm CO and ethanol as a function of the O_2 plasma etching (Reproduced from Moon et al. [6]), (**c**) Scanning electron microscopy images of Au-decorated WO_3 cross-linked hollow hemispheres and response of Au-decorated WO_3 cross-linked hollow hemispheres to 5 ppm NO_2, 50 ppm CH_3COCH_3 and C_2H_5OH at optimized temperatures (Reproduced from Shim et al. [7])

225 % change in its resistance when exposed to 50 ppm CO at 250 °C, whereas a gas sensor based on a flat TiO_2 film showed an 85 % change. Numerical analysis revealed that the enhancement of the gas sensitivity in the hollow-hemisphere gas sensor is the result of an increase in the effective surface area for the adsorption of gas molecules [4]. Hierarchically porous structures provide good accessibility of gas molecules to the bottom part of the sensing material, leading to high sensitivity and fast responding speed [5]. Room-temperature deposition of TiO_2 on the plasma-treated templates and calcination at 550 °C resulted in embossed films with tailored links between anatase TiO_2 hollow hemispheres [6]. Although all embossed TiO_2 films displayed a similar increase in the surface-to-volume ratio compared with a plain TiO_2 thin film, the response of embossed TiO_2 films with nanolinked hollow hemispheres to CO or ethanol gases was much higher than the response of films with close-linked or isolated hollow hemispheres. The strong correlation between gas sensitivity and the structure of links between the TiO_2 hollow hemispheres revealed the critical importance of tailoring links (namely transducer function) between individual oxide nanostructures for enhancing gas sensing properties of the ensemble of the individual nanostructures. Au-decorated WO_3 cross-linked hollow hemispheres were fabricated using soft templates composed of highly ordered polystyrene beads and self-agglomeration of Au [7]. The distribution and size of Au nanoparticles on the surface of WO_3 cross-linked hollow hemispheres are controlled by varying the thickness of the initial Au film. The responses of Au-decorated WO_3 cross-linked hollow hemispheres to various gases such as NO_2, CH_3COCH_3, C_2H_5OH, NH_3, CO, H_2, and C_6H_6 are at least 5 times higher than those of bare WO_3 cross-linked hollow hemispheres. The response enhancement by Au decoration was dependent on the target gas, which is attributed to an interplay

between electronic and chemical sensitizations. In particular, the Au-decorated WO_3 cross-linked nanodomes exhibited extremely high sensitivity and selectivity, and ppt-level detection limits to NO_2 and C_2H_5OH at 250 °C and 450 °C, respectively.

2.2 Improvement of Selectivity

As aforementioned, semiconductor gas sensors have poor selectivity toward a specific target gas since interfering gas molecules can also react with adsorbed oxygen ions, leading to false readings in gas detection. To overcome this, people make sensor array with multiple gas sensing elements. Like the mammalian olfactory sensing system, a specific target gas is recognized by creating and processing many signals generated by the sensor array. Hwang et al. [8] developed gas sensors with an array of near single crystalline TiO_2 nanohelices fabricated by using an oblique angle deposition method, as shown in Fig. 3. The nanohelix device showed remarkable improvements in gas-sensing performance, including ~10 times higher response at 50 ppm, approximately a factor of 5 lower detection limit, and much faster response time than conventional thin film gas sensors. These tremendous improvements are attributed to the large surface-to-volume ratio, extremely small dimension, and near single crystallinity of the nanohelices, as well as the unique top-and-bottom configuration of electrodes. In addition to the nanohelix gas sensor, they successfully demonstrated the prototype e-nose chip consisting of six gas sensors (TiO_2 thin film, TiO_2 nanohelices, indium-tin oxide slanted nanorods, SnO_2 thin film, SnO_2 nanohelices, and WO_3 nano zigzags) with different nanostructures or sensing materials, monolithically integrated into a sapphire wafer via conventional

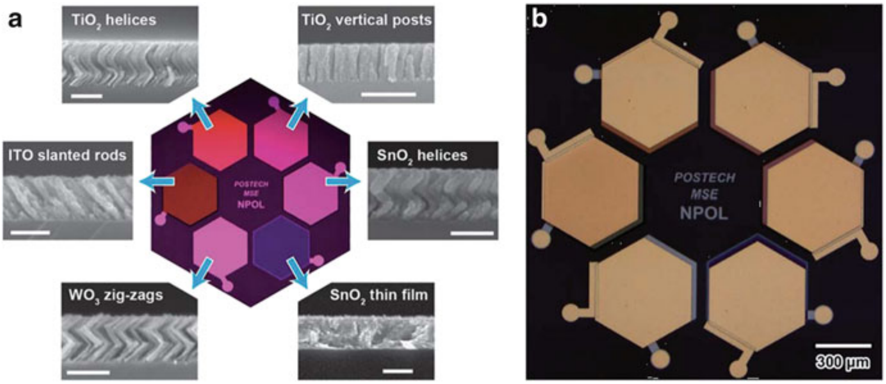


Fig. 3 (**a**) Optical microscopy image of the fabricated prototype e-nose chip after the deposition of six different sensing layers, and corresponding cross-sectional scanning electron microscopy images. (Scale bars represent 300 nm), (**b**) Optical microscopy image of the e-nose after the deposition of top and pad electrodes on the device (Reproduced from Hwang et al. [8])

microelectronics processes, showing an excellent sensitivity and selectivity towards H_2, CO and NO_2 gas species. The fabrication method and the configuration of the proposed gas sensors are promising building blocks not only for monolithically integrated e-nose chips but also for multifunctional smart sensor systems in which optical sensors, light emitters, driving electronic circuits, biosensors, and even solar cells for self-powering are monolithically integrated with e-noses for environmental monitoring.

If each element of sensor array has good selectivity to a specific target gas, the recognition of the target gas via signal processing would be much easier and more accurate. Recently Kim et al. [9] reported ultraselective and sensitive detection of xylene and toluene for monitoring indoor air pollution using Cr-doped NiO hierarchical nanostructures, while ultraselective detection of methyl benzenes with the ability to differentiate benzene, formaldehyde and ethanol has never been reported before. Pure and Cr-doped NiO hierarchical nanostructures were prepared, characterized, and utilized for the fabrication of ultraselective, highly sensitive, and reliable o-xylene and toluene gas sensors, as illustrated in Fig. 4. The gas response and sensor resistance in air of the sensors were both increased significantly by doping NiO flower-like hierarchical nanostructures with 1.15–2.56 at% Cr. Controlling the hole concentration via incorporation of Cr^{3+} into the NiO lattice results in Cr-doping-induced electronic sensitization of the NiO sensor and enhancement of the gas response. Moreover, the sensor consisting of Cr-doped NiO nanostructures showed superior selectivity to o-xylene and toluene, with negligible cross-responses to benzene, formaldehyde, ethanol, and other interferencing gases, while the sensor made of pure NiO nanostructures did not exhibit notable selectivity to a specific gas. The selective detection of methyl benzenes such as o-xylene and toluene was attributed to catalytic promotion of oxidation of methyl groups by the Cr component. Accordingly, the dual positive roles of Cr doping into NiO hierarchical nanostructures provide a novel method to design ultraselective, sensitive, and reliable xylene and toluene sensors that can discern benzene, formaldehyde, ethanol, and other indoor volatile organic compounds.

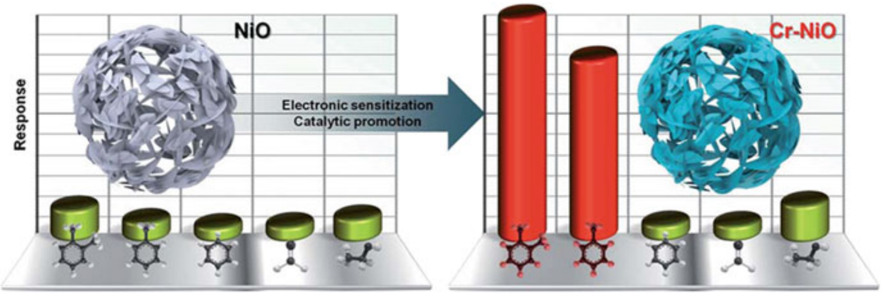


Fig. 4 Design of ultraselective and highly sensitive methyl benzene sensors for indoor air monitoring using Cr-doped NiO hierarchical nanostructures (Reproduced from Kim et al. [9])

2.3 Improvement of Stability

Besides poor selectivity, there are a number of other obstacles that need to be overcome in order to further expand the area of application of semiconductor gas sensors. The influence of humidity on the sensor response, the response/recovery speed, and the sensor resistance in an air atmosphere is one of most serious problems. This is because water vapor is present in most applications and strongly interacts with the oxide semiconductor surface, which leads to a significant deterioration of the sensor performance. Many strategies have been used to improve the selectivity and response/recovery times, such as using catalytic additives, increasing the operation temperature, using porous nanostructures, and extending the range of the measurement parameters.

Kim et al. [10] reported a new strategy to produce humidity-independent chemical sensors by combining readily gas-accessible hierarchical sensing layers and catalytic surface additives. The humidity dependence of the gas-sensing characteristics in SnO_2-based sensors was reduced to a negligible level by NiO doping. In a dry atmosphere, undoped hierarchical SnO_2 nanostructures prepared by the self-assembly of crystalline nanosheets show a high CO response and a rapid response speed. However, the gas response, response/recovery speeds, and resistance in air were deteriorated or changed significantly in a humid atmosphere. When hierarchical SnO_2 nanostructures were doped with 0.64–1.27 wt% NiO, all of the gas sensing characteristics remained similar, even after changing the atmosphere from a dry to wet one. According to diffuse-reflectance Fourier transform IR measurements, it was found that the most of the water-driven species are predominantly absorbed not by the SnO_2 but by the NiO, and thus the electrochemical interaction between the humidity and the SnO_2 sensor surface was totally blocked. NiO-doped hierarchical SnO_2 sensors exhibit an exceptionally fast response speed (1.6 s), a fast recovery speed (2.8 s) and a superior gas response [$R_a/R_g = 2.8$ at 50 ppm CO (R_a: resistance in air, R_g: resistance in gas)] even in a 25 % r.h. atmosphere. The doping of hierarchical SnO_2 nanostructures with NiO is a promising approach to reduce the dependence of the gas-sensing characteristics on humidity without sacrificing the high gas response, the ultrafast response and the ultrafast recovery.

2.4 Reducing Power Consumption

To insure rapid and reversible operation of semiconductor gas sensors, which rely on adsorption and desorption of molecules on the surface of semiconducting materials, the operation temperature is typically maintained between 150 °C and 400 °C. This requires the integration of heating elements within sensor devices with power consumption as high as hundreds of mW. Recently many manufacturers adopt micromachined silicon platforms to reduce the power consumption of semiconductor metal oxide gas sensors. Elmi et al. [11] reported the development of state-of-the-art metal oxide semiconductor gas sensors based on ultra-low-power consumption micro-machined hotplates and targeted to volatile organic compounds detection at ppb-level. The design, the process flow as well as a thorough characterization of the fabricated ultra-low-power hotplates was demonstrated. With a 78 μm diameter circular hotplate, only 8.9 mW were necessary to reach the typical 400 °C operating temperature. Transient temperature response evaluation has shown a very low thermal time constant of only 1.5 ms. On those ultra-low-power hotplates, complete ultra-low-power metal oxide sensors with high sensitivity to ppb-level volatile organic compound concentrations have been fabricated. The sensing layer deposition process as well as a functional characterization was reported. Responses of 300 % to 5 ppb of C_6H_6 in synthetic air with a relative humidity of 30 % have been obtained operating the sensor at a constant temperature of 415 °C, with a response time of 50 s. The ability to enhance selectivity towards various compounds by working under different operating conditions has been preliminarily demonstrated. Although micromachined silicon platforms reduce the power consumption of semiconductor gas sensors down to tens of mW or several mW, but the power consumption is still remains a burden for portable devices operating with batteries.

3 Ultralow Power Consumption Metal Oxide Gas Sensors

In addition to the use of micromachined hot plates, other approaches have been tested to enhance surface chemistry and to reduce the operational temperature and power consumption of semiconductor metal oxide sensors, such as surface decoration of the sensing layer with noble metal nanoparticles, the use of UV activation of reactants, and the use of electrostatic field. More interestingly, self-heated metal oxide nanostructures have been used for ultra-low-power gas sensors.

3.1 Self-Heated Nanowire Gas Sensors

Strelcov et al. [12] demonstrated the effect of Joule self-heating of the semiconductor metal oxide nanowire gas sensor on its surface reactivity and kinetics. The core idea of the experiment was to operate the nanowire conductometric sensor under such biasing conditions that the Joule heat released in the chemiresistor is sufficient to warm up the nanowire to the temperature required for surface redox reactions to occur. Thus the yield and the rate of the oxygen ionosorption and hydrogen oxidation reaction running on the surface of the SnO_2 nanowire would be proportional to the Joule power released in the nanostructure. Due to small thermal capacitance and hampered thermal losses from the nanowire to its surroundings, the sensor was able to operate without a heater, consuming only a few microwatts of power. The results demonstrated the importance of the self-heating effect in nanowire electronics and its potential use in chemical and bio-sensing, where the ultra-small size of the active element and minimal power consumption are crucial.

Independently, Padres et al. [13] demonstrated that self-heating in individual SnO_2 nanowires could be used to fabricate ultralow power consumption gas sensors. The proof-of-concept devices exhibited responses to different concentrations of NO_2, nearly identical to those obtained with an external microheater. Therefore, energy-efficient metal oxide sensors suitable for mobile devices could be obtained using this intrinsic effect. Moreover, the combination of experiments with both technologies (self- or external heating) allowed determining the effective temperature achieved during the measurements, demonstrating the feasibility of taking advantage of self-heating in nanowires to develop ultralow power consumption integrated devices.

3.2 Self-Heated Thin Film Nanostructure Gas Sensors

Ultralow power consumption of gas sensors based on self-heated individual metal oxide nanowires or nanobelts have been reported, but these typically are insufficiently sensitive and/or are difficult to integrate with low cost and high yield mass production processes.

Alternatively, Moon et al. demonstrated structurally simple but extremely efficient all oxide chemoresistive sensors using highly effective self-activation in anisotropically self-assembled nanocolumnar tungsten oxide thin films on glass substrate with indium-tin oxide electrodes [2]. The fabrication process of transparent chemical sensors in which commercially available indium-tin oxide (ITO)-coated glass was used as the substrate is shown in Fig. 5. ITO interdigitated electrodes

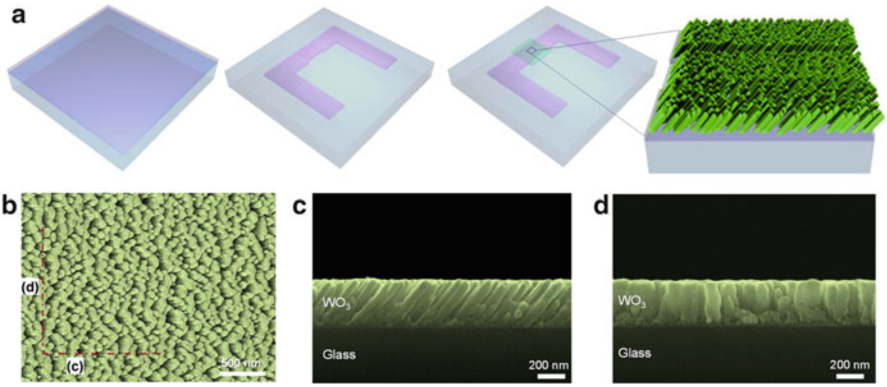


Fig. 5 (**a**) Schematic showing the fabrication process of transparent sensors based on nanocolumnar oxide films, (**b**) Plain-view SEM image of the nanocolumnar WO_3 film between ITO IDEs, (**c**, **d**) Cross-sectional SEM images of the nanocolumnar WO_3 film cut along and across the ITO IDEs as marked in (**b**), (**e**) 40°-tilted SEM image of nanocolumnar WO_3 film between and on ITO IDEs. Parts highlighted in *reddish color* indicate localized current pathways which meander with narrow necks (Reproduced from Moon et al. [2])

(IDEs) with interspacing of 5 mm were patterned by dry etching. Porous WO_3 thin films were deposited onto the IDEs by glancing angle deposition (GAD) via RF sputtering. GAD utilizes the self-shadowing effect of initial nuclei to grow inclined nanowires, nanorods or nanocolumns. The nature of the oxide nanostructures grown by GAD depends largely on the material being sputtered. For example, porous nanostructures were obtained for WO_3 and Nb_2O_5, whereas In_2O_3, TiO_2, SnO_2 and ZnO formed relatively dense films under the same deposition condition. The size and density of the initial nuclei, which depend on the metal element, appeared to be key parameters in determining the final nanostructure. Oxides composed of higher melting temperature metal elements resulted in more highly porous structures. Films with microstructural elements exhibiting high length to diameter aspect ratios are described as nanocolumnar. An example is the 330-nm-thick porous WO_3 thin film, with column diameters of 30–80 nm and aspect ratios, ranging from 5 to 8.

The nanocolumnar WO_3 sensing film on ITO IDEs looks smooth macroscopically. The total transmittance of the sensor over 450–900 nm was determined to be 90.2 %, nearly identical to the transmittance of the ITO/glass substrate by itself. Owing to this high transmittance, the fabricated sensors were barely visible, suggesting that these sensors could be embedded in next-generation smart displays equipped with transparent electronics for automobile, biomedical, military, aerospace, and consumer applications. With a maximum difference of 14 % at 440 nm, the 1.5 % lower specular transmittance, on average, compared to the total transmittance means that light scattering likely took place due to the presence of pores and local variations in the lengths of WO_3 nanocolumns.

Upon closer examination, it was apparent that the nanocolumnar WO_3 films exhibit structural anisotropy. The cross-sectional scanning electron microscopy (SEM) image of the films between the IDEs, cut parallel with the direction of the IDEs, showed arrays of canted nanocolumns separated from each other by elongated pores which extend over nearly the full thickness of the film. In contrast, the nanocolumns were connected each other and form walls with considerably higher density across the IDEs. The pores facilitate access of gaseous molecules to the bottom of the columnar film, while the dense connected nanocolumns across the IDEs provide efficient pathways for electrical current flow. The separation of the individual nanocolumns grown between the IDEs was seen more clearly by the cross-sectional transmission electron microscopy images.

Self-activation of dense-planar and nanocolumnar WO_3 thin film sensors was monitored by measuring resistance change with increasing applied voltage. Both devices displayed linear current-voltage characteristics, indicating ohmic behavior between the WO_3 films and ITO electrodes. When the applied voltage was increased from 0.1 V to 5 V, there was a decrease by a factor of 1.45 in sensor resistance, with time transient of $\sim$150 s, for the conventionally prepared thin film sensor. Thermographic images in Fig. 6 show that the increase in film temperature by self-activation was less than 2 °C. For the same condition, the resistance of the nanocolumnar film decreased by two orders of magnitude with transient time of $\sim$120 s. The film temperature rose to 139 °C due to self-activation. The reduced resistance of the WO_3 film at higher temperatures reflects the semiconducting nature

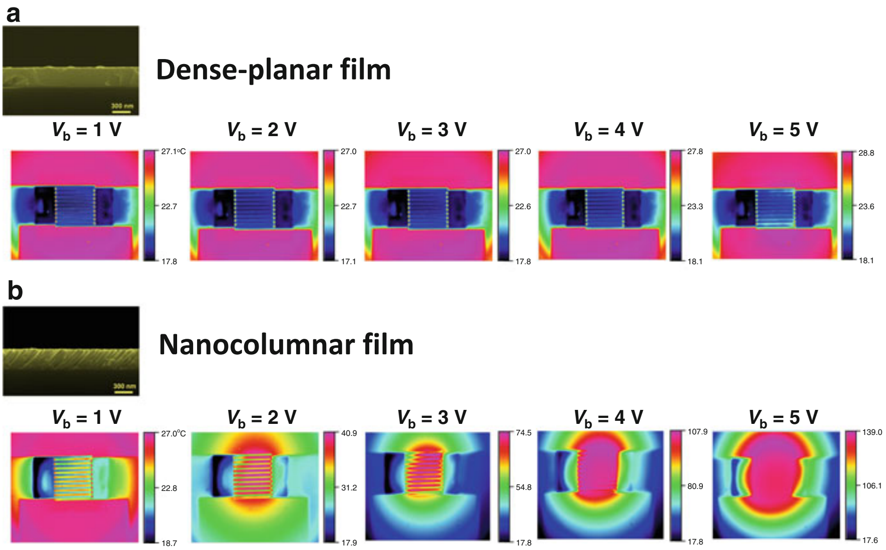


Fig. 6 (**a**, **b**) Thermographic images of dense-planar and nanocolumnar WO_3 gas sensors as a function of the input bias voltage (Reproduced from Moon et al. [2])

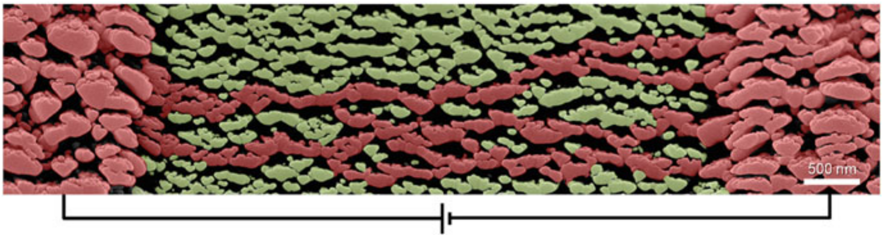


Fig. 7 40°-tilted SEM image of nanocolumnar WO3 film between and on ITO IDEs. Parts highlighted in *reddish color* indicate localized current pathways which meander with narrow necks (Reproduced from Moon et al. [2])

of WO_3. The nanocolumnar thin film device exhibited approximately two orders of magnitude higher resistance than the conventional film at the applied bias voltage of 0.1 V, but due to the pronounced self-activation of the nanocolumnar film, both devices exhibited similar values of resistance under 5 V bias.

Pronounced self-activation in the nanocolumnar film originates from the unusual geometry of the film. Figure 7 shows a 40°-tilted SEM image of the nanocolumnar WO_3 film deposited between and on the ITO IDEs. The porosity of the film is estimated to be 38 % on the basis of the analysis of the black and white contrast. A closer look reveals that the nanocolumns are disconnected from each other at many points and that some current pathways in the film are established only along localized regions. Even percolating pathways often meander through narrow

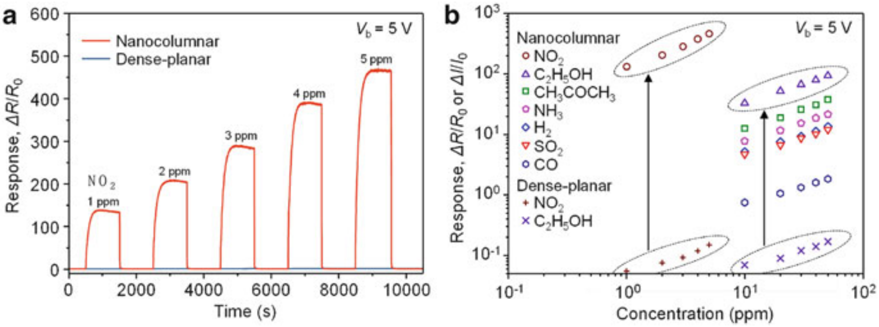


Fig. 8 (**a**) Sensing transients of the dense-planar and nanocolumnar WO_3 thin film sensors to 1–5 ppm NO_2 at an applied bias voltage of 5 V, (**b**) Response of the dense-planar and nanocolumnar WO_3 thin film sensors as a function of gas concentration for various target gases (Reproduced from Moon et al. [2])

necks (20–40 nm in width). For this geometry, electron flow is constricted, leading to increased joule heating. Furthermore, ambient air in the pores, providing an excellent thermal barrier, suppresses heat dissipation laterally to the IDEs, while the small contact area between nanocolumns and glass substrate minimizes heat dissipation to the substrate. This is in contrast to large heat losses in conventional sensors, connected with the incorporation of independent heaters screen-printed or sputtered beneath the substrate or between the sensing film and the substrate. These findings clearly demonstrate that self-assembled nanocolumnar WO_3 films can serve as very efficient self-activated microheaters with minimal heat loss and power consumption. In addition to thermographic imaging, joule heating in the nanocolumnar films was confirmed by application of nanoscale electrical contact resistance measurements.

Dynamic sensing transients of the dense-planar and nanocolumnar WO_3 thin film sensors to NO_2 without external heating are shown Fig. 8. In the case of the dense-planar sensor, the response to NO_2 (defined here as $\triangle R/R_0$, where R_0 and $\triangle R$ denote the initial resistance of the sensor in air and the resistance change of the sensor by exposure to the test gas, respectively) was as low as 0.15 at 5 V and the sensor does not show full recovery to the original resistance. In contrast, the nanocolumnar film sensor shows a dramatic enhancement in response with compared to the dense-planar sensor. At 5 V, the response was extremely high, 450; the highest value ever reported, as far as the authors are aware, for metal oxide thin film sensors. The resistance fully recovers to the initial value within several hundred seconds. The response time (t_{90}) that is the time for the sensor's response resistance to reach 90 % of its steady state value for the nanocolumnar film sensor was about 190 s, which was much faster than that of the dense-planar sensor ($t_{90} > 500$ s). The linear and ultra-high response of the nanocolumnar sensor to 125 ppm of NO_2 promises reliable detection of this environmentally important gas. In comparison, the response of the dense-planar sensor is negligibly small and sluggish, largely, attributable to the low sensing temperature and its reduced active surface area.

To verify the superior response of the nanocolumnar sensor relative to the dense-planar sensor more generally, the response to various gases such as NO_2, C_2H_5OH (ethanol), CH_3COCH_3 (acetone), NH_3, H_2, SO_2 and CO was studied. For the dense-planar sensor, the responses of the sensor to acetone, NH_3, H_2, SO_2, and CO ($\triangle I/I_0$, where I_0 and $\triangle I$ denote the initial current of the sensor in air at a fixed bias and the current change by exposure to the test gas, respectively) were close to zero while the responses to 5 ppm NO_2 and 50 ppm ethanol were lower than 0.2. In stark contrast, ultrahigh responses to all the gases were achieved by the nanocolumnar sensor. For NO_2 and ethanol, the response was three orders of magnitude higher than for that of the dense-planar sensor. This exceptional response of the nanocolumnar sensor is attributed to the combined effects of self-heating, the porous nanostructure with high surface-to-volume ratio ($\sim$33 times higher specific area than the dense-planar film) and the presence of narrow necks between the columns. Although the NO_2 concentration of 1 ppm was the lowest examined experimentally in the present study, the theoretical detection limit (signal-to-noise ratio > 5) was calculated to be approximately 5 parts per trillion (ppt). This value is much lower than the ambient air quality standard (AAQS) levels of the European Union, United States and Korea, which are at several ppb levels. For SO_2 and CO, the detection limits of the nanocolumnar sensor are also substantially lower than the AAQS levels, suggesting strong potential of this technology serving as the basis of highly responsive air quality sensors. Furthermore, detection limits of sub-ppb levels to ethanol and acetone demonstrate the potential of the sensor for use in high performance volatile organic compound sensors. Owing to the low working temperature ($\sim$140 °C), the nanocolumnar sensor shows very stable operation. From a comparison involving the same configuration of gas chemical sensors under external heating, the working temperature of the present sensor was estimated to be lower than 150 °C. Meanwhile, the temperature of the backside of the sensor was measured to be 43 °C during operation, at an applied bias voltage of 5 V, confirming that the present sensor design undergoes minimal heat dissipation through the substrate. With increasing the applied bias voltage, the authors could obtain higher responses to target gases. However, when the applied bias voltage is higher than 7 V, the working temperature of the sensor became higher than 200 °C. As we discussed earlier, such a high working temperature should come at a cost to long-term reliability probably along with morphological changes in the sensing film. This suggests that the applied bias voltage bias should be adjusted so that the working temperature might not exceed 200 °C.

In addition to high sensitivity and long-term reliability, low power consumption is required for practical applications of the nanocolumnar sensors as a component in handheld devices such as mobile phones. To address this issue, the authors have measured power consumption of the sensors with changing the area of sensing film and the duty cycle of pulsed bias voltage. Figure 9 shows the power consumption of nanocolumnar sensors as a function of sensing area. By reducing the area of the WO_3 sensing film from 1 mm × 1 mm to 100 μm × 170 μm, the response of the sensors remains nearly constant, while the power consumption decreases from 251 mW to 21.6 mW. These values are much lower than the power consumption

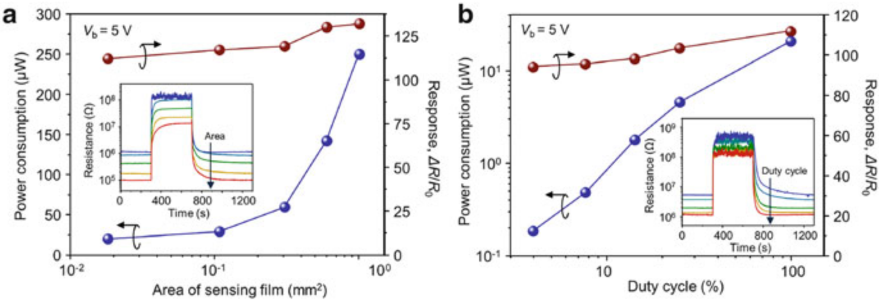


Fig. 9 (**a**) Power consumption and response of nanocolumnar WO_3 thin film sensors to 1 ppm NO_2 as a function of the area of sensing film. *Inset*: response transients to 1 ppm NO_2 of sensors with different areas of sensing film, (**b**) Power consumption and response of the nanocolumnar WO_3 thin film sensor with 100 μm × 170 μm sensing area to 1 ppm NO_2 as a function of the duty cycle of pulsed bias voltage. For each duty cycle, the pulse period is 5 ms. *Inset*: response transients to 1 ppm NO_2 of the sensor with changing the duty cycle (Reproduced from Moon et al. [2])

of even micromachined thin film sensors (5–200 mW) and comparable to those of self-heated single nanowire sensors (tens of mW). When the response of the sensor with 100 μm × 170 μm sensing area is normalized with respect to the power consumption, it exhibits incomparably superior performance to the state-of-art chemoresistive sensors. The power consumption of the sensor could be further lowered using pulsed mode operation, as shown in Fig. 13. When the duty cycle of the pulsed bias voltage is reduced down to 4 %, the sensor still exhibits excellent sensing performance with moderate decreases in response and response speed. The strikingly low power consumption, 0.18 μW, of the sensor at a 4 % duty cycle means that the sensors can operate for a half year using a commercially available lithium polymer cell phone battery (output voltage: 3.7 V DC, capacity 1,500 mAh). These results demonstrate the overall superior performance of the nanocolumnar WO_3 thin film sensors and their remarkably low power requirements point to the feasibility of embedding them into miniature portable devices.

The remarkable device performance, achieved with a facile fabrication process, considerably broadens the potential application of chemoresistive sensors to transparent electronics and highly miniaturized mobile devices. We believe that very high performance portable electronic noses with selective chemical sensing can be developed by integrating multiple sensors of our design onto a single sensor platform coupled with the adoption of surface decoration and temperature variation techniques. The approach, successfully demonstrated in this work for miniaturized chemoresistive gas sensors, can be extended, we believe, to improve the performance of other solid-state devices such as fuel cells, CO_2 reduction systems and other types of sensors that operate at elevated temperatures by removing separate heating elements and exploiting the self-activation capacity of active elements within the device structure.

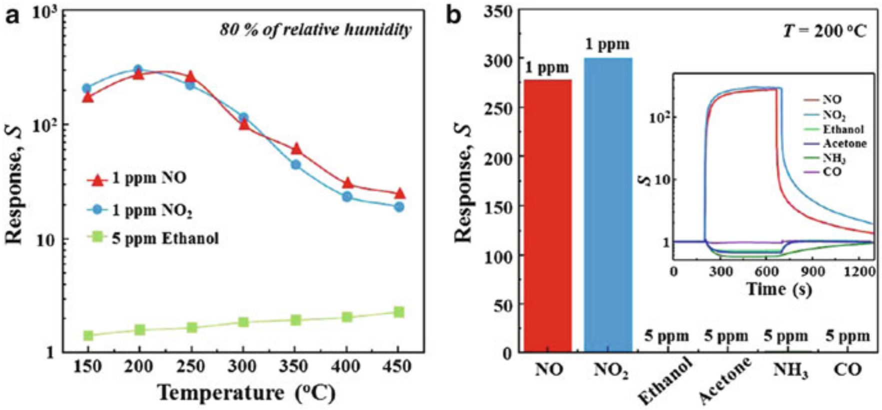


Fig. 10 (**a**) Responses of the VLNF sensor to 1 ppm NO, NO_2 and 5 ppm ethanol in 80 % of relative humidity atmosphere as a function of temperature, (**b**) Response and response curves (*inset*) of the VLNF sensor to 1 ppm NO_x, 5 ppm ethanol, acetone, NH_3, and CO at 200 °C in 80 % of RH atmosphere (Reproduced from Moon et al. [14])

Recently we found that the nanocolumnar WO_3 sensors are promising for extremely selective NO detection. Figure 10 shows responses of a chemoresistive sensor based on the 380-nm-thick WO_3 film with villi-like nanofingers (VLNF sensor) to 1 ppm NO, NO_2 and 5 ppm ethanol as a function of operating temperature in 80 % of RH atmosphere. The response, S, is defined as $R_{gas}/R_{ambient}$ for the oxidizing gas NO_x (NO, NO_2) and $R_{ambient}/R_{gas}$ for the reducing gases (ethanol) where R_{gas} and $R_{ambient}$ denote the sensors' resistances in the presence and absence of a test gas, respectively. Highly sensitive and selective NO and NO_2 sensing properties of the VLNF sensor is clearly observed. The responses to 1 ppm NO and NO_2 reach 278 and 300 at 200 °C, respectively. Under a dry atmosphere, the response of VLNF sensor to 1 ppm NO is 450, which is dramatically higher and considered to be the highest value compared with the response values of previous reported sensors based on WO_3. When the temperature increases over 250 °C, the response to NO_x decreases, while the responses to ethanol slightly increase. This indicates that the optimum temperature for the detection of NO_x is 200 °C, which is consistent with previous studies on chemoresistive NO and NO_2 sensors. The fact that the responses change according to the operating temperature show the amount of chemisorbed oxygen play an important role in the mechanism of NO and NO_2 detection. Generally, the operating temperature affects the kinetics of the adsorption on the active surface and leads to the change of gas response. At less than 200 °C, oxygen adsorption and surface reaction are generated by sufficient thermal energy, which is effective to overcome the activation energy barrier. Thus, the response of VLNF sensor increases to 200 °C. However, when the operating temperature increases (>200 °C), the desorption process on the active surface is dominant. Consequently, the responses tend to decrease as the increases of operating temperature due to the thinner depletion layer at high temperature.

For 5 ppm ethanol, acetone, NH_3, and CO, the VLNF sensor showed responses less than 2 at 200 °C. Thus the response ratios, S_{NOx}/S_{gases} are higher than 150. Because NO molecules have the better activity than oxygen for adsorption on the oxide surface, NO is not separated easily from the surface in the off state of the gas. In this reason, the recovery to the original baseline resistance by the adsorption of oxygen molecules on the surface is relatively lower than the response. In general, the operating temperature is a critical factor to improve the gas sensing performance of a metal oxide chemoresistive sensor because the amount of ionized oxygen species (O^-, $O_2{}^-$, O^{2-}) on the surface of the metal oxide changes with the operating temperature, leading to changes in both sensor resistance and response. For relatively low temperatures, there are much less ionized oxygen species on the surface of the metal oxide. Therefore, reducing gases such as ethanol, acetone, NH_3, and CO have a poor sensitivity at high temperature. On the other hand, oxidizing gases such as NO_x have a high sensitivity at relatively low temperatures ranging from 150–250 °C due to the electron-trapped force of the NO_x molecules.

An extremely selective NO_x sensing mechanism at low temperature can be explained as the following. The molecular NO_x has an unpaired electron and is known as a strong oxidizer than other gases. Upon NO_x adsorption, electron transport is likely to occur from nanostructure WO_3 to NO_x because of the electron-trapped force of the NO_x molecules at low temperature. Consequently, the vividly high selectivity of the VLNF sensor to NO_x with negligibly low cross-response to ethanol, acetone, NH_3, and CO, which are well-known reactive gases that might be included in human breath with concentrations ranging from several ppb to several ppm demonstrates a strong potential for detecting NO in human breath. A closer look reveals that the maximum responses to ethanol at 450 °C are much lower than the response to NO_x at 200 °C, reflecting that WO_3 itself has a high selectivity to NO_x relative to other metal oxide semiconductors including SnO_2, the most common material for chemoresistive sensors. A typical response curve of the VLNF sensor to 0.2–1 ppm NO at 200 °C in 80 % of RH is shown in Fig. 11. For comparison, the response of a dense plain sensor fabricated by rf-sputtering is also plotted. Upon exposure to oxidizing NO, the VLNF sensor quickly responds with increase in the resistance, which indicates that the WO_3 film is an n-type semiconductor. Compared with the reference sensor based on a dense plain WO_3 film (plain sensor), the VLNF sensor exhibits about 200 times higher responses to 1 ppm NO. Even at an extremely low concentration of 200 ppb, the VLNF sensor shows clear response, which is the first experimental demonstration on detecting ppb-level NO in highly RH atmosphere using a chemoresistive sensor in our best knowledge. In addition, for dynamic sensing transients and response to five consecutive pulses at NO concentration ranging from 0.2 to 1 ppm and 5 ppm, the resistances are completely recovered after reacting test gas. Therefore, the VLNF is possible for reusable sensor with very stable operation.

In order to estimate the NO detection limit of the VLNF and plain sensors, the response values, S−1, are plotted as a function of NO concentration in a linear scale in Fig. 15. The linear relationship between the response value and the concentration for the VLNF sensor demonstrates the feasibility and the operation capabilities of

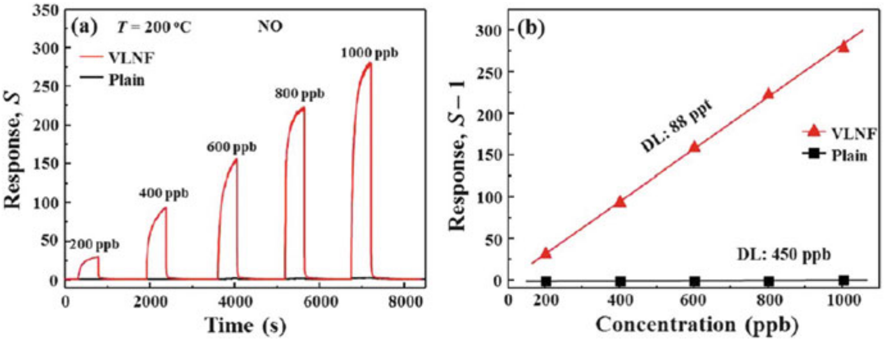


Fig. 11 (**a**) Dynamic sensing transients of the VLNF sensor and a sensor based on a dense plain WO_3 film (plain sensor) to 0.2−1 ppm NO at 200 °C in 80 % of relative humidity atmosphere, (**b**) Response of the VLNF and plain sensors as a function of NO concentration at 200 °C. Theoretical detection limits (DL) of both sensors are presented (Reproduced from Moon et al. [14])

the sensor for real applications. By applying linear least-square fits to the data, the theoretical detection limit of the VLNF sensor (signal-to-noise ratio > 3) is estimated to be as low as 88 ppt, whereas that of the plain sensor is only 450 ppb. The NO detection limit of the VLNF sensor is much lower than the asthma diagnostic standard levels for NO (30−50 ppb).

This result strongly suggests that a main factor for the ultrahigh response of the VLNF sensor is indubitably the porous nanostructure with narrow necks between intergrain boundaries of the WO_3 formed by annealing at 500 °C for 1 h. In general, semiconductor gas sensor utilizes the change of electrical resistance by gas absorption in potential barrier height between grain boundaries (transducer function), and it is well known that the response increases with decreasing the particle size. The whole region of the narrow necks becomes electron depletion area (space charge layer) by NO adsorption. This phenomenon significantly increases the double Schottky barrier heights in the intergrain boundaries of the WO_3 aggregates, which results in a large increase of the conductance for the nanocrystalline material upon exposure to NO. Therefore, we believe that the narrow necks with high chemoresistive variation as well as the gas accessible nanostructures with high surface area play a critical role in the remarkably enhanced sensing properties.

3.3 Room Temperature Gas Sensors

Adsorption and desorption of gas molecules on the surface of metal oxides are thermally activated processes, which cause the response and recovery times to be usually very slow at room temperature. Thus, gas sensors based on 1D oxide nanostructures operate at high temperature (150–400 °C) to enhance the surface molecular adsorption/desorption kinetics and continuously clean the surface. Development of room-temperature gas sensors might have very important advantages such as low power consumption, simple system configuration, reduced explosion hazards, and longer device lifetime.

Desorption of gas molecules typically requires much higher activation energy than adsorption. Law et al. [15] first demonstrated SnO_2-nanoribbon-based gas sensor operating at room-temperature by desorbing attached NO_2 gas molecules using ultraviolet (UV) irradiation. UV-assisted desorption of NO_2 was explained as follows: before UV illumination, oxygen species are adsorbed on the SnO_2 nanoribbon surface, taking free electrons from the n-type SnO_2 nanoribbon and forming a depletion region that extends into the thin nanoribbon. When the SnO_2 nanoribbon is illuminated by UV light with wavelength shorter than the bandgap energy of SnO_2, electron-hole pairs are generated. The positive holes discharge the negatively charged oxygen ions chemisorbed on the nanoribbon surface and eliminate the depletion region. Electrons produced at the same time increase the conductivity of the SnO_2 nanoribbon.

More recently, Prades et al. [16] demonstrated that illuminating metal oxide gas sensors with UV light is a viable alternative not only to activate but also to modulate their response towards oxidizing gases. Under dark (non-illuminated) conditions, nanowires exhibited extremely low responses S to NO_2 at $T = 25$ °C without any noticeable recovery of the resistance baseline. On the contrary, the same devices displayed significant and reversible responses to NO_2 pulses (concentrations from 100 ppb to 10 ppm) with characteristic response and recovery time constants of only a few minutes under constant UV illumination. It is noteworthy that sensor response to NO_2 scaled up with the energy of the impinging photons. The dependence of the response S of these devices on photon energy (E_{ph}) is directly related to the capacity of photons to transfer energy to adsorbed NO_2 molecules, which facilitate their desorption from the SnO_2 surface. If $E_{ph} > E_{bandgap}$, the sensor recovery time is minimized and the gas response S maximized, suggesting that band-to-band photoexcited pairs contribute to a fast desorption of NO_2 adsorbates after their separation by the surface built-in potential. Under UV illumination, photons partially desorb oxygen species from the surface, providing an increased number of adsorption sites available for gas molecules. Thus, the gas response of semiconductor metal oxide nanostructures with UV light illumination was about several hundred times higher than that without UV light illumination.

Alternatively, Fan et al. [17] reported gate-refreshable nanowire gas sensors. They developed highly sensitive room-temperature chemical sensors based on ZnO nanowire field-effect transistors for detection of NO_2 and NH_3. The electric field applied over the back gate electrode modulates the carrier concentration, which in turn significantly affects adsorption and desorption behaviors of gas molecules or gas sensitivity. A strong negative field was utilized to refresh the sensors by an electrodesorption mechanism. In addition, different chemisorbed species could be distinguished from the "refresh" threshold voltage and the temporal response of the conductance. Using the field-effect transistor sensor, they found that the total chemisorption coverage of NH_3 is less than that of NO_2, which has been confirmed by the observation that NW FETs tend to be less sensitive to NH_3 than to NO_2 and the desorption energy needed for NO_2 is greater than that for NH_3. This indicates the higher surface binding strength for NO_2 than for NH_3 and also accounts for the differences on their refresh voltage and recovery rate.

UV illumination and electrorefreshment are interesting methods to obtain room temperature semiconductor metal oxide gas sensors. However, there are still obstacles to be overcome. For UV assisted gas sensors, UV light source which is bulky, expensive, and consumes large electric energy seems to a big burden for application in portable personal devices. Gate refreshment requires relatively very high voltages and the recovery speed is quite slow even under the high bias voltages. Furthermore, most of room temperature metal oxide gas sensors ever reported were limited to the detection of several gases such as NO_2, H_2S, NH_3, and H_2. To realize the commercialization of room temperature gas sensors, higher response, faster response and recovery speed, the detection of volatile organic compounds, and the minimization of humidity influence on gas sensing performance should be accomplished at room temperature in absence of external activation sources.

4 Outlook

In 2008, Finnish company Nokia unveiled Morph which is a concept mobile phone (https://research.nokia.com/morph). Morph concept technologies could create fantastic opportunities for mobile devices: newly-enabled flexible and transparent materials blend more seamlessly with the way we live; devices become self-cleaning and self-preserving; transparent electronics offering an entirely new aesthetic dimension; built-in solar absorption might charge a device, whilst batteries become smaller, longer lasting and faster to charge; integrated sensors might allow us to learn more about the environment around us, empowering us to make better choices. Nanosensors would empower users to examine the environment around them in completely new ways, from analyzing air pollution, to gaining insight into biochemical traces and processes. New capabilities might be as complex as helping us monitor evolving conditions in the quality of our surroundings, or as simple as knowing if the fruit we are about to enjoy should be washed before we eat it, as shown in Fig. 12. Our ability to tune into our environment in these ways can help us make key decisions that guide our daily actions and ultimately can enhance our health.

Readily, NASA announced developing a proof of concept of new technology that would bring compact, low-cost, low-power, high-speed nanosensor-based chemical sensing capabilities to cell phones in 2009 (http://www.nasa.gov/centers/ames/news/features/2009/cell_phone_sensors.html). The device NASA developed is about the size of a postage stamp and is designed to be plugged in to a mobile device to collect, process, and transmit sensor data, as shown in Fig. 13. The device was able to detect and identify low concentrations of airborne ammonia, chlorine gas and methane. The device senses chemicals in the air using a sample jet and a

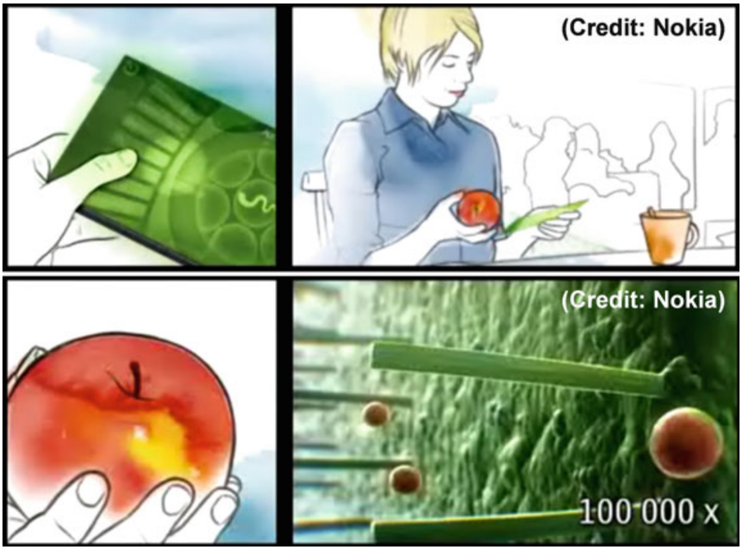


Fig. 12 Nokia's Morph, concept mobile phone with a gas sensor which can indicate the condition of a fruit (February, 2008)

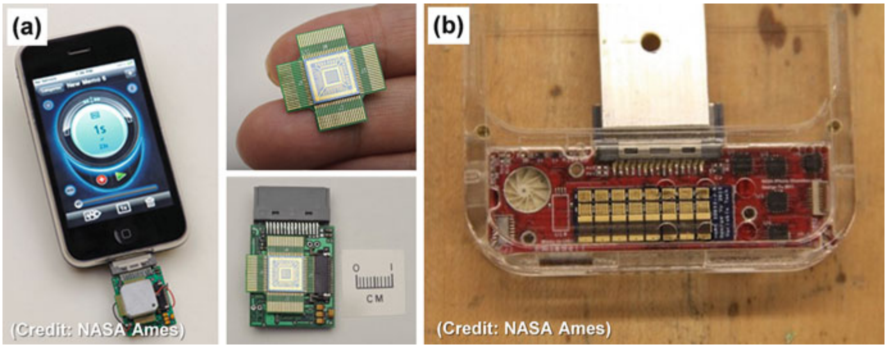


Fig. 13 (**a**) NASA's mobile phone-based gas sensor array (November, 2009), (**b**) NASA's cancer-sniffing cell phone sensor array (February, 2012)

multiple-channel silicon-based sensing chip, which consists of 16 nanosensors, and sends detection data to another phone or a computer via telephone communication network or Wi-Fi.

In 2102, the same group in NASA announced the development of a small chip sensor about the size of a postage stamp, which houses 32 nanosensor bars (http://gizmodo.com/5881097/this-is-nasas-cancer-sniffing-cellphone-sensor). Each bar is composed of a different nanostructure material. Because each sensor bar is unique it can respond to different chemicals in different ways, enabling it to not only differentiate between them, but also to monitor their relative levels, in real time.

In its state, it is housed in a small case that attaches to a smartphone. The idea was to develop a low-cost version so that consumers can afford to have them for health and safety applications. The cell phone implementation is aimed squarely at consumers. The chip only draws 5 mW, which means very little battery-drain (the smartphone they tested it with can use the sensor for 8 continuous hours on a single charge). It is primarily being developed to monitor carbon monoxide as well as chlorine, ammonia, and methane in your home. But these things could really be used anywhere because they're so small. An app could automatically send data back to the Department of Homeland Security or other emergency services agencies, which would give them a big-picture look at a larger area and let them know if a mass evacuation is needed.

The most exciting potential use, though, is how it could diagnose and monitor people with medical conditions. For example, for diabetes patients there is a direct correlation between the level of acetone in their breath and the level of sugar in their blood. The nanosensor could be used as a completely noninvasive diagnosis and measurement method. There is also a correlation between nitrous oxide and lung cancer. Breathing on your phone could give you the early warning you need to catch it in time. In the case of diabetes, acetone levels in the breath can indicate high or low blood sugar. It's feasible to think that this device could help to do away with having to test blood sugar levels carried out by collecting blood samples from a patient. The nanosensor device could help to catch the disease early on, something that could save a person's life.

In 2013, Samsung released the Galaxy S4 mobile phone which adopted a humidity sensor mounted in the phone. It is the first mobile phone that has a built-in chemical sensor. Although many commercialized plug-in types of gas sensors for mobile phones exist, there is no built-in mobile phone gas sensor yet. Nonetheless, developing built-in gas sensors is attracting great attention. In December 2012, IBM's 5-in-5 list predicted that the five sense-related technologies enabled by cognitive computing systems that would impact our lives in the next 5 years. One of the technologies is that computers will have a sense of smell, as shown in Fig. 14 (http://www-03.ibm.com/press/us/en/pressrelease/39685.wss). We already have electronic devices that can "smell"—the most obvious example is the breath analyzer that detects alcohol from a breath sample. But IBM said that electronic noses are set to become much more widespread and will provide a valuable tool for doctors. By examining the molecular biomarkers present in our breath, tiny sensors that are small enough to be integrated into mobile phones or other mobile devices will be able to provide valuable diagnostic information about our physical health. Similar technology already exists, such as an "artificial nose" that can sniff out bacterial infections, and another that can detect narcotics and explosives. IBM says it has already demonstrated the ability to measure biomarkers down to a single molecule using relatively simple sensing systems and believes it won't be long before the technology is sniffing out various ailments, such as liver and kidney disorders, diabetes and tuberculosis, amongst others.

It is apparent that the world market for semiconductor metal oxide gas sensors will expand rapidly next 10 years due to applications in personal handheld devices

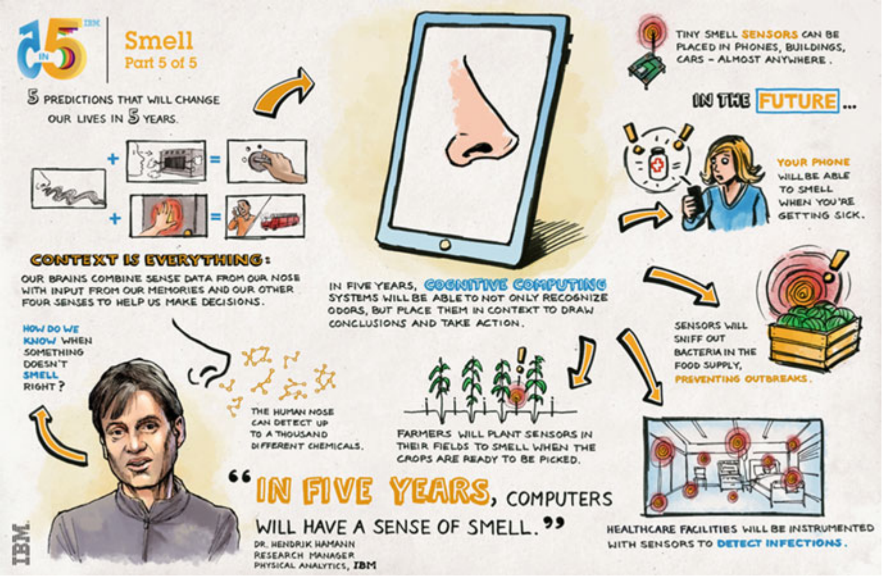

5 IN 5 IBM
Smell
Part 5 of 5
5 PREDICTIONS THAT WILL CHANGE OUR LIVES IN 5 YEARS.
+
=
+
=
CONTEXT IS EVERYTHING:
OUR BRAINS COMBINE SENSE DATA FROM OUR NOSE WITH INPUT FROM OUR MEMORIES AND OUR OTHER FOUR SENSES TO HELP US MAKE DECISIONS.
HOW DO WE KNOW WHEN SOMETHING DOESN'T SMELL RIGHT?
THE HUMAN NOSE CAN DETECT UP TO A THOUSAND DIFFERENT CHEMICALS.
IN FIVE YEARS, COGNITIVE COMPUTING SYSTEMS WILL BE ABLE TO NOT ONLY RECOGNIZE ODORS, BUT PLACE THEM IN CONTEXT TO DRAW CONCLUSIONS AND TAKE ACTION.
FARMERS WILL PLANT SENSORS IN THEIR FIELDS TO SMELL WHEN THE CROPS ARE READY TO BE PICKED.
"IN FIVE YEARS, COMPUTERS WILL HAVE A SENSE OF SMELL."
DR. HENDRIK HAMANN
RESEARCH MANAGER
PHYSICAL ANALYTICS, IBM
IBM
TINY SMELL SENSORS CAN BE PLACED IN PHONES, BUILDINGS, CARS - ALMOST ANYWHERE.
IN THE FUTURE...
YOUR PHONE WILL BE ABLE TO SMELL WHEN YOU'RE GETTING SICK.
SENSORS WILL SNIFF OUT BACTERIA IN THE FOOD SUPPLY, PREVENTING OUTBREAKS.
HEALTHCARE FACILITIES WILL BE INSTRUMENTED WITH SENSORS TO DETECT INFECTIONS.

Fig. 14 "In 5 years, computers will have a sense of smell" from IBM's 5-in-5 list of 2012 (December, 2012)

like cell phones and tablet computers. Thus, the development of highly sensitive, reliable, low cost, small size and low power consumption semiconductor gas sensors is very important to realize tiny sensors that smell can be integrated into cell phones and other mobile devices, feeding information contained on environmental gases, order, and flavor and exhaled biomarkers to a computer system that can analyze the data. The development of Internet of Things (IoT) that refers to the interconnection of uniquely identifiable embedded computing like devices within the existing Internet infrastructure has led to a sharp rise in demand for smart gas sensors and even for electronic noses.

Acknowledgments This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Liu X, et al. A survey on gas sensing technology. Sensors. 2012;12:9635–65.
2. Moon HG, et al. Self-activated ultrahigh chemosensitivity of oxide thin film nanostructures for transparent sensors. Sci Rep. 2012;2:588.
3. Yamazoe N, Shimanoe K. New perspectives of gas sensor technology. Sensors Actuators B. 2009;138:100–7.
4. Moon HG, Jang HW, Kim JS, Park HH, Yoon SJ. Mechanism of the sensitivity enhancement in TiO_2 hollow-hemisphere gas sensors. Electron Mater Lett. 2010;6:135–9.

5. Lee JH. Gas sensors using hierarchical and hollow oxide nanostructures: overview. Sensors Actuators B. 2009;140:319–36.
6. Moon HG, et al. Embossed TiO_2 thin films with tailored links between hollow hemispheres: synthesis and gas-sensing properties. J Phys Chem C. 2011;115:9993–9.
7. Shim YS, et al. Au-decorated WO_3 cross-linked nanodomes for ultrahigh sensitive and selective sensing of NO_2 and C_2H_5OH. RSC Adv. 2013;3:10452–9.
8. Hwang S, et al. A near single crystalline TiO_2 nanohelix array: enhanced gas sensing performance and its application as a monolithically integrated electronic nose. Analyst. 2013;138:443–50.
9. Kim HJ, et al. Ultraselective and sensitive detection of xylene and toluene for monitoring indoor air pollution using Cr-doped NiO hierarchical nanostructures. Nanoscale. 2013;5: 7066–73.
10. Kim HR, et al. The role of NiO doping in reducing the impact of humidity on the performance of SnO_2-based gas sensors: synthesis strategies, and phenomenological and spectroscopic studies. Adv Funct Mater. 2011;21:4456–63.
11. Elmi I, Zampolli S, Cozzani E, Mancarella F, Cardinali GC. Development of ultra-low-power consumption MOX sensors with ppb-level VOC detection capabilities for emerging applications. Sensors Actuators B. 2008;135:342–51.
12. Strelcov E, et al. Evidence of the self-heating effect on surface reactivity and gas sensing of metal oxide nanowire chemiresistors. Nanotechnology. 2008;19:355502.
13. Prades JD, et al. Ultralow power consumption gas sensors based on self-heated individual nanowires. Appl Phys Lett. 2008;93:123110.
14. Moon HG, et al. Extremely sensitive and selective NO probe based on villi-like WO_3 nanostructures for application to exhaled breath analyzers. ACS Appl Mater Interfaces. 2013;5:10591–6.
15. Law M, Kind H, Messer B, Kim F, Yang PD. Photochemical sensing of NO_2 with SnO_2 nanoribbon nanosensors at room temperature. Angew Chem Int Ed. 2002;41:2405–8.
16. Prades JD, et al. Equivalence between thermal and room temperature UV light-modulated responses of gas sensors based on individual SnO_2 nanowires. Sensors Actuators B. 2009;140:337–41.
17. Fan ZY, Lu JG. Gate-refreshable nanowire chemical sensors. Appl Phys Lett. 2005;86:123510.

Handheld Gas Sensing System

**Shih-Wen Chiu, Hsu-Chao Hao, Chia-Min Yang, Da-Jeng Yao,
and Kea-Tiong Tang**

Abstract Handheld gas sensing systems have drawn attentions recently for personal use and daily applications. However, commercially available gas detection devices are yet to satisfy the needs due to the challenging issues of system miniaturization, such as insufficient selectivity and sensitivity. In this chapter, we introduce an approach to achieve this goal. Based on an array of surface acoustive wave (SAW) gas sensors, a bio-inspired gas sensing system (also called electronic nose) could be realized to construct a robust system to identify gases. To increase the gas sensitivity, nanocomposites of polymers and ordered mesoporous carbons (OMCs) is introduced. The polymers are directly grown on the carbon material through a radical polymerization process, thus forming interpenetrating and inseparable composite frameworks with carbon. Furthermore, to reduce the system size and power consumption, the integrated circuits (IC) technology is adopted to implement the readout interface circuit to replace bulky instruments, such as frequency counter. Finally, several odor classification algorithms are introduced to perform gas classification.

1 Introduction

Organic vapors are colorless and have sharp, penetrating, and intensely irritating odors under pressure. Such odors could be corrosive and fatal if inhaled. Therefore, detecting organic vapors is essential for numerous processes such as human safety [1], environmental monitoring [2], and clinical diagnosis [3]. Although several commercial gas sensing products are available on the market, many of them are bulky, and require a desktop or laptop computer, which makes them unsuitable for portable

S.-W. Chiu • K.-T. Tang (✉)
Department of Electrical Engineering, National Tsing Hua University, Hsinchu City, Taiwan
e-mail: kttang@mx.nthu.edu.tw

H.-C. Hao • D.-J. Yao
Institute of NanoEngineering and MicroSystems, National Tsing Hua University,
Hsinchu City, Taiwan

C.-M. Yang
Department of Chemistry, National Tsing Hua University, Hsinchu City, Taiwan

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_8

or handhold applications. There are indeed a number of modern small electronic noses, such as the “Diagnose” from C-it of the Netherlands (11 × 18 × 7 cm) and the Artinose from SYSCA AG Germany (17 × 26 × 14 cm), but they are still too expensive for widespread adoption. In recent years, the requirement for handheld gas detection devices for personal use has increased considerably for nonexpert users in daily life, such as bad breath detectors for testing halitosis, alcohol detector, meat fresh meter and indoor air quality monitoring device. However, the personal handheld gas sensing system has yet to achieve the full potential. There are several obstacles to popularize the personal device, such as insufficiently robust selectivity and sensitivity after miniaturization. To achieve a widespread handheld gas sensing system, we can approach the miniaturized gas sensing system in aspects of sensor device, sensing material, readout electronics and signal processing methods.

Gas detection methods can be categorized into two groups: (1) direct methods, which involve monitoring a physical parameter of a target gas, and (2) indirect methods, which involve using a chemical reaction or indicator to determine the concentration of a gas being sensed. Various types of sensors, including conductive sensors (metal-oxide semiconductor; conducting polymer [4–7], piezoelectric sensors (quartz crystal microbalance; surface acoustic wave, SAW) [8–11], optical sensors [12], and spectroscopy-based sensors (mass spectrometer; ion mobility spectroscope) [13], have been used for sensing gases. Electrochemical sensors, such as conductive gas sensors and piezoelectric gas sensors, usually contain various electrodes, including sensing, counter, and reference electrodes, and a membrane that is porous to gas, but not to liquid. As air diffuses into the cell, the sensor adsorbs certain gases and a parameter differential is produced. The current produced by the chemical reaction is proportional to the concentration of the reacting gas.

1.1 Polymer-Coated SAW Gas Sensor

Polymer-coated SAW sensors are among the most favorable sensors used for achieving high sensitivity in applications that require detecting organic gasses. Currently, SAW devices are used for various chemical applications because of their high sensitivity, fully reversible behavior, and high signal-to-noise ratio [14]. For a portable gas detection system based on an SAW sensor [15], a piezoelectric substrate was used for transforming energy between mechanical strains and electric signals. For interdigital transducers (IDTs), input and output comb-like metal electrodes were used as energy transformation structures on the surface of a selected substrate. When an AC voltage was applied to the input IDTs, the IDTs induced dynamic strains in the substrate; these strains launched a surface wave on the top of the substrate. The induced surface wave was propagated through the active sensing region, after which it was received and transformed into electric signals by the output IDTs. According to the magnitude change and phase shift of the AC electric signals between the input and output IDTs, a change in mass was detected on the active sensing region. To increase the selectivity and sensitivity of the sensor, different polymers were coated onto the active sensing region of the SAW sensors to absorb target gas molecules. For example, poly-N-vinylpyrrolidone (PNVP) as one of the sensing films because of its high selectivity in sensing amine vapors. Several studies have been conducted to develop pernicious gas sensors by using organic layers [16], semiconducting metal oxides [17], and self-assembled layers [18, 19]. However, the persistence and reproducibility of sensing devices based on organic layers were poor. Although semiconducting metal oxide films have shown favorable performance, most of such films are suitable only for high-temperature sensing, which is not appropriate for a portable sensing system. However, polymer-based sensing films have shown favorable performance at room temperature, and are thus more appropriate for a portable E-nose system. For example, a self-assembled polymer layer—polydiacetylene/calyx[n]arene [16]—that was assembled on an ST-quartz (SiO_2) SAW sensor chip demonstrated a favorable linear response to amine vapors at room temperature [20].

1.2 Bioinspired Gas Sensing System

Inspired by the structure of mammalian olfactory systems [21, 22], electronic nose systems primarily consist of a sensor array, signal transducer, and pattern recognition engine (Fig. 1) [23–26]. In mammalian olfactory systems, olfactory receptor cells are crucial sensory cells for sensing odors. The nasal cavity has six to ten million olfactory receptor cells [27]; the human genome contains approximately 900 different olfactory receptor genes, whereas the mouse genome contains approximately 1,300 olfactory receptor genes [28]. When an odor enters the nasal cavity, olfactory signals are activated in the olfactory receptor cells. Olfactory bulbs collect and convert the olfactory signals into neurological signals and subsequently send these signals to the brain for odor identification. Although numerous types of olfactory receptor cells exist, the odor identification system of mammals is not based on one type of receptor cell for a specific odor. Conversely, scents are sensed and recognized by an array of multiple receptor cells; each combination senses a different odor that represents an odor “fingerprint.” Several permutations and combinations exist, and these combinations enable mammals to distinguish various odors. A similar system is adopted for the electronic nose. The unselected sensors that form the sensor array are used to detect odors, generating and identifying the odor “fingerprint”. However, for most electronic nose systems, the number of sensors is limited; typically, a number of sensors ranging from 4 to 32 is used depending on the application [29].

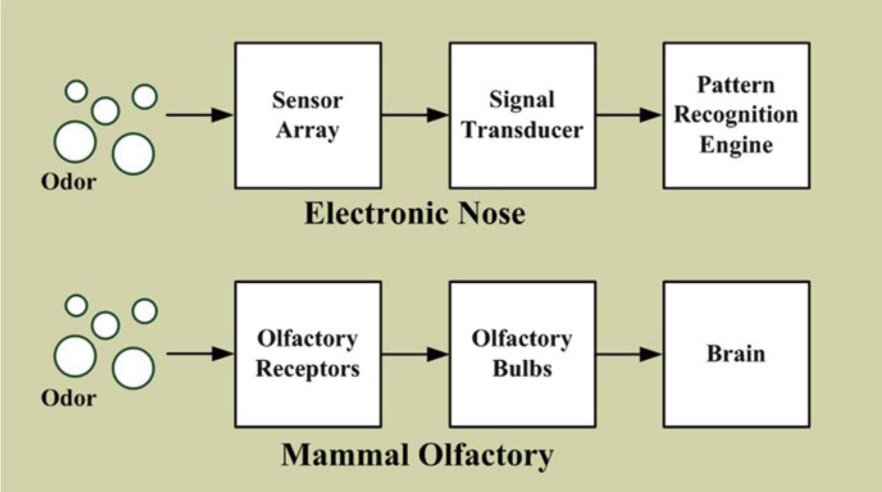


Fig. 1 The basic gas identification system blocks: an electronic nose and a mammal olfactory

1.3 Overview

This chapter introduces the methods used in implementing the handheld gas sensing system. Section 1 presents the introduction and an overview of bioinspired gas sensing systems based on SAW gas sensors. Section 2 presents the implementation of an SAW device. Section 3 presents the nanocomposite platelet sensing film for the SAW device. Section 4 presents the frequency sensor signal readout electronics and their application-specific integrated circuit (ASIC) as well as odor clustering methods. Finally, Sect. 5 presents the conclusion and discussion. The chemical SAW sensors were fabricated using MEMS technology, and oscillator circuits were designed to improve sensing results. An SAW delay line was constructed onto a 128° YX-lithium niobate ($LiNbO_3$) piezoelectric substrate with a high K^2 value to fabricate SAW devices, and the polymers were coated on the active sensing region by a self-assembled process to create 2×2 discontinuous SAW array chips. The polymer selection and preparation as chemical interfaces for the SAW array were tested and proved through a two-way hierarchical clustering analysis. The frequency shift readout electronics comprised a multiplexer, counter, and microprocessor. The frequency variation between the start and steady states of the SAW device was determined to be lower than 1 kHz. In addition, this study investigated the selectivity, sensitivity, reversibility, and chemical characteristics of the sensor at room temperature. Using ASICs to fabricate frequency readout circuits reduces the size of the system and cost of mass production [30]. A low-power, high-resolution ASIC chip was fabricated using a TSMC 0.18-μm 1P6M standard CMOS process. The ASIC chip was connected directly to an array comprising four SAW sensors to output frequency data representative of the sensor response.

To enhance gas sensitivity, nanocomposites of polymers and platelet CMK-5-like carbon were prepared; these materials demonstrated superior performance in detecting gravimetric gas. A zirconium-containing platelet SBA-15 was used as a hard template to prepare the CMK-5-like carbon, which was then used as a lightweight and high-surface-area scaffold for growing polymers through radical polymerization. Mesoporous nanocomposites composed of four distinct polymers were used as sensing materials for the SAW devices to detect ammonia gas in parts-per-million levels. The sensors showed higher sensitivity and reversibility than those coated with dense polymer films, and the sensor array generated a characteristic pattern for an analyte with a concentration of 16 ppm. The results showed that the nanocomposite sensing materials are promising materials for highly sensitive gravimetric-type gas detection applications.

2 SAW Gas Sensor

Acoustic wave devices have been in commercial use for more than 30 years. The telecommunications industry is the largest consumer of these devices and uses these devices primarily for base stations and cellular telephones. SAW devices typically function as bandpass filters in both radio and intermediate frequency sections. Several emerging applications of acoustic sensor devices might eventually have demanded in market for telecommunications. SAW-based gas sensing devices are widely used in applications such as food product quality control [31, 32], determining the quality of indoor air [33, 34], environmental monitoring [35–38], human safety [39, 40], industrial hygiene [41], clinical diagnosis [42], and military applications. SAW devices have attracted considerable attention because of their sensitivity to variations in physical and chemical properties at or near the surface of a transducer [43]. Any discussion of the application of acoustic waves in sensor technology must begin with the phenomenon called *piezoelectricity*. In 1880, the brothers Pierre and Paul-Jacques Curie discovered piezoelectricity, which received its name from Hankel in 1881. Piezoelectricity largely remained a curiosity until 1921, when Cady discovered that a SiO_2 resonator can be used to stabilize electronic oscillators [44, 45]. A piezoelectric device produces electrical charges by imposing a mechanical stress, which is created when an appropriate electrical field is applied to a piezoelectric material. Piezoelectric acoustic-wave sensors apply an oscillating electric field to create a mechanical wave that propagates through a piezoelectric crystal substrate and is then converted to an electric field for measurement [46, 47]. The parameter K^2 is an indicator of the capacity for translation between electric and mechanical potential. SAW-based Filters were first applied in radar and sonar. Numerous advancements have recently occurred in the field of telecommunications, in which SAW devices have become crucial elements used as filters, demodulators, and oscillators [47, 48]. King, in 1964, published the first report of a bulk-acoustic-wave device used as a selective analytical sensor [49, 50]. This device was coated with typical stationary-phase materials and incorporated as a detector into a gas-liquid chromatographic instrument. Wihltjen and Dessy proposed the SAW chemical sensor, which was like the bulk-wave device, in 1979 [29, 51].

Acoustic wave sensors are sensitive, inherently rugged, inexpensive, and intrinsically reliable. Some of these sensors can be passively and wirelessly integrated [52, 53]. Acoustic wave sensors are thus designated because their detection mechanism involves a mechanical or acoustic wave. When an acoustic wave is generated on or through the surface of a material, any variation of the characteristics of the generating path affects the velocity or amplitude of the wave. An altered velocity can be detected by measuring the frequency or phase characteristics of the sensor, and this velocity can then be correlated with the corresponding physical quantity being measured. Such devices comprise a piezoelectric crystal and at least one layer of a chemically interactive material deposited on the surface to produce sensitivity to a certain chemical compound. All acoustic-wave devices and sensors use a piezoelectric material to propagate the acoustic wave at a constant frequency that ranges from megahertz to gigahertz. When molecules are adsorbed directly onto the surface of an SAW device or a coated film, fluctuations of SAWs are observed in the frequencies at which they vibrate [54]. Fluctuations of SAW characteristics are proportional to the mass of the targeted molecules deposited on the surface of the crystal with respect to the central frequency of the piezoelectric crystal.

2.1 Design of an SAW Device

In 1887, Lord Rayleigh discovered the mode of propagation of SAWs [45, 55] and predicted the properties of these waves in his classic paper. Such coupling strongly affects the amplitude and velocity of the wave. This feature enables SAW sensors to sense mass and mechanical properties directly. The surface motion also enables the devices to serve as microactuators. An SAW has a velocity that is 10^{-5} to 10^{-4} times higher than that of a corresponding electromagnetic wave; thus, Rayleigh surface waves are among the slowest waves propagated in solids [49]. Before the application of SAW sensors to electronics became generally acknowledged, seismologists and researchers interested in nondestructive testing accumulated considerable information. The Rayleigh wave is nondispersive; its displacement decays exponentially from the surface, causing more energy (generally more than 95 %) to be confined within a depth equal to one wavelength [43, 45, 56, 57].

Materials that serve as a piezoelectric substrate can be used in acoustic-wave sensors and devices; the most common of such materials are SiO_2, lithium tantalate ($LiTaO_3$), and, to a lesser extent, $LiNbO_3$. Each material has specific benefits and disadvantages including high temperature dependence, cost, attenuation, and low propagation velocity. Table 1 lists some relevant specifications for each material, including the most common cuts and orientations [58]. The temperature dependence of SiO_2 is selectable according to the cut angle; the direction of wave propagation and first-order effect of temperature can be minimized. SAW temperature sensors

Table 1 Parameters of substrate material of various types and applications [64]

Material	Crystal cut	SAW axis	Velocity/ m s^{-1}	K^2/%	Temperature coefficient of delay/ppm °C^{-1}	Capacitance/ finger pair/unit length C_0/pF cm^{-1}	Application
Quartz	ST	*x*	3,158	0.11	∼0	0.55	Oscillator resonator
$LiNbO_3$	*Y*	*z*	3,488	4.5	94	4.6	Wideband IF filter
$LiNbO_3$	128°	*x*	3,992	5.3	75	5.0	Wideband IF filter
$Bi_{12}GeO_{20}$	110	001	1,681	1.4	120	–	Long delay line
$LiTaO_3$	77.1° rotated *Y*	*z*	3,254	0.72	35	4.4	Oscillator
GaAs	(100t)	<110>	2,841	0.06	35	–	Semi-conductor IC

are designed to maximize this effect. This condition is not applicable to $LiNbO_3$ or $LiTaO_3$ because, for these materials, a linear dependence on temperature exists for all material cuts and propagation directions. Other materials that have commercial prospects are lead zirconium titanate, silicon carbide, gallium arsenide, zinc oxide, aluminum nitride, and polyvinylidene fluoride [52, 59–63].

An SAW can be most conveniently excited on a piezoelectric crystal or piezoelectric thin film that has an IDT. IDTs are widely used to excite electrical signals and detect SAWs [55]. Each IDT period comprises multiple strips aligned and connected periodically to bus bars. The variable center frequency and phase response are based on the IDT design and are measured using a network analyzer. Using a voltage between alternately connected finger electrodes results in the imposition of a periodic electric field on the piezoelectric substrate. When an alternating voltage is applied, a periodic strain field that produces a SAW is generated in the piezoelectric substrate. This field generates multiple waves that are introduced from the transducer; the wave fronts are parallel to the transducer fingers. IDTs have been widely used for exciting electric signals and detecting SAWs (Fig. 2).

The electrical characteristics of an IDT are determined according to the geometry of the fingers in each period, number of finger pairs, and material used as the substrate. The central frequency of an SAW device can be calculated using a simple velocity equation (1),

$$v = f \cdot \lambda \tag{1}$$

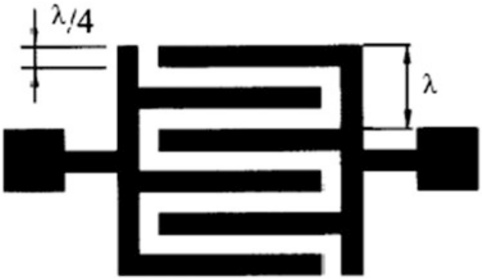


Fig. 2 Schematic illustration of an interdigital transducer

where *v* represents the velocity of the surface wave for the substrate used, *f* denotes the central frequency of the designed SAW device, and λ is the wavelength of the surface wave. The wavelength of the surface wave is a function of the width of the fingers and the spacing between them. The spacing of a single IDT is one quarter of the wavelength (λ/4), providing the definition required for photolithography. An IDT operates most efficiently when the SAW wavelength matches the periodicity of the transducer (Fig. 2); this condition occurs when the transducer is excited at the synchronous frequency. For simplicity, a simple velocity equation (2) is used to design the central frequency of an SAW device:

$$f_0 = \frac{v}{\lambda_T} = \frac{v}{4d} \tag{2}$$

where *v* is the velocity of the surface wave for the substrate used, f_0 denotes the central frequency at which the minimum insertion loss in the frequency response of the designed SAW device occurs, and λ_T represents the wavelength of the surface wave. The spacing of an IDT of a single electrode type is one quarter of the wavelength ($\lambda_T/4$), providing the required definition for photolithography.

An acoustic wave multiplying through a piezoelectric substrate is followed by a wave of electric potential. The acoustic wave is influenced by an applied electric field or the properties of a conducting or semiconducting surface layer. The two most crucial parameters of a material used in designing an SAW filter are K^2 and the SAW velocity *v*. The electromechanical coupling coefficient K^2 is a measure of the efficiency of a specified piezoelectric device in converting an applied electrical signal into mechanical energy related to a SAW. The short-circuit condition, which can be obtained by applying a thin guiding layer to the surface, is especially crucial to SAW devices because this condition influences both the performance of an IDT and operation of an SAW chemical sensor. The semiconducting layer partly shorts the piezoelectric field, thus changing the velocity and attenuating the wave. These effects can be characterized according to the electromechanical coupling parameter. The K^2 parameter indicates the capacity of translation between electric and mechanical potentials. The value of K^2 can be calculated as follows:

$$K^2 = \frac{e^2}{c\varepsilon} = 2\left|\frac{\Delta v}{v}\right| \tag{3}$$

where e is the piezoelectric coefficient, c represents the elastic coefficient, and ε denotes the dielectric coefficient of the substrate. The value of K^2, which can be calculated using Eq. (3), depends on the properties of the piezoelectric substrate or on the experimental results because of the velocity shift during metallization. Furthermore, Δv represents the magnitude of the SAW velocity change that occurs when the free surface of the piezoelectric is shorted with a thin and highly conductive metal film, and v denotes the unperturbed velocity of the SAW. $LiNbO_3$ was used as a chip substrate to produce a large acoustoelectric effect and, thereby, enhance the sensitivity of a chip compared with other substrates (Table 1).

When an SAW sensor is exposed to a target, the electrical and mechanical characteristics of the sensor vary according to the absorption capacity on the active sensing region; this phenomenon results in a frequency shift or phase shift. The effects of perturbation by mass loading include elastic loading and electric loading; the propagation characteristics of the surface wave are affected as follows [45, 65]:

An SAW sensor can detect a foreign mass present on its surface to a detection extent of 10^{-10} g. The frequency shift of an SAW delay-line oscillator fabricated using an acoustically thin and perfectly elastic thin film can be calculated according to perturbation theory as shown in Eq. (4) [66]:

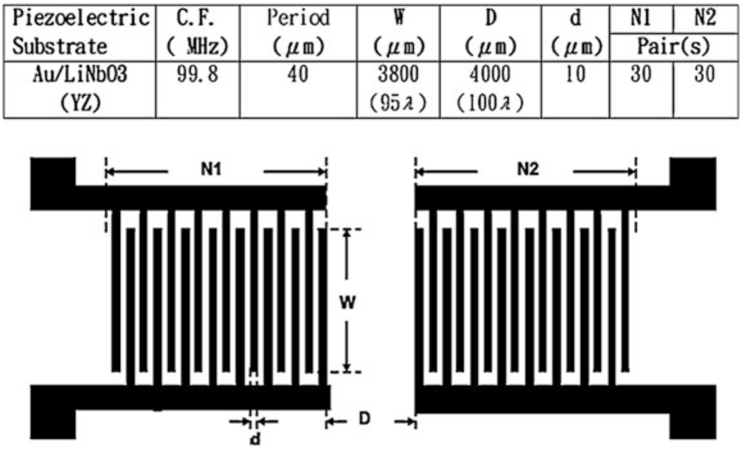
$$\Delta f \cong (k_1 + k_2)\ \ \mathrm{f_0}^2 m/A \tag{4}$$

where k_1 and k_2 represent piezoelectric parameters of the material, f_0 is the central frequency (Hz) of the SAW device, m denotes the mass of a foreign compound deposited on the surface of the crystal, and A represents the sensing area of the thin-film polymer.

2.2 SAW Device Fabrication

A chemical sensor generally comprises two major components: a transducer and a sensitive coating. A transducer is required to transform information into a measurable signal. The parameter K^2 provides the capacity of translation between electric potential and mechanical potential [55]. The design parameters, which can be calculated using Eqs. (1)–(4), depend on the properties of the piezoelectric substrate or on the experimental results; this is because of the velocity shift under metallization. $LiNbO_3$ was used as the chip substrate to produce a strong acoustoelectric effect compared with that of other substrates, and a sensitive coating was deposited on the wave propagation path of this substrate. The wavelength of the surface wave is a function of the spacing between the fingers and width of the fingers. The spacing of a single-electrode-type IDT was one-fourth the wavelength (λ/4) (Fig. 3), providing the required definition for photolithography.

According to the design parameters derived from simulation data on delta-function and cross-field models, the SAW device was fabricated on a 128° YX-$LiNbO_3$ piezoelectric substrate with IDTs composed of Cr/Au (20 nm/100 nm)



Piezoelectric Substrate	C.F. (MHz)	Period (μm)	W (μm)	D (μm)	d (μm)	N1	N2
						Pair(s)	
Au/LiNbO3 (YZ)	99.8	40	3800 (95λ)	4000 (100λ)	10	30	30

Fig. 3 The parameters of SAW sensor

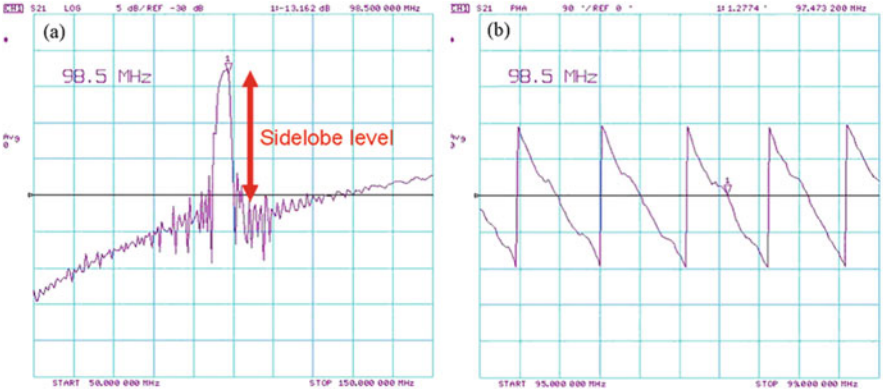


Fig. 4 The center frequency of SAW sensor is 98.5MHz, insertion loss is around −8.53 dB, and sidelobe rejection is 26dB, Q-value is about 50∼60 (**a**) Frequency response of SAW device (**b**) Phase response of SAW device

by using a standard photolithographic technique and operated at 99.8 MHz. The distance between the centers of two IDTs was 4 mm, the IDT aperture was 3.8 mm, the number of finger pairs was 30, and the period was 40 m, which is equal to the SAW wavelength. After the devices were fabricated, their electrical characteristics were measured using a network analyzer HP 8714C. To reduce energy loss, the network between the sensor and instrument must be matched. The center frequency of the SAW sensor was 98.5 MHz, the insertion loss was approximately −8.53 dB, the sidelobe rejection was 26 dB, and the Q-value was approximately 50–60 (Fig. 4). The characteristics of the sensor were extremely stable for temperatures between 30 and 80 °C and relative humidity (RH) between 45 % and 50 %. The simple regression equation for frequency shift (Y, MHz) and temperature (X, °C) was $Y = -0.10376 + 0.0032X$, $R^2 = 0.9961$.

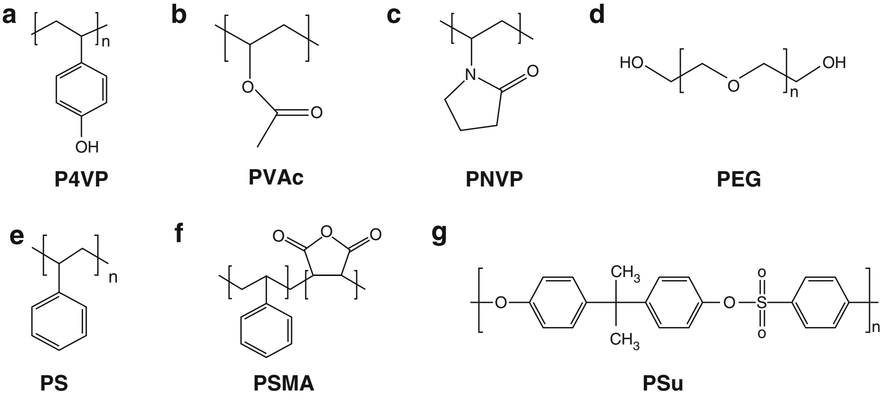


Fig. 5 Structural formula of polymer. (**a**) Poly-4-vinylphenol, P4VP; (**b**) poly-vinylacetate, PVAc; (**c**) poly-N-vinylpyrrolidone, PNVP; (**d**) polyethyleneglycol, PEG; (**e**) polystyrene, PS; (**f**) polystyrene-co-maleic anhydride, PSMA; and (**g**) polysulfone, PSu

2.3 Sensing Film: Polymer Coating

Seven polymers with different polarities and functional groups, namely poly-4-vinylphenol (P4VP), poly-vinylacetate (PVAc), poly-Nvinylpyrrolidone, polyethyleneglycol (PEG), polystyrene (PS), polystyrene-co-maleic anhydride (PSMA), and polysulfone (PSu) [18] (Fig. 5), were used to coat the sensing film. A simple spin-coating method was used to coat the surface of the SAW sensor with chemical interfaces. The thickness of the polymers was approximately 8–14 μm, and the frequency shift of the SAW was nearly 200–450 kHz. Dilute solutions of P4VP (in 95 % ethanol), PSMA, PVAc, PS (in methyl ethyl ketone), PSu, PEG, and PNVP (in tetrahydrofurane) were prepared. Table 2 shows the polymeric solutions that were deposited on the surface of a IDT device to spin-coat the thin film. After the spin-coating process, the device was allowed to rest for hours so that the solvent could evaporate completely.

The sensor was tested using ethanol for three consecutive times, and Fig. 6 shows the test results. In each test, the test chamber was vacuum-pumped for 10 min, and 50 μL of ethanol was then injected into the chamber for the subsequent 3–5 min; after 3–5 min, the valve was opened to allow the air to flow. Both the SAW sensor and sensor readout electronics exhibited high reproducibility.

Before the sensing layer was coated, the center frequency (denoted as f_0) of every SAW sensor was determined using a network analyzer. The seven polymers, PVAc, PNVP, P4VP, PS, Psu, PSMA, and PEG, were used as sensitive film materials to coat the surface of different sensors in the array through a spin-coating process. The frequency of the chip after coating (f_p) was then measured. The sensor chip was connected to the oscillator and sensing circuits, and the frequency of the chip f_c was measured. The sensor array was then tested to detect five vapors, and the frequency

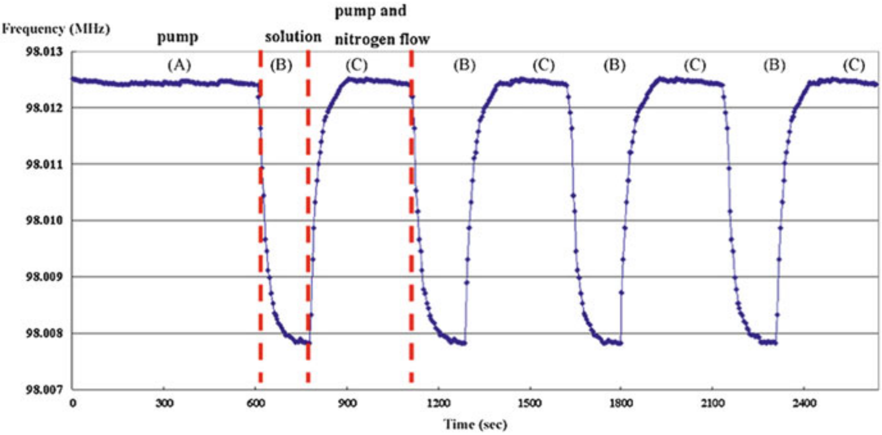


Fig. 6 Ethanol test results for the SAW sensor

Table 2 Measured frequency shifts

Polymer	Gas	Ethanol	Amine	TMA	Methanol	Acetone
P4VP	Average Δf_{gas} (kHz)	18.37	22.8	31.6	22.514	18.8
PNVP	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	3.877	4.87	7.0	4.62	3.813
	Average Δf_{gas} (kHz)	73.87	188.9	130.6	76.85	52.45
PVAc	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	14.423	35.941	25.795	15.263	10.35
	Average Δf_{gas} (kHz)	20.6	17.405	22.23	17.96	13.938
PS	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	3.074	2.635	3.332	2.7	2.11
	Average Δf_{gas} (kHz)	7.886	10.253	9.78	7.866	7.6755
PSMA	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	1.614	2.11	2.012	1.578	1.502
	Average Δf_{gas} (kHz)	10.375	13.239	12.421	8.304	9.536
PEG	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	1.644	2.13	1.935	1.299	1.503
	Average Δf_{gas} (kHz)	10.75	19.25	18.59	10.40	9.01
Psu	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	1.536	2.872	2.689	1.51	1.316
	Average Δf_{gas} (kHz)	12.18	15.9	14.95	10.709	11.49
	$\Delta f_{gas}/\Delta f_{polymer}$ (%)	1.749	2.31	2.11	1.515	1.637

Every data is repeated by at least eight times for statistics
Standard deviations of $\Delta f_{gas} < 3$ kHz

of the chip after vapor adsorption (f_m) was evaluated. Table 2 shows the measured frequency shifts ($\Delta f_{gas} = f_m - f_c$). A normalization method, $\Delta f_{gas}/\Delta f_{polymer}$ ($\Delta f_{polymer} = f_0 - f_p$), was used to eliminate the polymer and circuit effects (Fig. 7a). The gas concentration was subsequently estimated. Figure 7b shows radar plots of the normalized data for the seven membranes. A characteristic sensing fingerprint for the five gases is shown for each sensing polymer. These fingerprints can be used for gas recognition. Regarding the structure of these polymers, except for PEG and PSu, the main chains are composed of long sp^3 carbons. The side chains of most of the polymers used in this study (e.g., PNVP, PVAc, and PSMA) contained

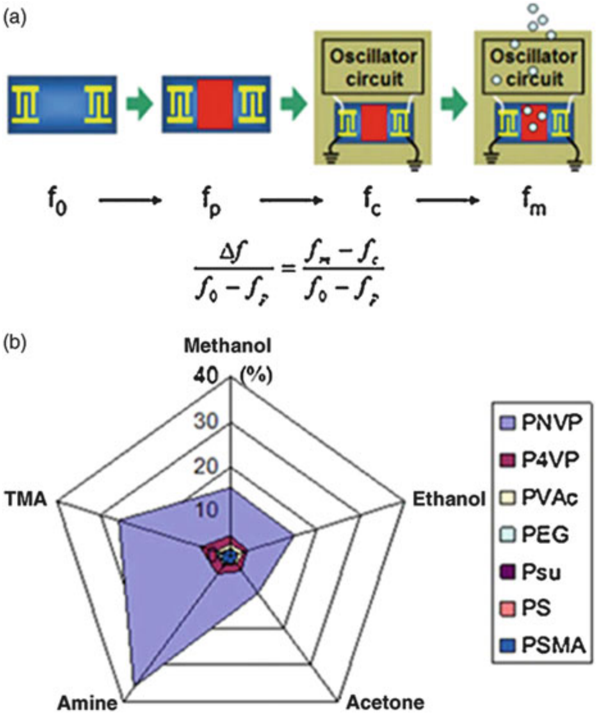


Fig. 7 The results of data analysis. (**a**) Schematic of the normalization method. (**b**) Radar plots of seven polymers to five gases

polarizable carbonyl groups; PSu contains polarizable sulfonyl groups in the main chain. The results in Fig. 7b show that the response to TMA and ammonia of polymers bearing carbonyl functional groups was stronger than that of polymers bearing other functional groups. PNVP, in particular, formed a more favorable sensing membrane than the other polymers and produced a higher frequency shift in sensing the two amine-type gases. These phenomena may be attributed to the strong interaction between the carbonyl groups and amines; these strong interactions caused the amount of amine adsorbed by sensors coated with polymers containing carbonyl groups to be higher.

3 Superior Nanocomposite Sensing Materials

Ordered mesoporous materials with high surface areas, large pore volumes, and uniform mesopores have attracted substantial attention [67–74]. Among such materials, ordered mesoporous carbons (OMCs) [75–94] exhibit high electric and thermal conductivities, low density, and high chemical as well as mechanical stability; these materials have been extensively applied in energy-related devices as sorbents for separation and storage and as catalyst supports. OMCs can be prepared by replicating established mesoporous "hard templates" [75–90] or by performing a self-assembly of carbon precursors with amphiphilic "soft templates" followed by condensation and carbonization processes [87–94].The OMC materials prepared using both templating methods have highly microporous frameworks [75–78]. In general, the hard template pathway provides a high degree of versatility for controlling the textural and morphological properties of OMCs. For example, CMK-3 is an OMC prepared through volume templating from a mesoporous silica SBA-15 and has a hexagonally ordered array of interconnected carbon nanorods [77]. The morphology of CMK-3 is determined according to that of SBA-15, as exemplified by the preparation of the rod-shaped CMK-3 material by using a rod-shaped SBA-15 [95]. Moreover, when surface templating is used instead of volume templating from SBA-15, CMK-5 composed of interconnected carbon nanopipes arranged in a hexagonal pattern can also be prepared [78–81]. SAW sensors are a type of gravimetric sensor that is highly promising for portable electronic noses because of their high sensitivity, low cost, small size [14, 15, 96, 97], and dense polymer films are usually used as a versatile sensing material [14, 15, 96]. During gas detection, analyte molecules are adsorbed on the surface of and then dissolved into a polymer film. The equilibrium concentration of an analyte in a polymer film, which corresponds to the maximum change in weight for gravimetric detection, is determined according to the solubility and partial pressure of the analyte [98]. However, as schematically shown in Fig. 8a, analyte molecules are adsorbed and desorbed much faster than they diffuse in a polymer film, and the adsorption–desorption equilibrium at the film interface is established much earlier than the equilibrium concentration is reached [98]. The situation deteriorates during the detection of low-concentration analyte gases; practically, the equilibrium concentration may never be reached within a limited detection time. This may result in a poor (e.g., parts per million level) gas concentration.

Polymer/OMC nanocomposites that serve as superior sensing materials for detecting gravimetric gas were demonstrated [99]. As shown in Fig. 8b, an OMC was used as a high-surface-area, lightweight, and mechanically stable scaffold for directly growing polymers [100] that form interpenetrating and inseparable composite frameworks with carbon. With this nanoscale design, mesoporous nanocomposites enable the diffusion of analyte molecules into the materials to interact directly with all polymer molecules in them. Therefore, nanocomposites may exhibit superior sensitivity and reversibility in gas detection compared with dense polymer films. The current study developed an approach for preparing platelet-shaped CMK-5-like OMCs that contain short carbon nanopipes to facilitate gas diffusion and adsorption further. The zirconium(IV) ion-assisted method for synthesizing platelet-shaped SBA-15 reported by Chen et al. [101] was improved to prepare a thinner and less aggregated platelet-shaped mesoporous silica (PMS) hard template. Without further incorporation of other acid catalysts [75, 77, 78], Zr^{4+} ions were maintained in PMS as acid catalysts to produce a uniformly distributed polymeric carbon precursor that transformed into a platelet OMC (PMC) after

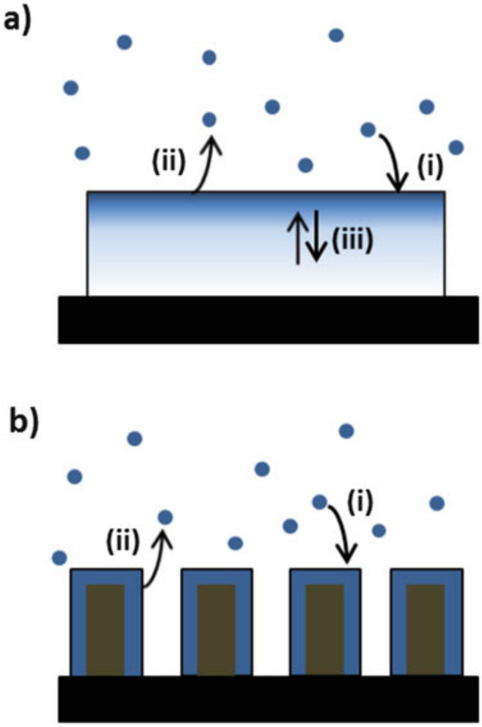


Fig. 8 Schematic comparison of the gas sensing processes with (**a**) a dense polymer film and (**b**) a polymer/OMC nanocomposite, involving the (i) adsorption and (ii) desorption of gas molecules at polymer surface, and (iii) the molecular diffusion in polymer. The *blue color* scale represents the concentration of the gas molecules in polymer

a pyrolysis process. Several polymer/PMC nanocomposites were prepared, and the SAW sensors deposited with the nanocomposites showed high sensitivity and reversibility in detecting ammonia gas at a parts-per-million level. Measuring NH_3 levels is of considerable interest in environmental monitoring and process control fields because of the high toxicity of the gas, and the exposure limits set by the U.S. National Institute for Occupational Safety and Health are 25 ppm over an 8-h period and 35 ppm over a 10-min period [102].

3.1 Material Synthesis and Characterization

PMS materials were synthesized by adding tetraethoxysilane (TEOS) or tetramethoxysilane (TMOS) into a HCl solution containing the triblock copolymer Pluronic P123 and $ZrOCl_2 \cdot 8H_2O$ [101]. The reaction mixture was stirred at 35 °C for 0.25–30 min before it was aged for 24 h at 90 °C. The solid was filtered, dried, and finally calcined at 540 °C. Furfuryl alcohol (FA) was used as a carbon source for preparing PMC [86, 87, 89]. Furthermore, 0.5 g of PMS was impregnated with 0.9 mL of FA and was subsequently heated at 105 °C for 2 h. Next, the mixture was heated at 900 °C in a nitrogen atmosphere for 3 h. The silica template was dissolved using a diluted solution of HF. The polymer/PMC nanocomposites were prepared through radical polymerization [100]. PMC was impregnated with a chloroform solution that comprised a mixture of a vinyl monomer, divinylbenzene, which served as a cross-linker, and 2,2′-azobis(isobutyronitrile), which functioned as a radial initiator. The polymerization reaction was initiated by heating the sample at 120 °C in an argon atmosphere. The resultant nanocomposite was washed extensively with chloroform and ethanol, and was subsequently dried under ambient conditions. The monomers used include N-vinylpyrrolidone, 4-vinylpyridine, styrene, and 4-tertbutoxystyrene; the resulting nanocomposites contained poly(N-vinylpyrrolidone) (PNVP/PMC), poly(4-vinylpyridine) (P4VP/PMC), polystyrene (PS/PMC), and poly(4-tertbutoxystyrene) (P4BS/PMC). Transmission electron microscope images were captured using a JEOL JEM-2010 microscope operated at 200 kV and equipped with an energy dispersion spectrometer. Scanning electron microscope images were obtained using a JEOL JSM-6330F microscope. X-ray diffraction (XRD) patterns were recorded using a Mac Science 18-MPX diffractometer and Cu Kα radiation. Nitrogen physisorption isotherms were measured at 77 K by using a Quantachrome Autosorb-1MP instrument. The isotherms were analyzed by using the nonlocal density functional theory method [103–105] and a model of nitrogen adsorbed on silica with cylindrical pores in which the adsorption branch is considered to evaluate pore sizes of the samples [106]. The Brunauer–Emmett–Teller surface areas were calculated from the adsorption branches in a relative pressure range of 0.05–0.20, and the total pore volumes were evaluated at a relative pressure of 0.95. Inductively coupled plasma mass spectroscopy data were obtained using a Perkin-Elmer SCIEX-ELAN 5000 device. Thermogravimetric analysis (TGA) was performed using a Linseis STA PT1600 analyzer.

Sensing materials were deposited on the $LiNbO_3$-based SAW devices, each of which was integrated with an IDT and signal readout system [81], by using a spin-coating process and ethanol as a solvent. Each SAW device exhibited a central frequency of approximately 116.2 MHz; after the coating process, the amount of sensing material deposited on each device was monitored according to the SAW frequency shift (Δf_s). Gas sensing measurements were performed in a 1-L chamber at 24 °C and a RH of 21 %. Laboratory air was first passed over the SAW sensor for 30 min to ensure stable SAW frequency signals. A stream of NH_3 gas with a calibrated concentration was subsequently passed through the sensor for 100–200 s; next, air was allowed to flow for 120–300 s to enable the sensor to recover to its initial frequency value. For each measurement, these procedures were repeated 3–5 times. The flow rate was maintained at 1 L min^{-1} during the measurements.

3.2 Analysis of the Mesoporous Materials

To improve the reported synthesis procedure [101] and, thus, fabricate thinner and less aggregated SBA-15 platelets, the silica precursor and stirring duration (for

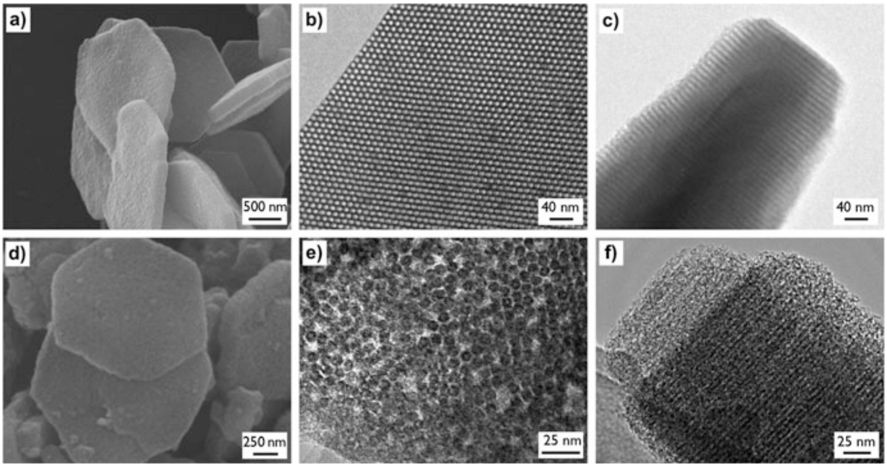
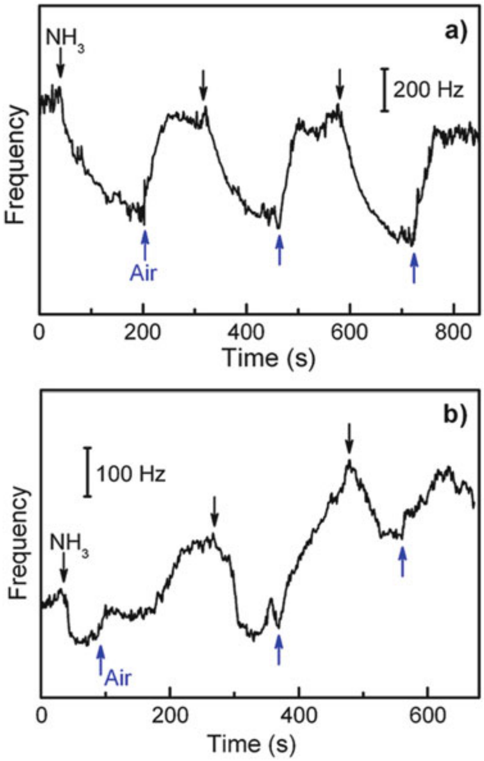


Fig. 9 SEM (**a**, **d**) and TEM (**b**, **c**, **e**, **f**) images of PMS (**a**–**c**) and PMC (**d**–**f**). The images in (**b**) and (**e**) are viewed along the channel axis, and those in (**c**) and (**f**) are taken perpendicular to the channel axis

mixing all reactants) were varied. When TEOS served as a silica precursor and a stirring time of 30 min was used [101], SBA-15 platelets that were approximately 0.5 μm thick and 1.0 μm wide were obtained; most of these platelets were aggregated and intergrown. When TMOS was used instead of TEOS, the hexagonal platelets became thinner (approximately 0.2 μm) and wider (nearly 2.0 μm). When the stirring time was shortened, the amount of aggregated and intergrown platelets decreased. As illustrated in Fig. 9a, under optimal conditions (with TMOS serving as a silica precursor and a stirring time of 1 min), the PMS sample mainly contained thin and separate platelets. These results can be understood by considering the formation mechanism of SBA-15 platelets. In the presence of Zr^{4+} ions, the faster hydrolysis of TMOS relative to that of TEOS may result in quicker silicate condensation around the ends of the copolymer micellar rods, leading to earlier termination of the growth of platelet SBA-15. Furthermore, when the stirring time is shortened while TMOS is allowed to be hydrolyzed and dissolved in the synthesis solution, the platelets may be grown under a static condition to prevent interconnection between platelets.

Controlled amounts of selected vinyl monomers, divinylbenzene and 2,2′-azobis(isobutyronitrile), were impregnated into PMC; the monomers were subsequently polymerized through a heating process to form polymer/PMC nanocomposites [100]. The nanocomposites had polymer loadings (determined according to TGA) of approximately 38–42 wt%; furthermore, these nanocomposites exhibited large surface areas (940–1,080 $m^2\ g^{-1}$), large pore volumes (0.86–0.98 $cm^3\ g^{-1}$), and relatively uniform mesoporosity. The pore sizes of these nanocomposites, compared with those of PMC, decreased by approximately 0.3–0.4 nm. To detect NH_3 gas at a parts-per-million level, the nanocomposites were deposited on the

Fig. 10 Response curves of the SAW sensors deposited with PNVP/PMC (**a**) or PNVP (**b**) to 16 ppm NH_3



integrated SAW sensors [102]. Figure 10a shows the response curves of the SAW sensor on which PNVP/PMC was deposited through three successive exposures to 16 ppm of NH_3 at 24 °C and a RH of 21 %. The SAW frequency decreased immediately when it was exposed to the analyte, and a frequency shift (Δf) of approximately 445 Hz was observed after 180 s. After the first exposure, the analyte was replaced with air for approximately 150 s, and the frequency increased to a level that was roughly maintained in subsequent recoveries from successive exposures to NH_3. The frequency shift was highly reproducible, and a standard deviation of approximately 44 Hz was discovered for successive exposures. The response was attributed to the interactions between NH_3 and PNVP in PNVP/PMC; in addition, PMC alone could not adsorb the analyte to produce detectable Δf, and this was confirmed in a control experiment. The relatively rapid and reproducible response to such a dilute NH_3 suggests that the polymer molecules in PNVP/PMC interacted directly with NH_3 molecules to establish the adsorption–desorption equilibrium rapidly. For comparison, an SAW sensor coated with PNVP was fabricated through a spin-coating process. As illustrated in Fig. 10b, the sensor exhibited an unstable response to 16 ppm of NH_3. Three successive exposures resulted in measurable but divergent values of Δf (approximately 100–210 Hz), and a slow but steady increase in SAW frequency was observed. Although increasing the amount of PNVP

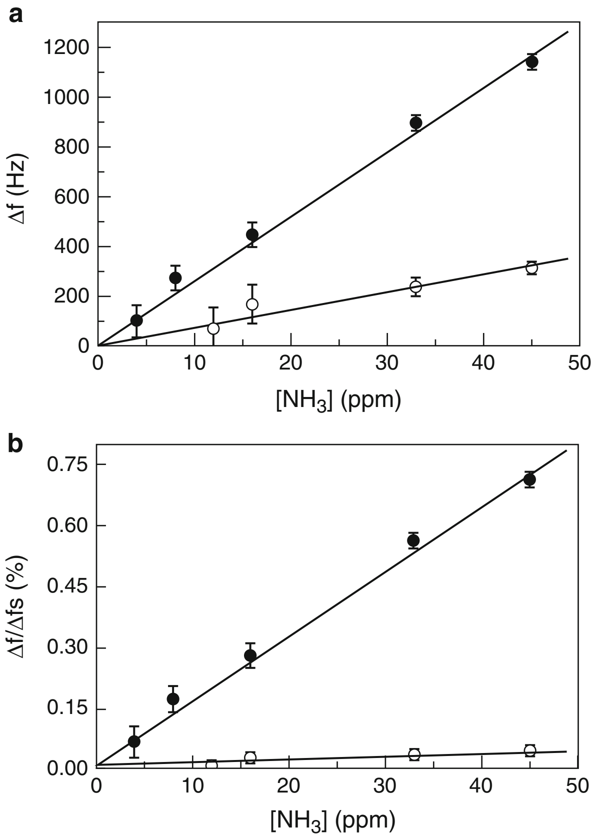


Fig. 11 Frequency shifts (**a**) and $\Delta f/\Delta f_s$ values (**b**) for the SAW sensors deposited with PNVP/PMC (*filled circle*) or PNVP (*unfilled circle*) to ppm-level NH_3. Error bars are reported as one standard deviation of the response averaged over eight exposures

coated on the sensor could not improve the stability and sensitivity of the sensor, reducing the coating amount deteriorated the signal-to-noise ratio of the signals. Obviously, for detecting gravimetric gas, the polymer/PMC nanocomposites showed more favorable sensitivity and reversibility than dense polymer films did.

Figure 11a shows the average response of SAW sensors on which PNVP/PMC or PNVP was deposited at different concentrations of NH_3 at the same temperature and RH; the results of a simple linear regression analysis are also shown in this figure. The PNVP/PMC-deposited sensor exhibited a linear response to NH_3 at a parts-per-million level with a high sensitivity value (defined as the slope of the regression line, $\Delta f/[NH_3]$) of 26.2 Hz/ppm. The detection limit of the sensor was lower than 5 ppm; this concentration is much lower than the exposure limits [102]. However, the PNVP-coated sensor exhibited a detection limit of approximately 12 ppm, and it showed much lower frequency shifts (corresponding to a sensitivity value of 7.1 Hz/ppm) and significantly higher standard deviations (77–85 Hz) in its response during eight exposures to the analyte with a concentration lower than 30 ppm. Furthermore, the amount of PNVP/PMC deposited was considerably lower than that of PNVP, and the corresponding frequency shifts (Δf_s) for PNVP/PMC and PNVP were 160 and 750 kHz, respectively. Therefore, the values for the two sensing materials normalized against Δf_s ($\Delta f/\Delta f_s$) showed a greater difference; the slope of the regression line for PNVP/PMC was nearly 17 times higher than that for the PNVP film (cf. Fig. 11b). The amount of polymer in the nanocomposite was approximately 40 wt%, and the frequency response per unit weight of PNVP for the SAW sensor deposited with PNVP/PMC became approximately 42 times higher than that for the PNVP-coated sensor.

The poor performance of the PNVP-coated SAW sensor in detecting NH_3 at parts-per-million levels may be attributed to the aforementioned sensing process; the poor performance could also be attributed to the interference caused by the adsorption of water by the hydrophilic polymer. Such interference was not obviously observed for PNVP/PMC, and this is probably because of the hydrophobic nature of the carbon scaffold [77–79]. Ammonia sensing at an RH of 44 % was performed, and although the PNVP-coated SAW sensor exhibited a higher standard deviation and an unfavorable detection limit of approximately 24 ppm, the PNVP/PMC-deposited sensor exhibited a nearly identical sensitivity and reversibility. Additional studies on the cross-sensitivity toward water are in progress. Finally, an array of SAW sensors on which four nanocomposites were deposited was used to detect NH_3 under the same conditions (24 °C, RH of 21 %). All sensors were more sensitive and exhibited more stable, reversible, and reproducible responses than the sensors coated with dense polymer films. Figure 12 shows the response pattern of the SAW sensor array to NH_3 at 16 ppm. All four sensors exhibited detectable but different frequency shifts to this dilute analyte, generating a characteristic fingerprint for NH3. The differences in response among the four nanocomposites may be correlated to the functional groups of the polymers. PS/PMC and P4BS/PMC produced weak responses mainly because of the lack of substantial interactions between the polar analyte molecules and hydrophobic PS and P4BS polymers. PNVP/PMC showed the largest frequency shifts among the four nanocomposites, indicating that the pyrrolidone moiety in PNVP may interact relatively strongly with NH_3 through the amide oxygen. For P4VP/PMC, the adsorption interaction between the pyridine group in P4VP and NH_3 was weaker than that for PNVP. These results imply that polymer/OMC nanocomposites can be used as superior sensing materials for highly sensitive and reversible gravimetric sensor arrays in electronic nose applications.

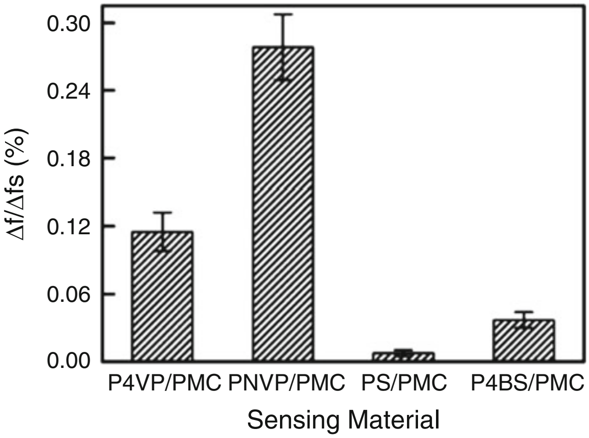


Fig. 12 Response pattern of 16 ppm NH3 generated by the SAW sensor array with four nanocomposite sensing materials

4 Sensor Signal Readout Electronics and Data Cluster Methods

To develop a gas detection system, readout electronics for replacing instruments such as frequency counters were designed and implemented. Numerous types of circuit implementations are used for detecting SAW frequencies and phases. One technique involves converting the SAW frequency into a voltage signal. This approach wastes a substantial portion of the voltage range because the frequency shift of the sensor does not exceed a certain amount [107]. The center frequency was nearly 100 MHz, and the maximum frequency shift did not exceed 1 MHz, representing less than 1 % of the center frequency. Therefore, when a direct frequency-to-voltage converter was adopted, under the assumption that the conversion from frequency to voltage is linear, approximately 99 % of the voltage range was wasted. Another approach involves measuring phase differences; however, the resolution is at only a megahertz level. The computing power requirements for portable gas detection systems vary according to the complexity of the application. In particular, programming-embedded-pattern-recognition software that is suitable for the application is the key to determine the computing power requirements. A training or database establishment procedure similar to the procedure executed in mammalian brains must be completed in the odor recognition algorithm, and these procedures represent the most complex phase in the algorithm.

4.1 Frequency Readout Electronics

Readout electronics for replacing instruments such as frequency counters were designed and implemented, and Fig. 13 shows a block diagram of the implemented readout electronics. Counter 1 is the main element that calculates the SAW output

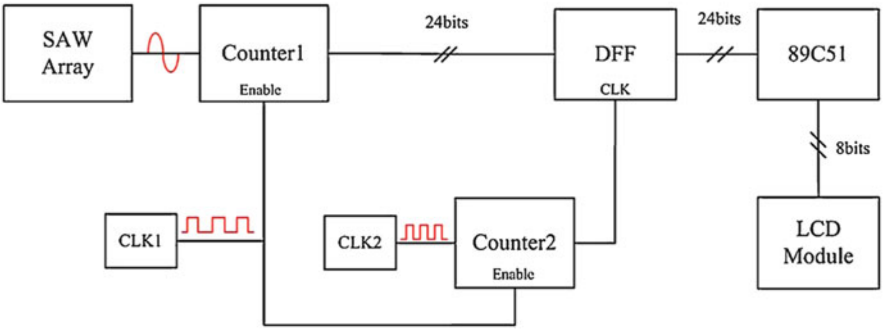


Fig. 13 Block diagram of the readout electronics

frequency. Because of the stabilization time of the SAW sensor, the sampling starting time was set to 0.5 s after the device was switched on. Theoretically, the accuracy of the calculated frequency increases with the increase in the bits of Counter 1. However, in consideration of the tradeoff between sampling time and circuit complexity, the number of bits in Counter 1 was set to 24, and the total sampling time was 0.12 s; the data in Counter 1 are collected by a 89C51 microprocessor. System Clock 1 (CLK1) activates Counter 1 and Counter 2 and simultaneously resets the previous data. System Clock 2 (CLK2) controls Counter 2 to trigger a D flip-flop (DFF) after Counter 1 begins operating for a fixed period of time. CLK1 and CLK2 are generated by connecting the counters with oscillators with fixed frequencies. The 89C51 microprocessor collects data from the DFF and then calculates the real SAW sensor frequency and stores it in its built-in memory. A reference sensor frequency is subtracted from these frequencies subsequently to obtain the frequency shift of the sensor; the result is then displayed on the LCD module. The LCD display is designed such that its first row shows the real-time sensor number and frequency, and the second row shows the frequency difference between a sensor and the reference sensor.

The readout electronics were implemented on a PCB and the 89C51 microprocessor board. The top level comprises the 89C51 microprocessor board and an LCD display module. The lower level comprises the PCB of the counters, CLKs, and DFF. To simulate the SAW signal, the circuit was first tested using a fixed sinusoidal signal generated by a function generator. This signal served as the input of Counter 1. The frequency data can be obtained by reading the LCD display. These data were used to analyze the accuracy of the circuit. The signal frequency was swept from 98 to 99 MHz in increments of 10 kHz, and the result was obtained from the LCD display. The frequency error (LCD displayed frequency—input frequency) was within ±20 Hz, which translated to $\pm 2 \times 10^{-5}$ % according to a 100-MHz baseline frequency. The source of error was the phase difference between CLK1 and CLK2. Because the two timing clocks are different, every measurement had a time difference of approximately 1 ns. This difference can be reduced by evaluating the

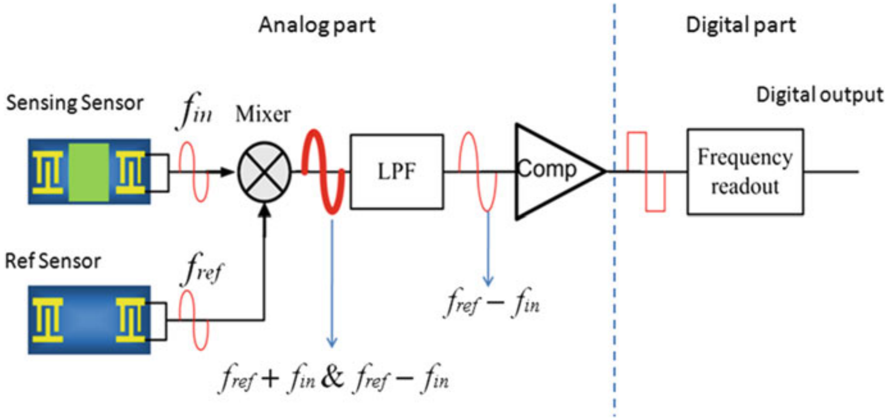


Fig. 14 Block diagram of the mixed-signal interface chip

frequency difference between an SAW sensor (signal) and a reference SAW sensor. For the next step, the function generator was replaced by a real SAW sensor and an operational amplifier.

A mixed-signal interface chip was implemented as shown in Fig. 14 [30]. The first stage of the chip is analog and contains a mixer, low-pass filter, and comparator. The analog stage modulates, filters, and converts the signal provided by the sensor into a square wave. The second stage of the chip comprises a digital readout for detecting changes in the sensor frequency. To read out the sensor signal (frequency change), the frequency of the sensor *fin* was subtracted from that of a reference sensor *fref*. This resulted in a decrease in the frequency change *fref* − *fin* compared with *fref*. Because the amplitude of the signal that must be converted was lower, the resolution could be improved and the power consumption could be reduced. The reference sensor was not coated with a polymer membrane; therefore, the variations in its frequency caused by the input gas were extremely low. Nevertheless, the reference sensor was still influenced by the same environmental parameters, such as temperature and humidity, as the sensing sensor was. Subtracting *fin* from *fref* ensured that the background effect of the sensor was eliminated.

A mixer was integrated into the front end of the analog section; the output of this mixer has a high-frequency term $f_{ref} + f_{in}$ and low-frequency term $f_{ref} - f_{in}$. After the mixer, a low-pass filter passes the low-frequency term (*fref* − *fin*) to a comparator to generate a square wave output to the digital stage. A digital frequency readout circuit receives the output from the comparator and reports the frequency data. The main idea was to use three counters with DFFs for storage (Fig. 15). Counter 2 provides a fixed sampling time $\text{TS} = 217 \times \text{T}_{\text{CLK2}}$, where T_{CLK2} represents the clock period of CLK2. The input signal (F_{in}) is transmitted into Counter 3 as the clock. When Counter 3 counts for a time of 217 × TCLK2, the MSB of Counter 2 triggers the DFF and stores the current output data (D) of Counter 3. Counter 1 generates control

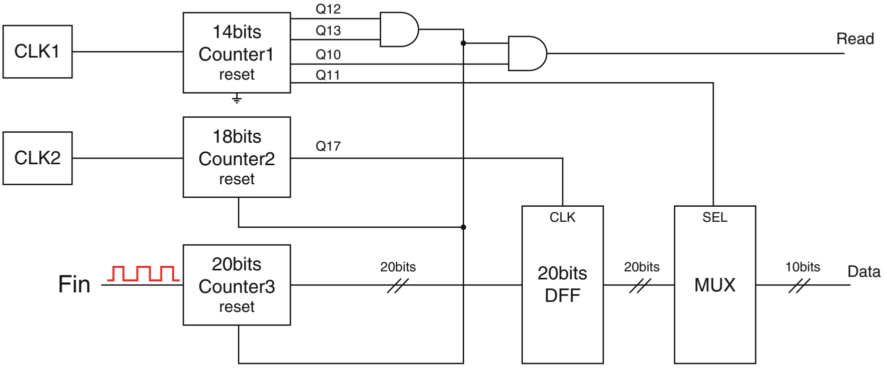


Fig. 15 Schematic of digital frequency readout circuit

signals for the multiplexor, reset signal of Counter 3, and Read signal processing units. The input signal (F_{in}) can be calculated as follows:

$$F_{in} = \frac{\mathrm{D}}{2^{17}} \times f_{CLK2} = \frac{D}{T_S} \tag{5}$$

According to Eq. (5), the resolution of the mixed-signal interface chip depends on the sampling time T_S. Theoretically, a finer resolution can be achieved by using a longer sampling time. For example, a 1-s sampling time corresponds to a resolution of 1 Hz (each bit in Counter 3 represents 1 Hz), and a sampling time of 100 ms corresponds to a resolution of 10 Hz (each bit in Counter 3 represents 10 Hz). Because each sensor was switched on for 1 s, a sampling time of 1 s would be too long. Moreover, a 10-Hz resolution was ideal for detection. Consequently, a sampling time of 100 ms was used.

Ethanol and acetone were used to test the SAW array. Figure 16 shows a typical response of the sensors (coated with PMSA, PEG, and PNVP) that were exposed to ethanol. The experiment was repeated three times. The baseline drifted because of the temperature variation. SAW devices are sensitive to environmental parameters, especially temperature [108]. Currently, numerous SAW devices are manufactured with materials such as $LiNbO_3$ or lithium tantalite. The advantage of using these two materials is their high K^2 values, and the disadvantage is their high temperature coefficient. In this study, the temperature effect was compensated using a reference SAW device without a membrane.

To verify the accuracy of the interface chip, a signal generator was used to emulate the SAW signals. A spectrum analyzer was used to measure the outputs of the mixer and low-pass filter. The two input signals to the mixer were 117- and 118-MHz sinusoidal waves with an amplitude of 0.5 V. Therefore, the spectrum analyzer indicated that the output spectrum of the mixer had two frequency components,

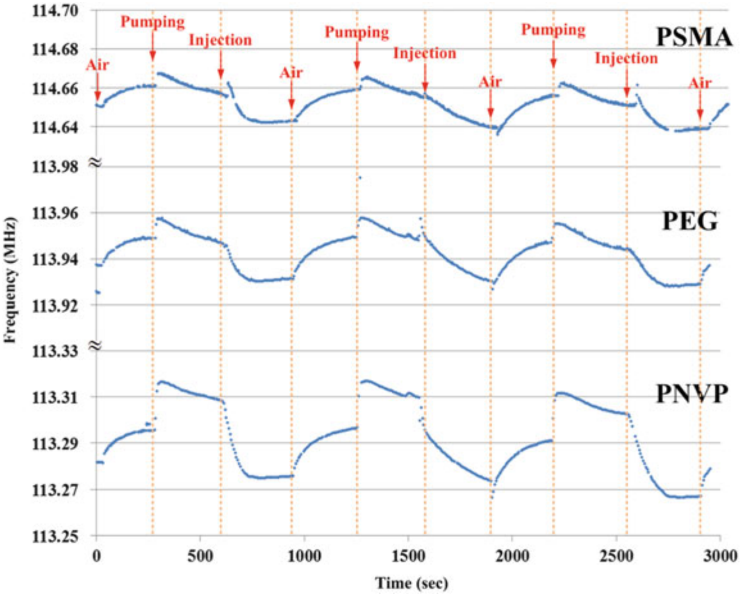


Fig. 16 Typical sensor response (gas: ethanol, membrane: PNVP)

1 and 235 MHz, and the low-pass filter output spectrum had one frequency component, 1 MHz. These results verified the function of the mixer and low-pass filter.

The output of the low-pass filter was sent to a comparator to convert the signal into a square wave, which was passed to the digital frequency readout circuit. To test the frequency readout circuit, a square wave was input into the circuit and a frequency counter simultaneously. The test frequencies were 10, 100, 1 kHz, 10 kHz, 100 kHz, and 1 MHz. Each frequency was sampled 1,000 times, and the total test time was 100 s. In all tests, the frequency outputs of the readout circuit were the same as those of the frequency counter. The power consumed by the entire chip was 1.48 mW; a power supply of 3.3 V was used. Figure 17 shows a photo of the die used in the low-power mixed-signal SAW interface ASIC. For comparison, Table 3 shows with the results of other studies of the SAW interface circuit [107, 109].

Figure 18 shows the setup of the gas test. The SAW array was connected to the interface chip, and the output was connected to a wireless module. Both the interface chip and wireless module were operated using two 1.5-V batteries. The wireless module transmitted the sensor data to a base station PC. The measured output was compared with the output of the frequency counter according to Eq. (6):

$$\sum_{i=1}^{n} A_i - B_i / n = error \tag{6}$$

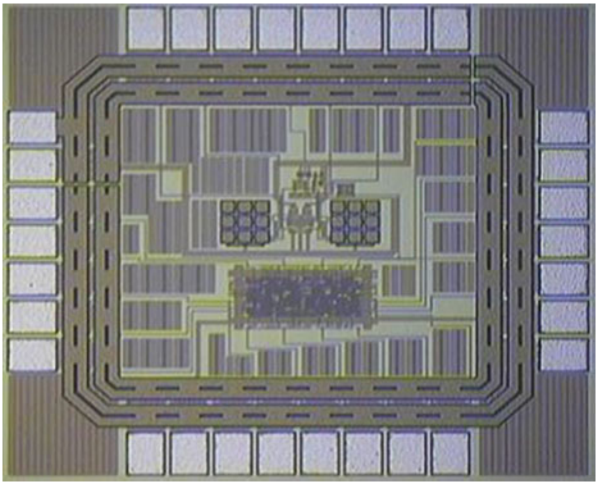


Fig. 17 Die photo of the low-power mixed-signal SAW interface ASIC

Table 3 Benchmark of the SAW interface circuit

	[108]	[110]	This work
Year	2005	2000	2010
Supply voltage	3.3 V	2.5 V	3.3 V
Process technology	0.35 μm	GaAs	0.18 μm
Power consumption	38.35 mW	225 mW	1.48 mW
Resolution	10 MHz	3 MHz	10 Hz
Input frequency	354 MHz	690 MHz	99.8 MHz

where Data List A represents the output of the frequency counter, Data List B denotes the data transmitted to the base station PC, and n is the data number used for comparison. Table 4 shows a summary of the mean error and standard deviation of the two data lists, indicating that both the mean error and standard deviation between the data transmitted from the sensor node and output of the frequency counter were less than 4 Hz.

4.2 Odor Analysis and Cluster

Although typical electronic noses require a PC to acquire and process the signals from a sensor array, the PC, which is used as a pattern recognition engine, can be replaced with a powerful microcontroller equipped with an embedded odor classification program. This reduces the volume and weight of a digital apparatus that is used as an individual smart portable electronic nose device. The computing power requirements for an embedded electronic nose system vary according to the complexity of the application. In particular,

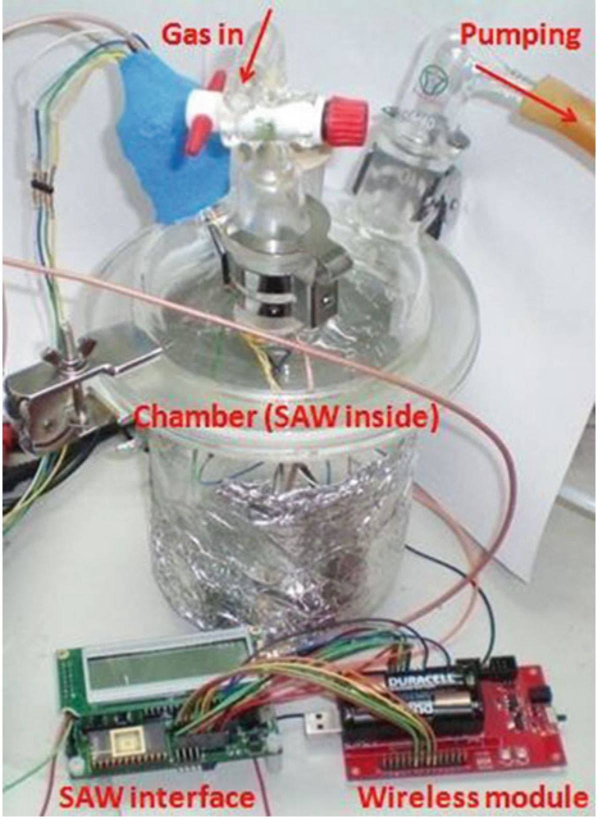


Fig. 18 Test setup of the sensor and its electronics

Table 4 Mean error and standard deviation of the two lists of data

	Gas			
	Ethanol		Acetone	
	Readout data (Hz)			
Membrane	Mean error	Standard deviation	Mean error	Standard deviation
PNVP	2.18	2.56	2.41	2.47
PS	2.47	3.24	1.24	1.89
PSMA	2.48	3.39	2.17	2.52
PEG	2.13	3.25	1.25	1.44
P4VP	2.31	3.54	0.74	0.94
PVAc	1.53	2.69	1.23	1.85
PSu	1.97	2.58	3.68	2.94

programming-embedded-pattern-recognition software that is suitable for the application is the key to determining the required computing power. A training or database establishment procedure similar to the procedure executed in mammalian brains must be completed in the odor recognition algorithm; these procedures represent the most complex phase in the algorithm. Furthermore, the microprocessor must be evaluated according to the power consumption, system operation frequency, data capacity, instrument size limitations, manufacturing cost, and compatibility with other electronic devices through connections such as Ethernet connections, buses, ports, and display interfaces. Table 5 shows the architectural options for designing an embedded electronic system arranged from simple to complex in form [110].

The sensing signals can be analyzed using a combined system of five classifiers, namely multilayer perceptron, Gaussian mixture models, radial basis functions, K-nearest neighbors, and probabilistic principal component analysis [111]. For example, a neural network algorithm can be applied to a sensor array for odor analysis and classification [7]. In [112], the performance of a portable electronic nose based on a sensor array was analyzed using principal component analysis, and the device exhibited 100 % accuracy in a probabilistic neural network.

Furthermore, clustering methods can be divided into two general classes, supervised and unsupervised clustering [40]. The algorithm used in this study sorts through all data to determine simultaneously the families of polymers and gases that exhibit similar behaviors in each experiment. Using a mathematical similarity relation is crucial [113, 114]. Data on reasonable similarity measures regarding the behavior of polymers or gases, such as the Euclidean distance, angle, or dot products of the two vectors representing a series of measurements, can be used. This study showed that the sensible Spearman's rank correlation coefficient (r_s) with average-linkage clustering conforms adequately to the chemical notion of the similarity of two polymers or gases. In this study, clustering analysis was performed using CLUSTER (http://rana.lbl.gov/EisenSoftware.htm) [115], a hierarchical clustering program obtained from the Eisen Laboratory at the University of California, Berkeley. The hierarchical clustering algorithm used is based closely on the average-linkage method of Sokal and Michener [116], which was developed for clustering correlation matrices such as those used in this study. The objective of this algorithm is to compute a dendrogram that assembles all elements into a single tree. For any set of polymers and gases, a similarity matrix was computed by using the previously described metric, which contains similarity scores for all values.

Before clustering, the data were normalized according to $\mathrm{f_{polymer}}$ $(\mathrm{f_0 - f_p})$. To assign polymers that were in the data sets to specific functional groups, information on the molecular functions and characteristics of each polymer provided in the chemical engineering databases were used. Differential frequencies of the SAW array as the target gas was absorbed by the coated polymers were normalized according to $\mathrm{f_{polymer}}$ $(\mathrm{f_0 - f_p})$ and clustered using the hierarchical clustering program CLUSTER. The clusters were then visualized using the TREEVIEW (http://rana.lbl.gov/EisenSoftware.htm) [115] program. This program identified polymers and gases, and the results were consistent with of analysis of seven polymers to five

Table 5 Mean error and standard deviation of the two lists of data [110]

Architecture	Typical configuration (bit/speed/RAM)	Pros/Cons	Typical programming language	Available processing
Sensor array + μC (PIC)	8 bit/10 MHz/k bytes	Easy, small, low power, portable, cheaper	ASM/C	Easy algorithms with few data, KNN, easy NN, mostly trained off-system, linear classifiers, quadratic classifiers
SA + high perf. MC	8–16 bit/10–33 MHz/k bytes	Small, low power, portable, cheap	ASM/C	Some small matrix manipulation available, linear (PCA, LDA, PCR), KNN, easy fuzzy interface systems
SA + μP or DSP	16–32 bit/20–100 MHz/Mb	Very fast, medium size, portable, huge power consumption	ASM/C/C++	Linear (PCA, LDA, PCR, PLS), KNN, easy neural and fuzzy system, standard feature extraction/selection (PCA, LDA)
SA + embedded PC	32 bit/80–233 MHz/Mb	Fast, medium size, portable, huge data capacity, high consume expensive	Any	Linear, complex learning algorithms (GA, NeuroFuzzy systems, mixture models, APR, FIS optimization algorithms), advanced feature extraction/selection (SFS, SFFS)
SA + desktop PC	32–64 bit/700 MHz/Mb	Fast, medium size, portable, huge data capacity, consume not critical, expensive, not portable	Any/visual	Linear, complex learning algorithms (GA, NeuroFuzzy systems, mixture models, FIS optimization algorithms), advanced feature extraction/selection (SFS, SFFS), etc.

gases in a cluster showing that such polymers and gases have similar functional groups or common elements, most of which are identifiable according to their chemical properties. The data on the polymer responses were classified and clustered

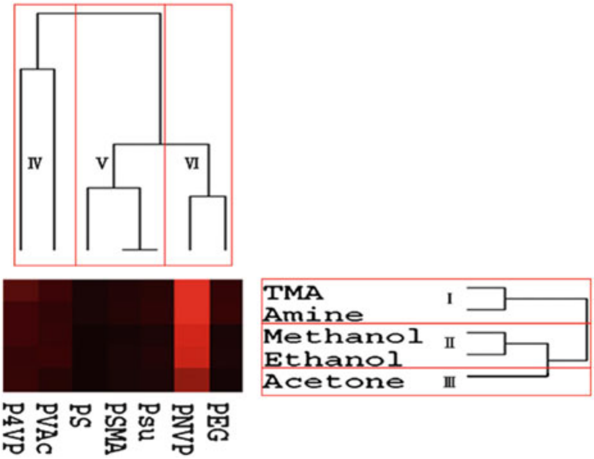


Fig. 19 Two-way hierarchical clustering: the algorithm grouped gases into three families (I, II, III), and polymers into three families (IV, V, VI). Polymers and gases are divided into a family conform to function group or chemical property. The intensity of the *color* indicates the magnitude of fgas/fpolymer (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article)

according to the type of gas, and Fig. 19 shows the two-way hierarchical clustering results [15]. The clustering was successful: ammonia and TMA were grouped as the amine family, whereas methanol, ethanol, and acetone were classified into the family of volatile nitrogen-free compounds. In the family of volatile nitrogen-free compounds, the two alcohols could be further identified and grouped according to acetone. In addition, despite the complex and diverse interactions between the polymers and target gases, the chemical structure and properties of the polymers were correlated with the responses to different gases. Both PNVP and PEG contain hydrophilic side chains, whereas P4VP, which contains acetate groups, and PVAc, which contains phenol groups, are less hydrophilic. PS, PSMA, and PSu contain benzene groups either in their side chains or main chains, making them more hydrophobic than other polymers. Thus, these clusters provide a foundation for understanding the mechanisms through which polymers adsorb a target gas.

5 Summary

This study presents the overall process flow involved in developing a portable gas detection system that comprises a packaged SAW device and readout electronics with an 89C51 microprocessor. The frequency and insertion loss of the SAW devices fabricated on PCB were measured using a network analyzer and determined to be −8.53 dB and 98.5 MHz, respectively. The operating mechanisms of vapor sensors with seven spin-coated polymer films responding to five vapors were studied. The results showed that the SAW devices coated with PNVP films exhibited higher sensitivity than those coated with other films. To increase the sensitivity in gravimetric gas detection, nanocomposites of polymers and OMCs were used. The platelet-shaped CMK-5-like carbon was prepared using zirconium containing SBA-15 platelets as a hard template; the polymers were directly grown on the carbon material through a radical polymerization process, thus forming interpenetrating and inseparable composite frameworks with carbon. The platelet nanocomposites exhibited high surface areas, large pore volumes, and relatively uniform mesoporosity. The two-way hierarchical clustering was instrumental in extracting information regarding vapor–coating interactions from the multidimensional data set. The results clearly indicated the ability of the sensors to recognize distinct components of gases belonging to similar families. In addition, this study clearly demonstrated that combining multiple coated SAW devices with an appropriate recognition algorithm provides a sensing system that can be selective.

Acknowledgements The authors would like to acknowledge the support of the National Science Council of Taiwan, under Contract No. NSC 101-2220-E-007-006, NSC 102-2220-E-007-006, and Ministry of Science and Technology, under Contract No. MOST 103-2220-E-007-023. We also acknowledge the support of the Chung-Shan Institute of Science and Technology for technical consulting.

References

1. Sun Y, Ong KY. Detection technologies for chemical warfare agents and toxic vapors. Boca Raton: CRC; 2004.
2. Yinon J. Field detection and monitoring of explosives. Trends Anal Chem (TrAC). 2002;21(4):292–301.
3. Cao W, Duan Y. Breath analysis: potential for clinical diagnosis and exposure assessment. Clin Chem. 2006;52(5):800–11.
4. Yamazoe N, Sakai G, Shimanoe K. Oxide semiconductor gas sensors. Catal Surv Jpn. 2003;7(1):63–75.
5. Panpan F, Zhixue Y. Metal oxide semiconductor gas sensor. Chem Ind Times. 2013;3:022.
6. Miasik JJ, Hooper A, Tofield BC. Conducting polymer gas sensors. J Chem Soc Faraday Trans 1. 1986;82(4):1117–26.
7. Bai H, Shi G. Gas sensors based on conducting polymers. Sensors. 2007;7(3):267–307.
8. Ding B, et al. Electrospun nanofibrous membranes coated quartz crystal microbalance as gas sensor for NH_3 detection. Sensors Actuators B Chem. 2004;101(3):373–80.
9. Koshets IA, et al. Calixarene films as sensitive coatings for QCM-based gas sensors. Sensors Actuators B Chem. 2005;106(1):177–81.
10. Grate JW, Patrash SJ, Abraham MH. Method for estimating polymer-coated acoustic wave vapor sensor responses. Anal Chem. 1995;67(13):2162–9.
11. Grate JW, et al. Determination of partition coefficients from surface acoustic wave vapor sensor responses and correlation with gas-liquid chromatographic partition coefficients. Anal Chem. 1988;60(9):869–75.
12. Ando I, Furuki M, Pu LS. Optical gas sensor. US patent No. 5,030,009. 1991 July 9.
13. Sharpe SW, et al. Gas-phase databases for quantitative infrared spectroscopy. Appl Spectrosc. 2004;58(12):1452–61.

14. Gardner JW, Varadan VK, Awadelkarim OO. Microsensors, MEMS, and smart devices, vol. 1. Chichester: Wiley; 2001.
15. Hao HC, et al. Development of a portable electronic nose based on chemical surface acoustic wave array with multiplexed oscillator and readout electronics. Sensors Actuators B Chem. 2010;146(2):545–53.
16. Dai E, et al. Organic vapor sensors based on SAW resonator and organic films. IEEE Trans Ultrason Ferroelectr Freq Control. 1997;44(2):309–14.
17. Penza M, Vasanelli L. SAW NOx gas sensor using WO_3 thin-film sensitive coating. Sensors Actuators B Chem. 1997;41(1):31–6.
18. Wells M, Crooks RM. Interactions between organized, surface-confined monolayers and vapor-phase probe molecules. 10. Preparation and properties of chemically sensitive dendrimer surfaces. J Am Chem Soc. 1996;118(16):3988–9.
19. Wang C, He X-W, Chen L-X. A piezoelectric quartz crystal sensor array self assembled calixarene bilayers for detection of volatile organic amine in gas. Talanta. 2002;57(6):1181–8.
20. Dermody DL, Crooks RM, Kim T. Interactions between organized, surface-confined monolayers and vapor-phase probe molecules. 11. Synthesis, characterization, and chemical sensitivity of self-assembled polydiacetylene/calix [n] arene bilayers. J Am Chem Soc. 1996;118(47):11912–7.
21. Sankaran S, Khot LR, Panigrahi S. Biology and applications of olfactory sensing system: a review. Sensors Actuators B Chem. 2012;171:1–17.
22. Gopel W, Weiss T. Design for smelling. IEEE Spectr. 1998;35(9):32–4.
23. Craven MA, Gardner JW, Bartlett PN. Electronic noses—development and future prospects. Trends Anal Chem (TrAC). 1996;15(9):486–93.
24. Gardner JW, Bartlett PN. A brief history of electronic noses. Sensors Actuators B Chem. 1994;18(1):210–1.
25. McCartney W. Olfaction and odours. An osphresiological essay. Berlin: Springer; 1968.
26. Lammerink TSJ, et al. Intelligent gas-mixture flow sensor. Sensors Actuators A Phys. 1995;47(1):380–4.
27. Firestein S. How the olfactory system makes sense of scents. Nature. 2001;413(6852):211–8.
28. Breer H. Olfactory receptors: molecular basis for recognition and discrimination of odors. Anal Bioanal Chem. 2003;377(3):427–33.
29. Röck F, Barsan N, Weimar U. Electronic nose: current status and future trends. Chem Rev. 2008;108(2):705–25.
30. Tang K-T, Li C-H, Chiu S-W. An electronic-nose sensor node based on a polymer-coated surface acoustic wave array for wireless sensor network applications. Sensors. 2011;11(5):4609–21.
31. Lozano J, Santos JP, Horrillo MC. Enrichment sampling methods for wine discrimination with gas sensors. J Food Compos Anal. 2008;21(8):716–23.
32. Lozano J, et al. Electronic nose for wine ageing detection. Sensors Actuators B Chem. 2008;133(1):180–6.
33. Wu T-T, Chen Y-Y, Chou T-H. A high sensitivity nanomaterial based SAW humidity sensor. J Phys D Appl Phys. 2008;41(8):085101.
34. Ricco AJ, Martin SJ, Zipperian TE. Surface acoustic wave gas sensor based on film conductivity changes. Sensors Actuators. 1985;8(4):319–33.
35. Huang F-C, Chen Y-Y, Wu T-T. A room temperature surface acoustic wave hydrogen sensor with Pt coated ZnO nanorods. Nanotechnology. 2009;20(6):065501.
36. Gardner JW, et al. An electronic nose system for monitoring the quality of potable water. Sensors Actuators B Chem. 2000;69(3):336–41.
37. Shin HW, et al. Classification of the strain and growth phase of cyanobacteria in potable water using an electronic nose system. IEE Proc Sci Meas Technol. 2000;147(4):158–64.
38. Yao D-J, et al. A biochemical sensing system using an 11-MUA/calix [6] arene bilayer to sense amine vapors. J Micromech Microeng. 2007;17(8):1435.
39. Cai Q, et al. Kinetic assay of antitrypsin in human serum by a surface acoustic wave (SAW)-impedance sensor. Fresenius J Anal Chem. 1996;356(1):96–7.

40. Casalinuovo IA, et al. Application of electronic noses for disease diagnosis and food spoilage detection. Sensors. 2006;6(11):1428–39.
41. Shen C-Y, Huang C-P, Huang W-T. Gas-detecting properties of surface acoustic wave ammonia sensors. Sensors Actuators B Chem. 2004;101(1):1–7.
42. Groves WA, Zellers ET, Frye GC. Analyzing organic vapors in exhaled breath using a surface acoustic wave sensor array with preconcentration: selection and characterization of the preconcentrator adsorbent. Anal Chim Acta. 1998;371(2):131–43.
43. D'amico A, Verona E. SAW sensors. Sensors Actuators. 1989;17(1):55–66.
44. Andle JC, Vetelino JF. Acoustic wave biosensors. Sensors Actuators A Phys. 1994;44(3): 167–76.
45. Ballantine Jr DS, et al. Acoustic wave sensors: theory, design, & physico-chemical applications. San Diego: Academic; 1996.
46. Vellekoop MJ, et al. Integrated-circuit-compatible design and technology of acoustic-wave-based microsensors. Sensors Actuators A Phys. 1994;44(3):249–63.
47. Gizeli E, et al. A love plate biosensor utilising a polymer layer. Sensors Actuators B Chem. 1992;6(1):131–7.
48. Freudenberg J, et al. A contactless surface acoustic wave biosensor. Biosens Bioelectron. 1999;14(4):423–5.
49. Länge K, Rapp BE, Rapp M. Surface acoustic wave biosensors: a review. Anal Bioanal Chem. 2008;391(5):1509–19.
50. Ferreira GN, da-Silva A-C, Tomé B. Acoustic wave biosensors: physical models and biological applications of quartz crystal microbalance. Trends Biotechnol. 2009;27(12): 689–97.
51. Bender F, et al. Improvement of surface acoustic wave gas and biosensor response characteristics using a capacitive coupling technique. Anal Chem. 2004;76(13):3837–40.
52. Huang Y-S, Chen Y-Y, Wu T-T. A passive wireless hydrogen surface acoustic wave sensor based on Pt-coated ZnO nanorods. Nanotechnology. 2010;21(9):095503.
53. Greve DW, et al. Surface acoustic wave devices for harsh environment wireless sensing. Sensors. 2013;13(6):6910–35.
54. Enguang D, et al. Organic vapor sensors based on SAW resonator and organic films. IEEE Trans Ultrason Ferroelectr Freq Control. 1997;44(2):309–14.
55. Drafts B. Acoustic wave technology sensors. IEEE Trans Microwave Theory Tech. 2001;49(4):795–802.
56. Caliendo C, et al. Surface acoustic wave humidity sensor. Sensors Actuators B Chem. 1993;16(1):288–92.
57. Zhou R, et al. Phthalocyanines as sensitive materials for chemical sensors. Appl Organomet Chem. 1996;10(8):557–77.
58. Zellers ET, White RM, Wenzel SW. Computer modelling of polymer-coated ZnO/Si surface-acoustic-wave and lamb-wave chemical sensors. Sensors Actuators. 1988;14(1):35–45.
59. Tseng C-C. Elastic surface waves on free surface and metallized surface of CdS, ZnO, and PZT-4. J Appl Phys. 1967;38(11):4281–4.
60. Bohrer FI, et al. Comparative gas sensing in cobalt, nickel, copper, zinc, and metal-free phthalocyanine chemiresistors. J Am Chem Soc. 2008;131(2):478–85.
61. Li P, Li Y, Yang M. Hyperbranched polycarboxylates and their nanocomposites with ZnO: investigations on the humidity-sensitive properties. J Appl Polym Sci. 2011;120(4):1994–2000.
62. Kukushkin IV, et al. Ultrahigh-frequency surface acoustic waves for finite wave-vector spectroscopy of two-dimensional electrons. Appl Phys Lett. 2004;85(19):4526–8.
63. Fu YQ, et al. Microfluidics based on ZnO/nanocrystalline diamond surface acoustic wave devices. Biomicrofluidics. 2012;6(2):024105.
64. Morgan DR. Surface acoustic wave devices and applications: 1. Introductory review. Ultrasonics. 1973;11(3):121–31.

65. Penza M, Milella E, Anisimkin VI. Gas sensing properties of Langmuir-Blodgett polypyrrole film investigated by surface acoustic waves. IEEE Trans Ultrason Ferroelectr Freq Control. 1998;45(5):1125–32.
66. Thompson M, Stone DC. Surface-launched acoustic wave sensors: chemical sensing and thin-film characterization. New York: Wiley; 1997.
67. Yanagisawa T, et al. The preparation of alkyltrimethylammonium-kanemite complexes and their conversion to microporous materials. Bull Chem Soc Jpn. 1990;63(4):988–92.
68. Kresge CT, et al. Ordered mesoporous molecular sieves synthesized by a liquid-crystal template mechanism. Nature. 1992;359(6397):710–2.
69. Horvath G, Kawazoe KJ. Generalized synthesis of periodic surfactant/inorganic composite materials. Chem Eng Jpn. 1983;16:470–7.
70. Zhao D, et al. Triblock copolymer syntheses of mesoporous silica with periodic 50 to 300 angstrom pores. Science. 1998;279(5350):548–52.
71. Ciesla U, Schüth F. Ordered mesoporous materials. Microporous Mesoporous Mater. 1999;27(2):131–49.
72. Soler-Illia GJ, Sanchez C, Lebeau B, Patarin J, et al. Chemical strategies to design textured materials: from microporous and mesoporous oxides to nanonetworks and hierarchical structures. Chem Rev. 2002;102(11):4093–138.
73. Boettcher SW, et al. Harnessing the sol–gel process for the assembly of non-silicate mesostructured oxide materials. Acc Chem Res. 2007;40(9):784–92.
74. Trewyn BG, et al. Synthesis and functionalization of a mesoporous silica nanoparticle based on the sol–gel process and applications in controlled release. Acc Chem Res. 2007;40(9): 846–53.
75. Ryoo R, Joo SH, Jun S. Synthesis of highly ordered carbon molecular sieves via template-mediated structural transformation. J Phys Chem B. 1999;103(37):7743–6.
76. Oh SM, Kim KB. Synthesis of a new mesoporous carbon and its application to electrochemical double-layer capacitors. Chem Commun. 1999;21:2177–8.
77. Jun S, et al. Synthesis of new, nanoporous carbon with hexagonally ordered mesostructure. J Am Chem Soc. 2000;122(43):10712–3.
78. Joo SH, et al. Ordered nanoporous arrays of carbon supporting high dispersions of platinum nanoparticles. Nature. 2001;412(6843):169–72.
79. Kruk M, et al. Synthesis and characterization of hexagonally ordered carbon nanopipes. Chem Mater. 2003;15(14):2815–23.
80. Solovyov LA, et al. Comprehensive structure analysis of ordered carbon nanopipe materials CMK-5 by X-ray diffraction and electron microscopy. Chem Mater. 2004;16(11):2274–81.
81. Lund K, Muroyama N, Terasaki O. Accidental extinction in powder XRD intensity of porous crystals: mesoporous carbon crystal CMK-5 and layered zeolite-nanosheets. Microporous Mesoporous Mater. 2010;128(1):71–7.
82. Che S, et al. Synthesis of large-pore Ia3d mesoporous silica and its tubelike carbon replica. Angew Chem. 2003;115(33):4060–4.
83. Kim S-S, et al. Nanocasting of carbon nanotubes: in-situ graphitization of a low-cost mesostructured silica templated by non-ionic surfactant micelles. Chem Commun. 2003;12:1436–7.
84. Kim J, Lee J, Hyeon T. Direct synthesis of uniform mesoporous carbons from the carbonization of as-synthesized silica/triblock copolymer nanocomposites. Carbon. 2004;42(12): 2711–9.
85. Yang C-M, et al. Facile template synthesis of ordered mesoporous carbon with polypyrrole as carbon precursor. Chem Mater. 2005;17(2):355–8.
86. Lee J, Kim J, Hyeon T. Recent progress in the synthesis of porous carbon materials. Adv Mater. 2006;18(16):2073–94.

87. Liang C, Li Z, Dai S. Mesoporous carbon materials: synthesis and modification. Angew Chem Int Ed. 2008;47(20):3696–717.
88. Inagaki M, Orikasa H, Morishita T. Morphology and pore control in carbon materials via templating. RSC Adv. 2011;1(9):1620–40.
89. Xia Y, Yang Z, Mokaya R. Templated nanoscale porous carbons. Nanoscale. 2010;2(5): 639–59.
90. Shi Y, Wan Y, Zhao D. Ordered mesoporous non-oxide materials. Chem Soc Rev. 2011;40(7):3854–78.
91. Moriguchi I, et al. Micelle-templated mesophases of phenol-formaldehyde polymer. Chem Lett. 1999;11:1171–2.
92. Liang C, et al. Synthesis of a large-scale highly ordered porous carbon film by self-assembly of block copolymers. Angew Chem Int Ed. 2004;43(43):5785–9.
93. Meng Y, et al. Ordered mesoporous polymers and homologous carbon frameworks: amphiphilic surfactant templating and direct transformation. Angew Chem. 2005;117(43):7215–21.
94. Tanaka S, et al. Synthesis of ordered mesoporous carbons with channel structure from an organic–organic nanocomposite. Chem Commun. 2005;16:2125–7.
95. Yu C, et al. High-yield synthesis of periodic mesoporous silica rods and their replication to mesoporous carbon rods. Adv Mater. 2002;14(23):1742–5.
96. James D, et al. Chemical sensors for electronic nose systems. Microchim Acta. 2005;149 (1–2):1–17.
97. Harris CM. Product review: seeing SAW potential. Anal Chem. 2003;75(15):355-A.
98. Kesting RE, Fritzsche AK. Polymeric gas separation membranes. New York: Wiley; 1993.
99. Ku P-H, et al. Polymer/ordered mesoporous carbon nanocomposite platelets as superior sensing materials for gas detection with surface acoustic wave devices. Langmuir. 2012;28(31):11639–45.
100. Choi M, Ryoo R. Ordered nanoporous polymer–carbon composites. Nat Mater. 2003;2(7):473–6.
101. Chen S-Y, et al. A facile route to synthesizing functionalized mesoporous SBA-15 materials with platelet morphology and short mesochannels. Chem Mater. 2008;20(12):3906–16.
102. National Institute for Occupational Safety and Health (NIOSH) and United States of America. NIOSH Pocket Guide to Chemical Hazards; 1997.
103. Neimark AV, Ravikovitch PI. Capillary condensation in MMS and pore structure characterization. Microporous Mesoporous Mater. 2001;44:697–707.
104. Ravikovitch PI, Neimark AV. Characterization of micro-and mesoporosity in SBA-15 materials from adsorption data by the NLDFT method. J Phys Chem B. 2001;105(29):6817–23.
105. Ravikovitch PI, Neimark AV. Characterization of nanoporous materials from adsorption and desorption isotherms. Colloids Surf A Physicochem Eng Asp. 2001;187:11–21.
106. Hsu Y-C, et al. Facile synthesis of mesoporous silica SBA-15 with additional intra-particle porosities. Chem Mater. 2007;19(5):1120–6.
107. Yasin FM, Tye KF, Reaz MBI. Design and implementation of interface circuitry for CMOS-based SAW gas sensors. In: IEEE International Proceedings SOC Conference, 2005. Herndon, VA: IEEE; 2005.
108. Wolff U, et al. SAW sensors for harsh environments. IEEE Sensors J. 2001;1(1):4–13.
109. Casalnuovo SA, Hietala VM, Heller EJ, Frye-Mason GC, Baea AG, Wendt JR. Monolithic integration of GaAs SAW chemical microsensor arrays and detection electronics. In: Proceedings of Solid-State Sensor and Actuator Workshop; 2000 June 4–8. Hilton Head Island, SC; 2000.
110. Perera A, et al. A portable electronic nose based on embedded PC technology and GNU/Linux: hardware, software and applications. IEEE Sensors J. 2002;2(3):235–46.
111. Adhikari B, Majumdar S. Polymers in sensor applications. Prog Polym Sci. 2004;29(7): 699–766.
112. Kim SJ. The effect on the gas selectivity of CNT-based gas sensors by binder in SWNT/silane sol solution. IEEE Sensors J. 2010;10(1):173–7.

113. Hsu H-P, Shih J-S. Multi-channel surface acoustic wave sensors based on principal component analysis (PCA) and linear discriminate analysis (LDA) for organic vapors. J Chin Chem Soc. 2006;53(4):815–24.
114. Hsu H-P, Shih J-S. Multi-channel surface acoustic wave (SAW) sensor based on artificial back propagation neural (BPN) network and multivariate linear regression analysis (MLR) for organic vapors. J Chin Chem Soc. 2007;54(2):401–10.
115. Eisen MB, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95(25):14863–8.
116. Sokal RR. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;38:1409–38.

Odor Sensing Technologies for Visualization of Odor Quality and Space

Chuanjun Liu and Kenshi Hayashi

Abstract In recent years, sensors for objective evaluation of quality and quantity of odor substances have shown a wide range of potential applications in many fields. However, the odor quality is difficult to be expressed by quantitative data because the odor sensation is brought about by a variety of volatile compounds, which form a complicated, subjective olfactory sense. Recent progress in molecular biological research of the olfactory system have shown that an odor cluster map produced on the surface of olfactory bulb through olfactory receptors presents essential information for brain to perceive odorants. The clustering perception model provides us with a new concept to design odor sensors with performance equivalent to mammalian olfactory system. The biological-inspired odor sensing based on various molecular recognition technologies, such as partial structure recognized water membrane/Pt electrodes, benzene-patterned self-assembled monolayer (SAM) layers, size and polarity selected molecular sieve materials, and molecularly imprinted polymer (MIP) adsorbents, are introduced to construct an artificial odor map and to evaluate the odor quality. On the other hand, odorants in our living environment can only be perceived by the sense of our olfactory, and odor space is invisible to eyes. The temporal and spatial distribution of odorants in environment is also important information for human and other animals. However, the visualization of odor space by using conventional sensor technologies is a difficult task due to the limited spatiotemporal resolution. Here optical sensing technologies based on fluorescence imaging and localized surface plasmon resonance (LSPR) are developed to visualize the spatiotemporal distribution of odorants in environment. In addition, the application of the developed sensors in the visualization of human body odor and odor release from fragrance encapsulated cyclodextrin inclusion complexes are presented also.

C. Liu • K. Hayashi (✉)
Department of Electronics, Graduate School of Information Science
and Electric Engineering, Kyushu University, 744 Motooka, Nishi-ku,
819-0395 Fukuoka, Japan
e-mail: hayashi@ed.kyushu-u.ac.jp

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_9

1 Introduction

We are living in an environment surrounded by various odorants. These odors are closely linked to the basic necessities of our life. The acquisition, processing, and analysis of odor information are extremely helpful for our life (Fig. 1). Although great achievements have been made in the development of odor sensing technologies, challenges still remain in objective evaluation of odors because the performance of existing odor sensors is far less than that of the human olfactory system. This is like because our knowledge and understanding of the model of the olfactory receptors of the mammalian olfactory system is not yet sufficient. Another reason is that it is impossible for us to develop sensing materials with molecular recognition ability comparable to olfactory protein receptors. It has been estimated that more than 400,000 different compounds are odorous to human nose. Unlike the primary colors in visual perception or primary tastes in gustatory perception, there are no primary odors in the case of olfactory perception. Therefore, the sensing methodology based on the assumption that there exist fundamental or elemental odorants is not appropriate for the qualitative evaluation of odor sensation.

Recent progress in molecular biology showed that there exist approximately 400 kinds of odor receptors in human olfactory system [1–3]. Different from the rigid lock-and-key model of common biological binding, odor receptors recognize not whole chemical structures of odorants but their partial structures; i.e. odotope theory of smell [4]. As a result, one receptor can recognize multiple odorants, and each odorant is recognized by multiple receptors in the olfactory system [5, 6]. After the odorants binding to receptors, the olfactory receptor cells are activated and the electrical signals are delayed to glomerulus cells located in olfactory bulb. The glomerulus cells activated by odorants with similar molecular features are located in close proximity to the cells and form molecular-feature clusters, namely odor map. In the odor map, glomeruli in the same cluster (or subcluster) tend to represent a similar combination of molecular features which likely share similar "odor quality" to human nose [3]. Therefore, the odor map of glomeruli might be used to define the basic odor quality although human cannot recognize primary odors in the olfactory sense.

The clustering perception model of the mammalian olfaction inspired us to find new approach to design odor sensors with performance equivalent to human nose. The realization of the sensors, however, is obstructed by the recognition ability of sensing materials in connection with odorant molecular features. Although the sensor arrays of most electronic noses (e-noses), such as quartz crystal microbalance (QCM), metal oxide semiconductor (MOX) gas sensors, and conductive polymer sensors, show cross-selectivity on certain group of chemicals, their selectivity cannot be represented in the molecular level of odorants. Thus, it is difficult to capture the molecular features as well as the qualitative information of odorants by using the conventional e-noses, although their difference can be discriminated

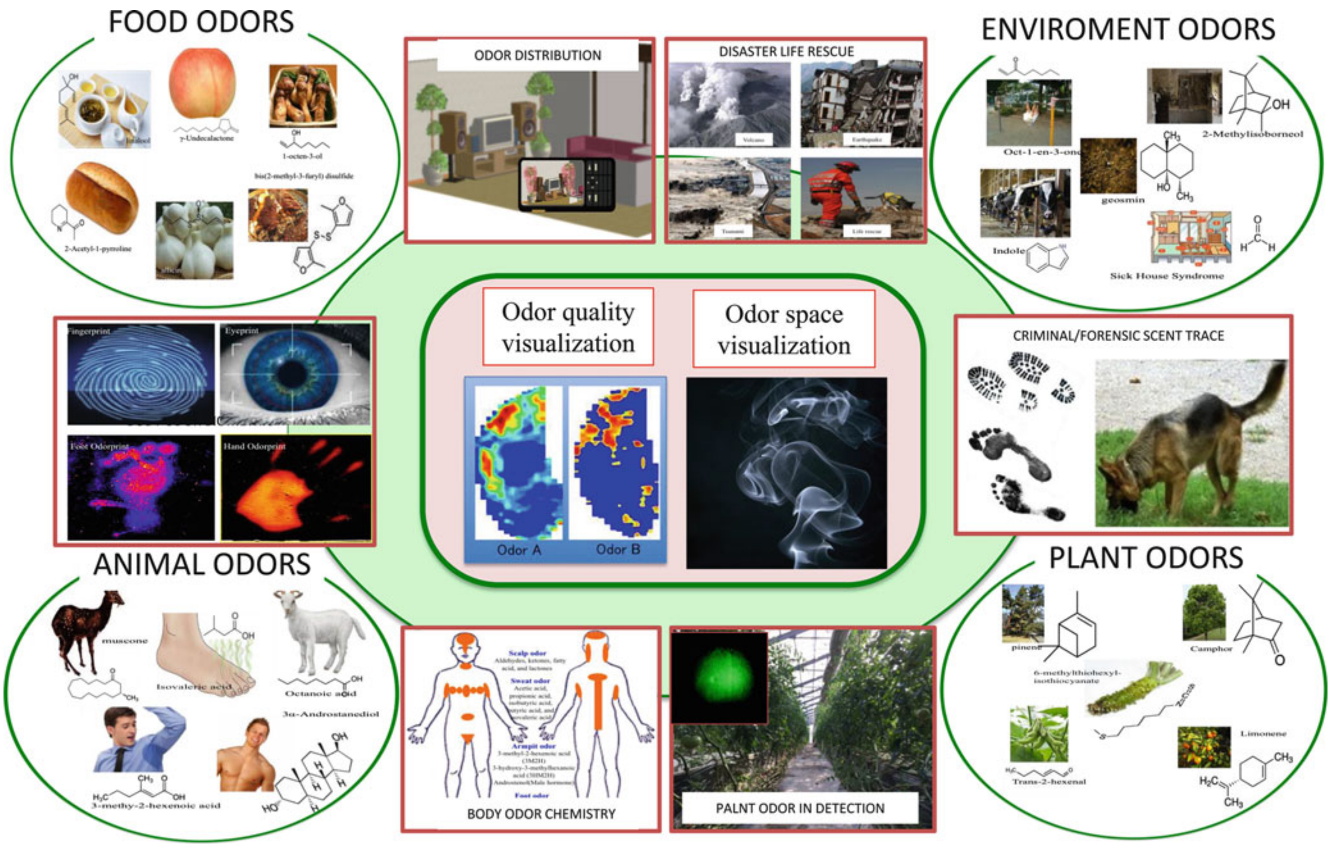


Fig. 1 Odor quality and space visualization and possible applications in our lives

by pattern recognition of the sensor output signals [7, 8]. In order to objectively evaluate odor quality, it is necessary to develop sensing materials or technologies to recognize the odorant molecular features.

2 Odor Quality Visualization

2.1 Odor Map and Odorant Molecular Features

2.1.1 Odor Map Analysis

Although we know the recognition rule of olfactory receptor with odorants, it remains unclear what kinds of molecular feature can be detected by olfactory receptors, and where the exact location of activated glomeruli is on the odor map. In order to figure out the detailed relationship between the molecular features and the odor map, we have carried out an odor map analysis on the basis of glomerular activity response archive of rat olfactory bulb, which is an odor map database provided publicly on the website by Leon group (http://gara.bio.uci.edu/). The published image database depicts the spatial distribution of 2-deoxyglucose uptake evoked in the glomerular of the rat olfactory bulb in response to wide range of defined odorant stimuli. Up to now, it have been updated approximately 400 odorants with 600 response patterns. Although the recorded odorant number is limited considering the amount of odorous substances, it provides a good example for researchers to investigate the relationship between molecular structure of odorants and the odor map response pattern.

The odor map analysis was carried out by the extraction of fingerprints of clusters and dimensionality reduction using principle component analysis (PAC) [9]. 321 matrices of 197 × 357 pixels were prepared from the gray scale images of 321 odorants. After the basis vector was calculated from the images with PCA, pixels having no contributions to the activity patterns were removed. The autoscaled data matrices were analyzed by PCA again. The principal component (PCn) scores and factor loadings of images were then calculated. Pixels corresponding to each PCn with smaller absolute value of factor loading were removed also. The above treatment finally afforded a PCA result with 80 % cumulative contribution from PC1 to PC80. Because the calculated factor loadings were small for the components after PC7, we just paid our attention to important principles from PC1 to PC6.

Odor maps based on restored pixel date of the six principle components are listed in Fig. 2a. These figures suggest that the odor images can be divided into nine independent areas. Quantitative evaluations were performed to compare the difference of odors in these areas [9]. The distribution of images depends on both positive and negative values of loadings. After a further comparison with principle score plots of different odor clusters, it is found that the nine activity regions divided by PCA are classified according to information related to molecular size and functional groups. The nine classified clusters and the common characteristics

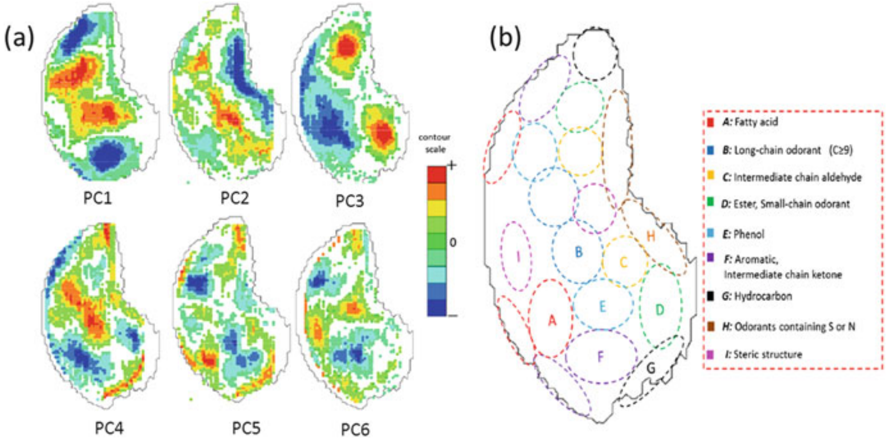

(a)
PC1
PC2
PC3
PC4
PC5
PC6
contour scale
+
0
–
(b)
A
B
C
D
E
F
G
H
I
🟥 A: Fatty acid
🟦 B: Long-chain odorant (C≥9)
🟨 C: Intermediate chain aldehyde
🟩 D: Ester, Small-chain odorant
🟦 E: Phenol
🟪 F: Aromatic, Intermediate chain ketone
⬛ G: Hydrocarbon
🟫 H: Odorants containing S or N
🟪 I: Steric structure

Fig. 2 (**a**) PCA analysis results based on glomerular activity response archive of rat olfactory bulb. Odor map images distributed by factor loading of PC1–PC6 were colored by the contour scale. (**b**) PCA-combined odor cluster map classified by functional groups of odorants (Reprinted with permission by MYU)

of odorants activated in each part are described in Fig. 2b. The nine clusters were obtained through a combination of areas corresponding to the six described principal components. The above odor-clustering map shows no discrepancy with results of other studies. For instance, the mapped cluster A area agrees with the activation by fatty acids [3, 10].

2.1.2 Extraction of Key Parameters for Odor Map Description

As we mentioned that the recognition of olfactory receptor with odorant is based on the partial structure instead of the whole chemical structure. From the standpoint of sensors, however, it is difficult to measure only the partial structure of odorants in most cases. To apply odor clustering to the sensor technology, measurable molecular parameters corresponding to each cluster were explored [9]. According to the calculation using computational package, MOPAC and GAMESS of ChemBio3D software (CambrisgeSoft), there were 76 types of molecular parameter could be obtained for the above 321 types of odorants. After the extraction of key parameters by PCA, the correlation coefficients between PCn and the key molecular parameters were obtained (as shown in Table 1). The results indicate that the parameters correlated with positive and negative principle component can be explored. Because the number of rotatable bonds and molecular size have more associations with PC1, higher or lower region in Fig. 2a are potentially activated by odorants with benzene rings or larger carbon chain. Moreover, it is found that the polarity information, e.g., acid dissociation constant (pK_a) and partition coefficient (log P), can become key information parameters. These analysis results are consistent with the practical activation pattern of glomerulus cells, although the exact match cannot be reached for all parameters and cluster position.

Table 1 Correlation results between principal component scores and molecular parameters

	PC1		PC2	
PCn	+	−	+	−
Correlated parameters	Number of rotatable bonds	Molecular length	Water solubility	Parameters correlated to elemental analysis (I, S)
Maximum correlation coefficient	0.56	0.56	−0.72	0.42
	PC3		PC4	
PCn	+	−	+	−
Correlated parameters	Parameters correlated to ester structure	• $\mathrm{p}K_\mathrm{a}$ • Critical temperature • Boiling point	• Percent hydrophilic surface • Straight carbon number • Log P • HLB	• Polar surface area • LogP • Partition coefficient
Maximum correlation coefficient		−0.62	0.63	−0.69
	PC5		PC6	
PCn	+	−	+	−
Correlated parameters	$\mathrm{p}K_\mathrm{a}$	Vapor pressure		$\mathrm{p}K_\mathrm{a}$
Maximum correlation coefficient	0.35	0.51		0.42

Correlated parameters and maximum correlation coefficients are described (Reprinted with permission by MYU)

Figure 3 describes the geometric progressions in odor maps of rats correlated with key parameters. Arrows indicate the activity progression by various incremental differences in structures. As can be seen, the regions of odor maps are classified by the degree of hydrophobicity and hydrophilicity of volatile chemicals, and the active parts progressively move according to increasing log P or hydrophilic-lipophilic balance (HLB) (blue arrow). Furthermore, chemotopic progressions by $\mathrm{p}K_\mathrm{a}$ and molecular size also occur in hydrophobic and hydrophilic regions. The results from the odor map analysis and molecular parameter extraction demonstrates that artificial odor-clustering maps can be roughly constructed using key parameters as long as they can be measured by corresponding recognition materials. Thus, it is possible to obtain the qualitative information of odorants via the artificial map because they show activity pattern close to the biological olfactory.

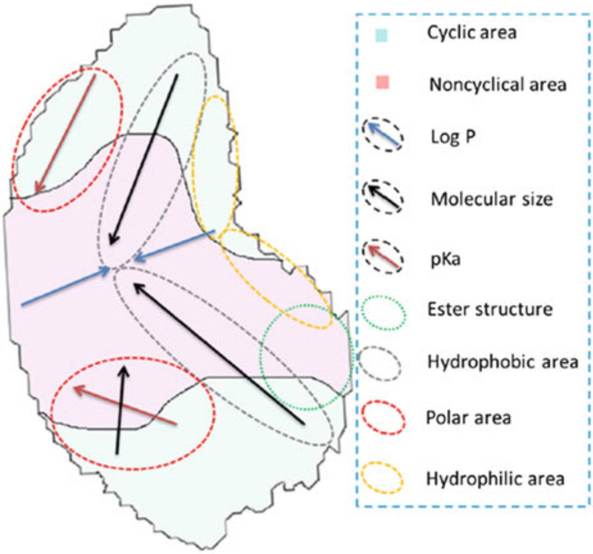


Fig. 3 Odor map images and geometric progression described by correlated key parameters with principal component scores (Reprinted with permission by MYU)

2.2 Odor Cluster Sensing Based on Partial Structure Recognition

2.2.1 Electrochemical Surface Polarity Controlling Analysis on Partial Structure of Odor Molecules

Our first olfactory-inspired odor sensor mimicked the odorant reception occurred in nasal mucus [11]. In the process of the biological odor reception, an odor molecule should be firstly dissolved into nasal mucus, and then its partial structures are received by transmembrane receptor proteins expressed from odorant receptor multi-genes [1]. We developed a water-membrane based surface polarity controlling electrochemical sensor to recognize the partial structure of odor molecules. In this sensor system, the thin water membrane worked as artificial mucus of olfactory system and platinum (Pt) electrode worked as the transducer to detect the partial structure of odorant molecules. Pt is a well-known material that can adsorb much amount of aromatic ring, hydroxyl, alkane and oxide nitrogen, which are partial structures included in many odorant molecules [12]. The adsorption character of odorants on the Pt electrode is different depending upon the polarity and the roughness of the surface, which can be reflected by electrical impedance called as a constant phase element (CPE) [13]. The interactions between odorants and the surface can be altered by controlling the surface polarization. Therefore, information about adsorbing odorants can be obtained by measuring CPE impedance through cyclic surface polarization impedance (cSPI) based on electrochemical impedance spectroscopy [14].

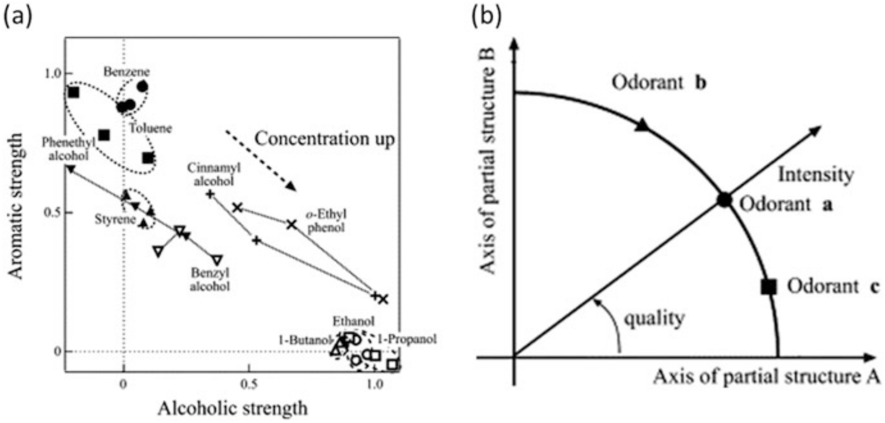


Fig. 4 (**a**) Map of odorants with the regression models of the responses to two partial structures (alcoholic hydroxyl group and aromatic ring). (**b**) Odor wheel which stands for odor quality and intensity by axes of partial structures (Reprinted with permission from Elsevier)

Fifteen odorants, including typical molecular structures of alcohols, aromatic compounds, aromatic alcohols and other odorants with different features, were examined. The change in both real part (ΔR_e) and imaginary part (ΔX_e) of the electrode impedance (Z_e) were calculated and plotted versus electrode polarity potential. The analysis results demonstrate that odorants with same functional groups show similar response profiles, while the response profiles of odorants with different functional groups are different from each other. For aromatic alcohols having both an alcoholic hydroxyl groups and an aromatic ring, they show response profiles with the combined characters of two groups (Fig. 4a). A linear multiple regression analysis based on the principal component scores of the three group odorants was carried out to build an odor map with the regression models of the response to two partial structures (alcoholic hydroxyl group and aromatic ring).

As shown in Fig. 4, it is possible to achieve the summarization of the features of the response to two partial structures, and the orthogonalization of the two regression models. The label of x-axis indicates the strength of the odor quality depending upon alcoholic hydroxyl group, and that of y-axis indicates the strength of odor quality upon aromatic ring. It is found that all alcohols are plotted around (1, 0), and those of benzene and toluene are plotted around (0, 1). For aromatic alcohols, they are plotted at the range of 0–1 against each axis. This suggests that normalization of alcoholic/aromatic compounds can be achieved and used to evaluate the odor quality and intensity of molecules with partial structures. For example, the calculated percentages of alcoholic hydroxyl and aromatic ring for four aromatic alcohols were 9 % and 91 % for phenethyl alcohol, 34 % and 66 % for benzyl alcohol, 57 % and 43 % for cinnamyl alcohol, and 59 % and 41 % for o-ethyl phenol, respectively. The above results indicate that the odor quality

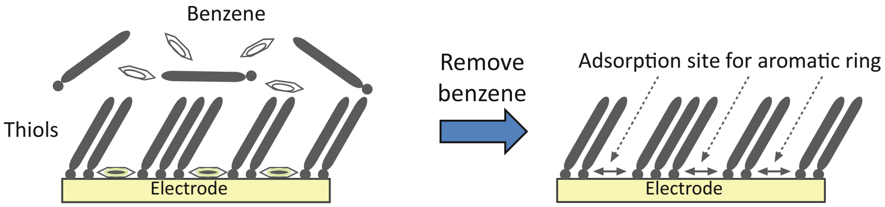


Fig. 5 Formation of benzene-patterned self-assemble monolayer (SAM) electrode

of benzyl alcohol and phenethyl alcohol is mainly dominated by aromatic rings, while that of o-ethyl phenol and cinnamyl alcohol is dominated mainly by alcoholic hydroxyl groups.

2.2.2 Benzene-Patterned Self-Assemble Monolayer (SAM) Electrode for Odor Recognition

Although the water membrane/Pt based odor sensor is effective in the recognition of partial structure, its resolution of response profile obtained by cSPI is not enough when several different odorants are adsorbed on the Pt electrode surface simultaneously. In order to improve the selectivity to specific odorants, especially for aromatic ring molecules, benzene-patterned SAMs electrodes were designed for the partial structure recognition [15, 16].

Figure 5 shows the formation process of benzene-patterned SAM electrode. The surface of the benzene-patterned SAM was characterized by AFM, ellipsometry, eletrochemical method (stripping volammetry), and contact angle [17]. As a result, the coverage and thickness of obtained benzene-patterned SAM were about 90 % compared with perfect coverage SAMs. It mean small amount of defects in the SAM structures that were functionalized as aromatic ring recognition sites. A cyclic surface-polarization impedance (cSPI) spectroscopy was used to transduce the surface adsorption to electric signals [11, 14, 18, 19].

Compared with the response of a bare gold electrode, a concentration dependent response was observed for the benzene-patterned SAM electrode. In addition, no affinity was shown for benzene patterned SAM electrode on ethanol and sucrose. This result confirms that the benzene-patterned SAM electrode shows specificity to odorants with aromatic rings. In order to improve the efficiency of the odor measurement, an integrated multi-channel sensor was developed to work as an artificial olfactory epithelium chip [16]. The water membrane of the artificial chip consisted of a mixture of 40 % 100 mM KCl and 60 % glycerin, which guaranteed a stable electrochemical measurement compared with the former water membrane/Pt electrode sensor system. In additional, each electrode was designed corresponding to a certain property of the odorant. For example, Fig. 6 shows a response profile of a multichannel chip on ethanol, phenethyl alcohol and benzene. The chip array was composed of an unspecific bare channel, a cellulose phosphate (P-cellulose) channel

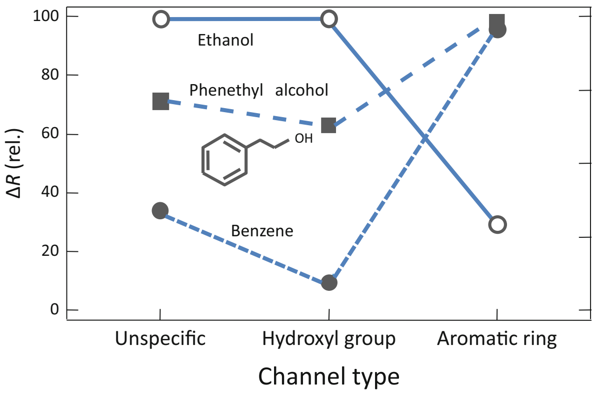


Fig. 6 Response profiles of a multi-channel sensor on different odorants

for hydroxyl group detection and a benzene-pattern SAM channel for aromatic ring group detection. It was found that the addition of P-cellulose to the glycerin membrane enhanced the specificity of the channel to ethanol detection. The different response profiles demonstrate that the artificial olfactory epithelium chip based on benzene-patterned SAM electrode is effective in the recognition of molecular partial structure of odorants.

2.3 Odor Cluster Sensing Based on Odor Adsorption/Separation System

2.3.1 Size and Polarity Clustering of Odorants

The previously described odor sensing based on partial structure recognition is not a complete imitation of the biological olfactory receiving process, but a physicochemical process in olfactory mucous membrane. This imitation is confined by limited molecular structures as well as odorant types. In order to design a sensor with high molecular recognition ability, it requires various kinds of nanostructures with high specificity and molecular profile recognition abilities to detect enormous numbers of odor molecules. On the other hand, there have existed various kinds of adsorbents with high affinities to different chemical compounds, e.g., column materials for gas chromatography (GC). In view of this, we developed an odor separating/sensing system by using conventional adsorbent materials and metal oxide (MOX) gas sensors [20, 21]. This is another approach that imitates the odor receptive mechanism of biological olfaction based on conventional gas detecting methods.

The system mainly consists of adsorbing and separating cells that enable classification of measured odorants based on molecular size and polarity (Fig. 7). Different materials, such as zeolite or carbon molecular sieves, gas chromatographic

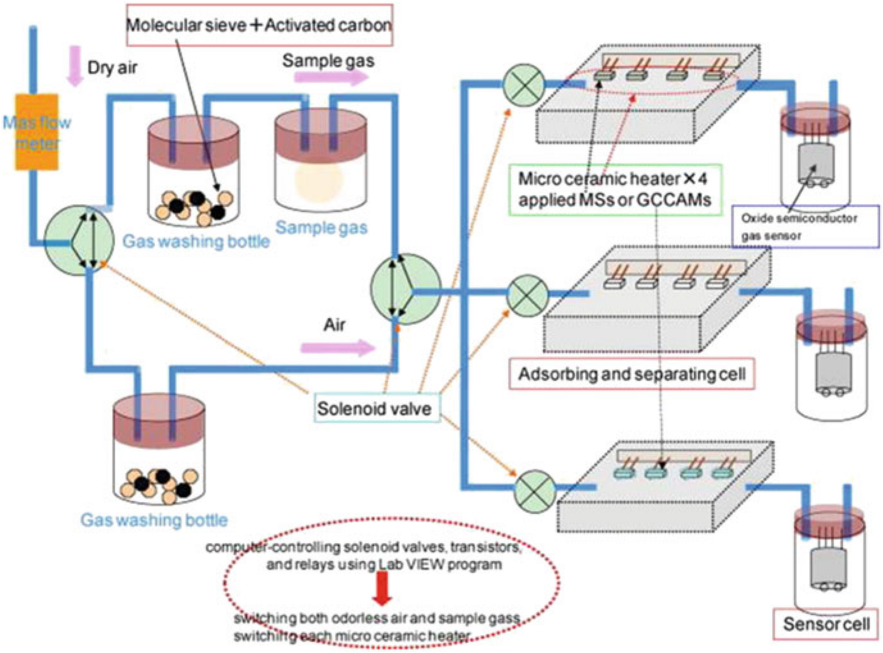

Molecular sieve + Activated carbon
Dry air
Sample gas
Mass flow meter
Gas washing bottle
Sample gas
Air
Solenoid valve
Gas washing bottle
Micro ceramic heater × 4
applied MSs or GCCAMs
Oxide semiconductor gas sensor
Adsorbing and separating cell
Sensor cell
computer-controlling solenoid valves, transistors, and relays using Lab VIEW program
switching both odorless air and sample gass, switching each micro ceramic heater.

Fig. 7 Response profiles of a multi-channel sensor on different odorants (Reprinted with permission from Elsevier)

column adsorbent (dimethylpolysiloxane: PDMS, polyethylene glycol: PEG, et al.), were used as adsorbent that enable classification of measured odorants according to their molecular size and polarity. Ceramic micro-heaters were set in each adsorbing and separating cell, in which the adsorbent material was adhered on the surface the heaters. Mixed odorants were delivered through computer controlled mass flow controller and solenoid valve system. During a programmed heating process, the trapped odorants were desorbed from the adsorbent and delivered to the connected sensor cell and measured with the MOX gas sensor. After the analysis of time course response pattern of the MOX gas sensors, the size and polarity information of the odorants could be extracted and used to construct an odor map with clusters (A–D) similar to that created on the olfactory bulb (Fig. 8). In addition, it was confirmed that mixed odors could be discriminated and decomposed into the elemental clusters by using the above odor adsorption/separation system.

2.3.2 Odor Cluster Sensing Based on Molecularly Imprinted Polymer (MIP) Materials

MIP is a polymer prepared by the molecular imprinting technique that leaves cavities in polymer matrix with affinity to template molecules. The imprinted cavities can fit the size, shape, and functional groups of target molecules. Therefore,

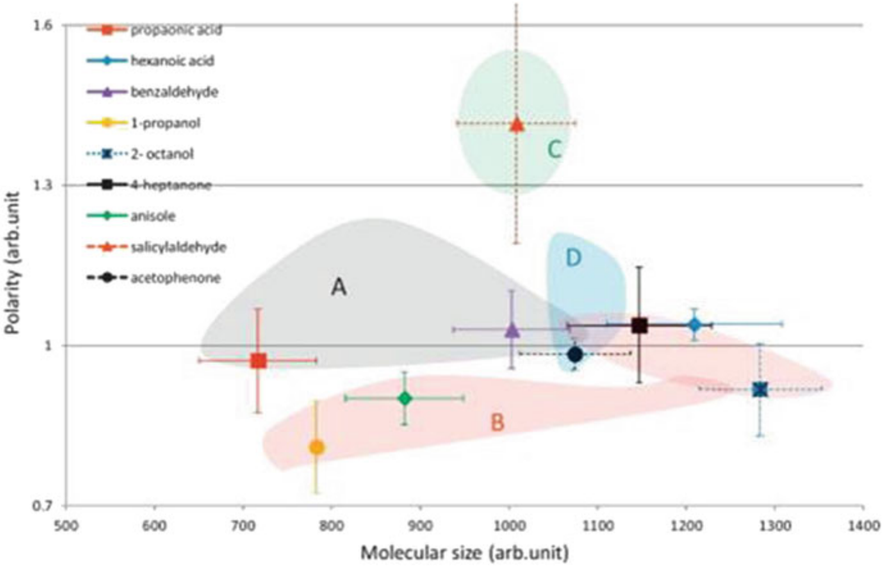


Fig. 8 Mapping of odorant using estimated molecular size and polarity obtained using the odor adsorption/separation system (Reprinted with permission from Elsevier)

the MIP material has gained considerable attention as an artificial receptor material with high molecular recognition ability [22, 23]. Active studies have been carried out for the application of MIP materials in the field of chromatographic separation [24], solid-phase extraction (SPE) [25, 26], and chemical sensors [27]. However, challenges remain for the application of MIP materials as gas or odor absorbance due to the particularity of odorant molecules such as low molecular weight and volatility. In addition, the non-specific adsorption occurred on the matrix polymer instead of recognition sites becomes outstanding if compared with the liquid phase adsorption. In order to overcome these problems, molecularly imprinted filtering adsorbents (MIFAs) were developed for highly selective gas detection [28, 29].

Figure 9 shows the design of the structure and preparation process of MIFAs. Polydimethylsiloxane (PDMS) was selected as concentrating substrate due to its excellent adsorbing performance in solid-phase microextraction (SPME). A titanium oxide (TiO_2) monolayer was firstly assembled on the PDMS surface using the surface sol-gel process. Then, poly(acrylic acid) (PAA)/template complex was chemically bound to the TiO_2 modified PDMS. The remove of the template molecules finally afforded the MIFAs. The thickness of the TiO_2/PAA layer was evaluated with a few nanometers. Specific binding sites were formed between PAA matrix or Ti-O layer and template molecules on the basis of the intermolecular forces such as hydrogen bonds and van der Waals force. The adsorption evaluation experiment by SPME-GC-MS confirmed the selectivity of MIFAs on a variety of odorants such as fatty acids and aldehydes. When MIFAs were embedded into the

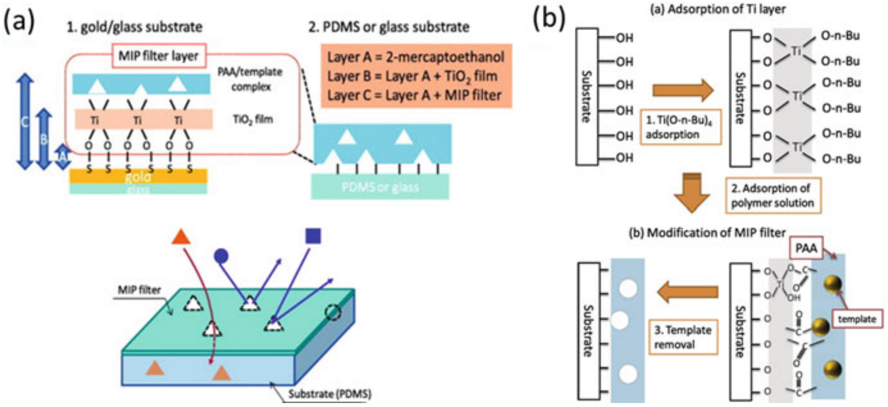

(a)
1. gold/glass substrate
MIP filter layer
PAA/template complex
TiO₂ film
Ti
O
S
gold
glass
C
B
A
2. PDMS or glass substrate
Layer A = 2-mercaptoethanol
Layer B = Layer A + TiO₂ film
Layer C = Layer A + MIP filter
PDMS or glass
MIP filter
Substrate (PDMS)
(b)
(a) Adsorption of Ti layer
Substrate
OH
O
Ti
O-n-Bu
1. Ti(O-n-Bu)₄ adsorption
2. Adsorption of polymer solution
(b) Modification of MIP filter
PAA
template
3. Template removal

Fig. 9 Schematic illustrations of the structure of MIFAs (**a**) and the preparation of an MIP filter by surface sol-gel process (**b**) (Reprinted with permission from Elsevier)

adsorption/separation system of Fig. 7, the sensor response to single vapor and odor mixtures were measured and artificial maps were constructed successfully (Fig. 10), which can be used in the objective evaluation of odor quality [29].

3 Visualization of Odor Space

3.1 Significance of Odor Space Visualization

We are living in an environment surrounded by a variety of odors. The odors are caused by volatile organic compounds, most of which are invisible to eyes and can only be perceived by the sense of our olfaction. Generally, odors exist in the environment as the form of "plume", which is carried out by wind to form an instantaneous, complicated concentration structure [30]. The detection of the spatiotemporal distribution of odor in environment occupies an important position in the life of human and other animals. For example, animals localize odor source (such as foods, mates, and predators) by perceiving the spatiotemporal structure of odor concentration, especially the concentration gradient along the direction of odor current [31]. Sensors with the abilities to detect or visualize the odor space have many important applications. They can be loaded on mobile robots to imitate the odor-locating behavior of animals, and thus used in searching for toxic gas leak, hazardous chemicals, and pollutant sources [32–38].

On the other hand, olfactory information is generally difficult to be transmitted compared with other sensory information, such as vision and hearing, because the sense of smell is a very primitive and personal feeling for our brain. If the odor space can be visualized, the olfactory perception can be conveyed as visual

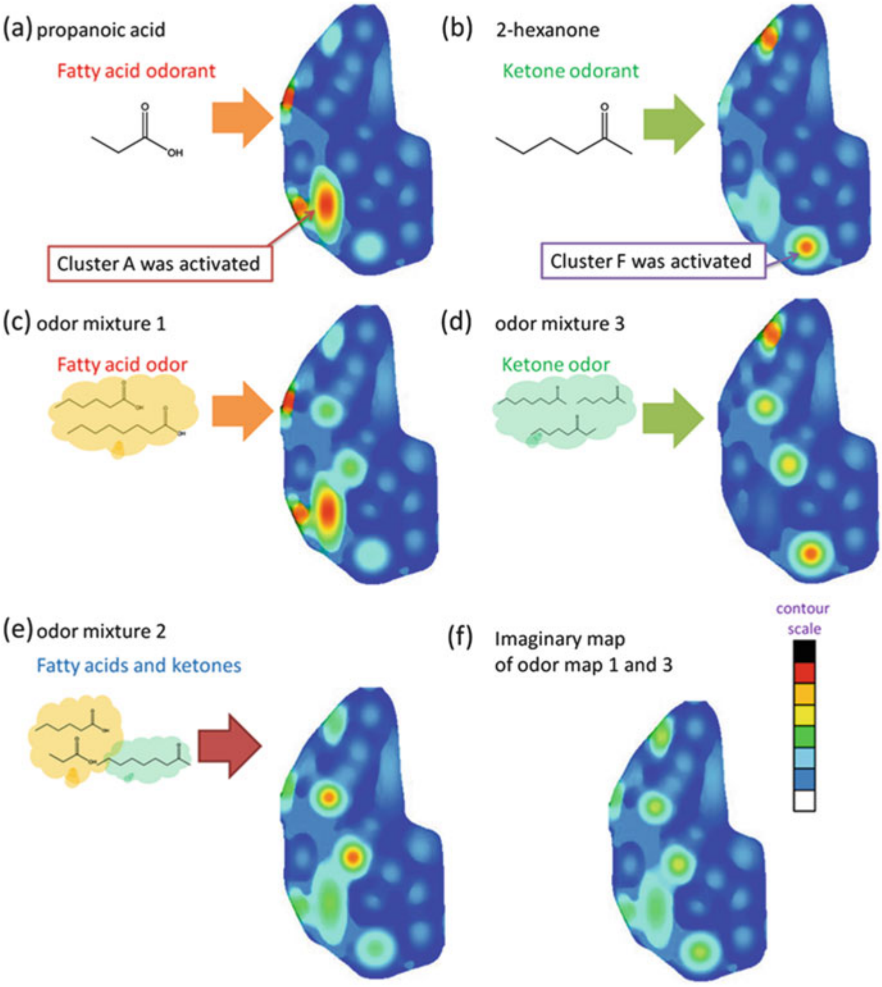


Fig. 10 The artificial odor maps of different single and mixed odors based on response of the sensor system shown in Fig. 7

perception, which will be a more intuitive approach to express and transmit odor information. Although great achievements have been made in the development of gas/odor sensors, the odor visualization based on conventional sensors is hindered by the poor temporal and spatial resolution. Therefore, it is necessary to develop new types of sensing systems to meet the requirement of the odor visualization. In this work, two odor imaging sensors based on fluorescence imaging and localized surface plasmon resonance (LSPR) are proposed, and their research and application for odor visualization are introduced.

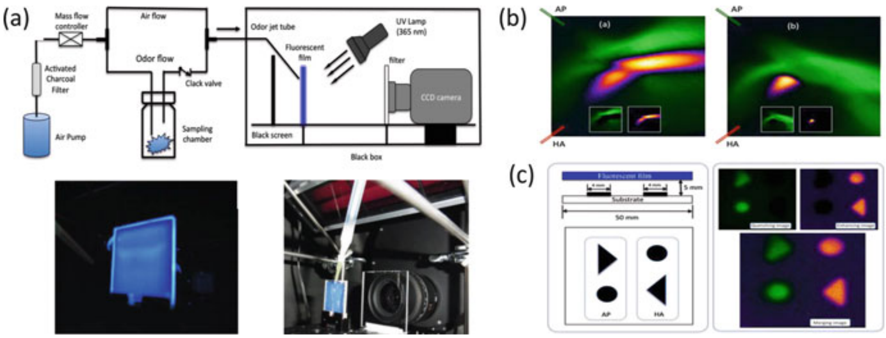


Fig. 11 (**a**) Experiment setup for the odor visualization. (**b**) Merged RGB images for the visualization of two odor flows. (**c**) Schematic diagram for visualizing the shape of different odor marks (Reprinted with permission from IEEE sensors)

3.2 Odor Visualization by Fluorescence Imaging

Fluorescence imaging is a well-known technology having wide application in many fields. The odor visualization by fluorescence imaging is based on the change in fluorescence intensity of fluorescent probes by odorants according to various interactions, such as solvent effect, pH effect, fluorescence resonance energy transfer (FRET), and so on. Investigations by fluorescence spectrophotometer demonstrated that there exist fluorescence quenching or enhancing interactions between typical odorants and fluorescent probes [39]. These phenomena were utilized to fabricate the fluorescence image sensor for the odor visualization. As shown in Fig. 11a, the sensor system consisted of an excitation light source, an agarose gel film containing fluorescent probes, and a high sensitive, cooled charge-coupled device (CCD) camera. The fluorescence intensity and spectrum changes of the sensing film derived from the interactions with odor molecules were recorded by the CCD camera. The fluorescence images were analyzed by using ImageJ software to visualize odors.

The sensor can be used to visualize not only the existence of odorants, but also the temporal and spatial distribution of odors in the environment. For example, Fig. 11b represents a typical visualized result for two different odor flows impacting on the sensing film surface [40]. It can be seen that two domains with clear boundary were successfully visualized by the fluorescence imaging. When the flow rate of one odor was changed, the boundary and shape were also changed. The detection of the spatial boundary of different odor flows is an interesting founding because it is generally difficult to measure by using conventional sensor arrays with high resolution. In addition, we developed a technique called "odor exposure" to record odors remained in environment. In the odor exposure experiment, the sensing film was close (instead of contact) to the odor source to absorb the volatilized odorants. After a certain exposure time, the fluorescence image of the film was taken and compared with that before the exposure. The result shown in Fig. 11c demonstrate that not only the shape of remained odor sources can be factually recorded, but also the type of odors can be distinguished by specific image treatment such as quenching or enhancing. Now, we are committed to develop fluorescent probes with higher sensitivity and selectivity upon various odorants. We hope in the future the complicated odor fingerprint can be identified and visualized by using such kinds of technology with much higher performance.

3.3 Odor Visualization by Localized Surface Plasmon Resonance (LSPR) Sensor

The most prominent characteristic of the fluorescence imaging is that they can record the shape and distribution of odors remained in environment via "odor exposure" for a long period of time, which is difficult to be realized by conventional sensor technologies. The visualization based on the fluorescence gel film, however, has shortcoming that its recovery time is greatly larger than its response time due to the slow desorption process of odorants from the film. The visualized images are a time-average spatial distribution of odorants in space, and thus, lack of sufficient time resolution. In order to realize the real time visualization of odor space, we paid our attention to odor sensors based on localized surface plasmon resonance (LSPR) phenomenon of metal nanoparticles. LSPR is a well-known optical phenomenon of metal nanoparticles (MNPs), which is generated by a light wave trapped within conductive nanoparticles smaller than the wavelength of light. The interaction between the incident light and surface electron of MNPs in conduction band results in coherent localized plasmon oscillation with a resonant frequency that strongly depends on the composition, size, geometry, dielectrical environment and particle–particle separation distance od MNPs. The gas sensing principle of LSPR sensor is based on a fact that the plasmon resonant frequency of MNPs is highly sensitive to the refractive index of surroundings. A small change in refractive index will result in a shift in the resonant frequency or intensity of transmittance/absorbance light (Fig. 12a). Since the sensor response has no need of chemical reactions, the most notable features of the LSPR odor sensors are fast response and fast recovery, which make it very suitable for real time visualization of odor space.

The LSPR sensor was fabricated by vacuum deposition of AuNPs on glass substrate. The influence of preparation conditions (such as current and time of deposition, temperature and time of annealing) on the gas sensing sensitivity was extensively studied [41]. In the odor visualization experiment, we investigated the response of the sensor on an odor flow under different measurement modes (transmittance, reflect, and scattering). The results demonstrated that the measurement in the reflect mode showed the highest response (Fig. 12b). Also, we investigated the real time response ability of the LSPR sensor [42]. The time scale of response and recovery was confirmed in the order of seconds (Fig. 12c), which meets the requirement for real time visualization. In order to further increase the sensitivity and visualization effect, we endeavor to develop LSPR sensors based on novel

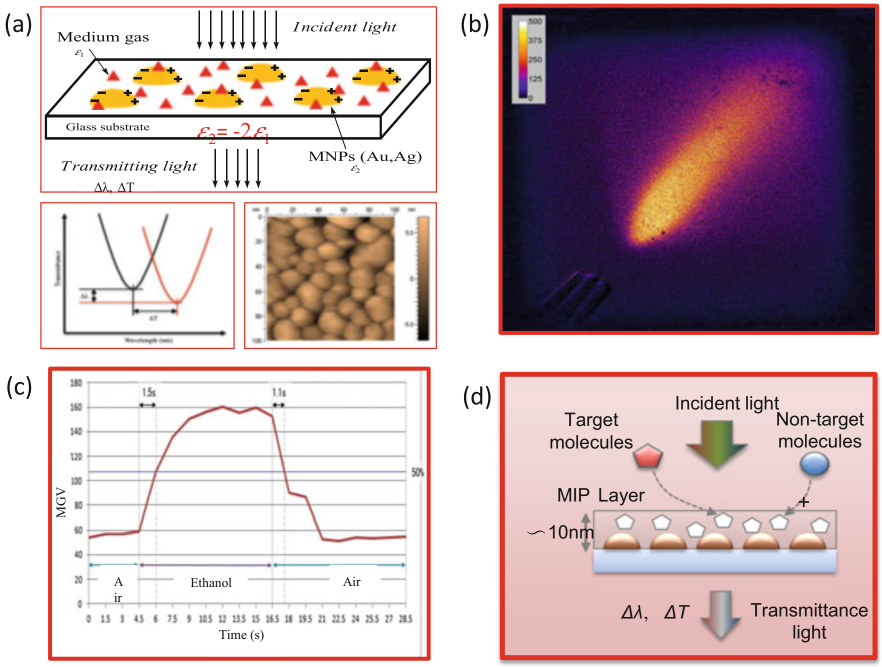


Fig. 12 (**a**) Schematic response mechanism of LSPR sensors. (**b**) Visualization result of a ethanol flow by using an AuNPs LSPR sensor. (**c**) Real-time response character of the LSPR sensor. (**d**) Schematic structure of an AuNPs LSPR sensor combined with MIP layer to enhance the selectivity

nanostructures [43, 44]. At the same time, as shown in Fig. 12d, molecularly imprinted polymer technologies were introduced on the surface of MNPs to acquire molecular recognizing ability in LSPR sensor with selective response upon various odorants [45].

4 Applications of Odor Visualization

4.1 Visualization of Human Body Odor

Human body emits a variety of volatile compounds (VOCs) contributing to a person’s body odor. The components of individual body odor are determined by many factors involving health status, hereditary features, food habits and even gender and age. Therefore, the body odors contain a large amount of information related to the body chemistry of individuals. The assay of human body odors has attracted much research interest in many fields such as cosmetics and medicines. The most commonly used analysis method for body odor is GC/MS combined with SPME. Due to its high cost and complicated operation skills, its application is limited in the laboratory level. With the increasing importance of body odor field assay, it is imperative to develop odor sensors as an alternative of GC/MS. However, it still lacks effective evaluation methods based on sensor technologies, and most field assays of human body odor at the present stage have to rely on human panelist or animals (such as canine) [46, 47].

The developed fluorescence image sensing was applied to detect the sweat odor of human body [48]. Volatile organic acid substances with low molecular weight were supposed to be one of the key contributors to the odor of a sweat sample collected from the human body. Quinine sulfate, a pH-dependent fluorescent compound that shows fluorescence enhancing in the presence of organic acids, was utilized to prepare the gel sensing film. The existence of organic acid components with high abundance in the collected sweat sample was confirmed by the analysis of SPME-GC/MS. Sequence images of the fluorescent film in the absence and presence of odors were recorded by a high-resolution CCD camera. The response of the sensor to typical organic acids in sweat odor, such as acetic acid, isovaleric acid, hexanoic acid and octanoic acid, was evaluated by the analysis of subtraction images as well as the real-time change of mean gray value (MGV). The fresh sweat is odorless. But after 54 h incubation (38 °C, sealed bottle), the odor of the sweat sample was successfully visualized (Fig. 13a), which indicates the application potential of the developed fluorescence imaging sensor in the detection of human body odor.

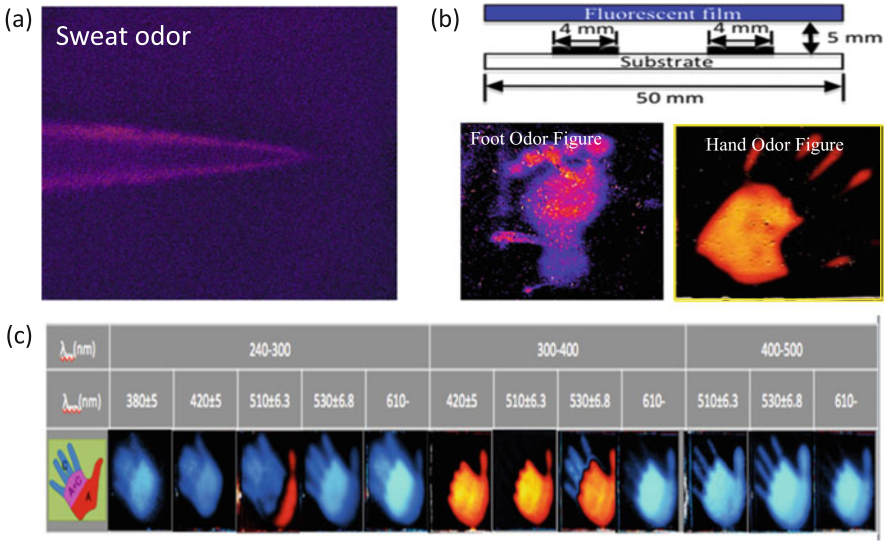


Fig. 13 (**a**) Visualized detection of human body sweat odor. (Reprinted with permission from Elsevier) (**b**) Visualization of foot and hand odor figures. (**c**) Multispectral fluorescence imaging results for hand odor figure with complicated odor composition and shape

Also, we used the approach of “odor exposure” to visualize the odor shapes of hand and foot. In the hand odor visualization experiment, the odor of hexanoic acid was coated on a rubber glove, while in the foot odor shape visualization; an isovaleric acid wetted paper with foot figure was used. The odors were exposed to the fluorescent film with a close distance. Both the hand and foot odor shapes were clearly visualized as shown in Fig. 13b, which demonstrates the high spatial resolution of the fluorescence image sensing. In addition to organic acids, human body odor contains a lot of other odorants, such as ketones and aldehydes, which mix together to form the complicated human body odors. In order to visualize and discriminate the complicated components and shape of human body odors, multispectral fluorescence imaging techniques were developed as shown in Fig. 13c, in which multiple fluorescent probes having versatile interactions were designed to visualize the mixed human body odors.

4.2 Visualization of Odor Release from Fragrance Inclusion Complexes

Evaluation of odor release is important for the development of encapsulated fragrance products. However, the use of existing instrumental analysis or gas-sensor technologies for odor-release evaluation is time- and labor-consuming. We here introduce the fluorescence imaging, which can be used as a simple but effective tool to visualize a controlled release of fragrance from cyclodextrin inclusion complexes [49]. The visualization was based on the fluorescence changes of samples caused by the release of fragrance molecules from the cavity of cyclodextrin. Fluorescence spectroscopy investigations proved that different strategies could be used in the visualization depending on the molecular characters of the encapsulated fragrances. In the case of fragrance molecules with fluorescence (typically methyl anthranilate: MA), the change of fluorescence intensity was derived from the changed microenvironment of MA—from the encapsulated state to the released state. For the fragrance molecules without fluorescence (typically menthol), an appropriate fluorescent probe (7-amino-4-methylcoumarin: AMC) was used, and the fluorescence change was observed due to the interactions between AMC and the released menthol molecules. The real-time releases were visualized by two-dimensional/three-dimensional surface plots of the fluorescence images using image-processing software. Both the moisture-activated burst release occurring of encapsulated MA on the short time scale and the slow, natural release in ambient conditions of encapsulated menthol can be demonstrated by the proposed visualization technique (Fig. 14).

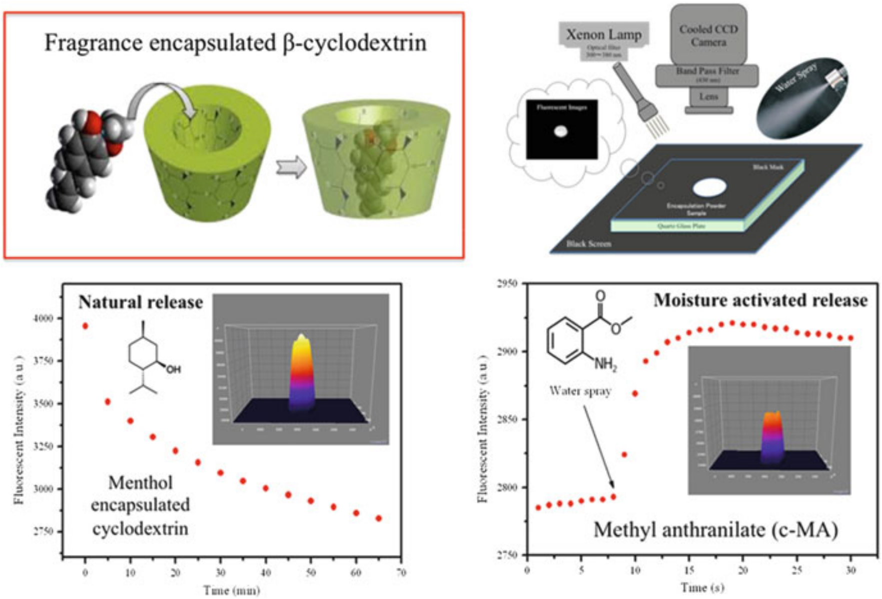


Fig. 14 Visualization of the odor release from fragrance encapsulated β-cyclodextrin (Reprinted with permission from Wiley)

5 Conclusions

Visualization sensing technologies aimed to evaluate odor quality and odor space are introduced in this chapter. A concept based on olfactory inspired odor cluster sensing was proposed to construct artificial odor map, from which qualitative information of odorant molecules can be obtained in metric space. Odor sensing based on various molecular recognition materials and sensor systems confirmed that the odor cluster sensing is an effective approach to evaluate odor quality in a manner close to the biological olfactory. Different from the conventional odor sensor aiming to obtain the odor intensity and quality information, the target of odor imaging sensor is to obtain the temporal and spatial information of odorants in environment. The odor image sensor can be used as an initiative method to "see" invisible odors in our living space. It is believed that more and more novel applications can be expected for the developed image sensor in the field of odor sensing.

References

1. Buck L, Axel R. A novel multigene family may encode odorant receptors – a molecular-basis for odor recognition. Cell. 1991;65:175–87.
2. Axel R. Scents and sensibility: a molecular logic of olfactory perception (Nobel lecture). Angew Chem Int Ed. 2005;44:6110–27.

3. Mori K, Takahashi YK, Igarashi KM, Yamaguchi M. Maps of odorant molecular features in the mammalian olfactory bulb. Physiol Rev. 2006;86:409–33.
4. Mori K, Shepherd GM. Emerging principles of molecular signal processing by mitral/tufted cells in the olfactory bulb. Semin Cell Biol. 1994;5:65–74.
5. Malnic B, Hirono J, Sato T, Buck LB. Combinatorial receptor codes for odors. Cell. 1999;96:713–23.
6. Araneda RC, Kini AD, Firestein S. The molecular receptive range of an odorant receptor. Nat Neurosci. 2000;3:1248–55.
7. Harper WJ. The strengths and weaknesses of the electronic nose. In: Headspace analysis of foods and flavors, vol. 488. New York: Springer; 2001. p. 59–71.
8. Rock F, Barsan N, Weimar U. Electronic nose: current status and future trends. Chem Rev. 2008;108:705–25.
9. Imahashi M, Hayashi K. Odor clustering based on molecular parameter for odor sensing. Sens Mater. 2014;26:171–80.
10. Matsumoto H, Kobayakawa K, Kobayakawa R, Tashiro T, Mori K, Sakano H, Mori K. Spatial arrangement of glomerular molecular-feature clusters in the odorant-receptor class domains of the mouse olfactory bulb. J Neurophysiol. 2010;103:3490–500.
11. Izumi R, Hayashi K, Toko K. Odor sensor with water membrane using surface polarity controlling method and analysis of responses to partial structures of odor molecules. Sens Actuators B. 2004;99:315–22.
12. Montilla F, Huerta F, Morallon E, Vazquez JL. Electrochemical behaviour of benzene on platinum electrodes. Electrochim Acta. 2000;45:4271–7.
13. Rammelt U, Reinhard G. On the applicability of a constant phase element (CPE) to the estimation of roughness of solid metal-electrodes. Electrochim Acta. 1990;35:1045–9.
14. Hayashi K, Ju MJ, Hayama K, Toko K. Chemical sensor using polarity controlled fractal surface. In: Transducers '01: Eurosensors Xv. Digest of Technical Papers, Volumes 1 and 2; 2001. p. 1770–3.
15. Hayama K, Hayashi K, Toko K. Functionalization of gold surfaces using benzene-patterned self-assembled monolayers for surface-polarization controlling method. Sens Mater. 2003;15:403–12.
16. Izumi R, Etoh S, Hayashi K, Toko K. Evaluation of the odor quality by substructures of odor molecules using integrated multi-channel odor sensor. In: Transducers '05. Digest of Technical Papers, Volumes 1 and 2; 2005. p. 1884–7.
17. Masunaga K, Michiwaki S, Izumi R, Ivarsson P, Bjoerefors F, Lundstrom I, Hayashi K, Toko K. Development of sensor surface with recognition of molecular substructure. Sens Actuators B. 2008;130:330–7.
18. Hayama K, Tanaka H, Ju MJ, Hayashi K, Toko K. Fabrication of a flow cell for electrochemical impedance measurements. Sens Mater. 2002;14:443–53.
19. Ju MJ, Hayama K, Hayashi K, Toko K. Discrimination of pungent-tasting substances using surface-polarity controlled sensor with indirect in situ modification. Sens Actuators B. 2003;89:150–7.
20. Imahashi M, Hayashi K. Odor clustering and discrimination using an odor separating system. Sens Actuators B. 2012;166:685–94.
21. Jha SK, Hayashi K. A novel odor filtering and sensing system combined with regression analysis for chemical vapor quantification. Sens Actuators B. 2014;200:269–87.
22. Alexander C, Davidson L, Hayes W. Imprinted polymers: artificial molecular recognition materials with applications in synthesis and catalysis. Tetrahedron. 2003;59:2025–57.
23. Ge Y, Butler B, Mirza F, Habib-Ullah S, Fei D. Smart molecularly imprinted polymers: recent developments and applications. Macromol Rapid Commun. 2013;34:903–15.
24. Owens PK, Karlsson L, Lutz ESM, Andersson LI. Molecular imprinting for bio-and pharmaceutical analysis. Trends Anal Chem. 1999;18:146–54.
25. Lanza F, Sellergren B. The application of molecular imprinting technology to solid phase extraction. Chromatographia. 2001;53:599–611.

26. Martin-Esteban A. Molecularly imprinted polymers: new molecular recognition materials for selective solid-phase extraction of organic compounds. Fresenius J Anal Chem. 2001;370: 795–802.
27. Algieri C, Drioli E, Guzzo L, Donato L. Bio-mimetic sensors based on molecularly imprinted membranes. Sensors. 2014;14:13863–912.
28. Imahashi M, Hayashi K. Concentrating materials covered by molecular imprinted nanofiltration layer with reconfigurability prepared by a surface sol-gel process for gas-selective detection. J Colloid Interface Sci. 2013;406:186–95.
29. Imahashi M, Watanabe M, Jha SK, Hayashi K. Olfaction-inspired sensing using a sensor system with molecular recognition and optimal classification ability for comprehensive detection of gases. Sensors. 2014;14:5221–38.
30. Murlis J, Elkinton JS, Carde RT. Odor plumes and how insects use them. Annu Rev Entomol. 1992;37:505–32.
31. Reidenbach MA, Koehl MAR. The spatial and temporal patterns of odors sampled by lobsters and crabs in a turbulent plume. Integr Comp Biol. 2009;49:E141.
32. Ishida H, Nakamoto T, Moriizumi T. Remote sensing of gas/odor source location and concentration distribution using mobile system. Sens Actuators B. 1998;49:52–7.
33. Ishida H, Tanaka H, Taniguchi H, Moriizumi T. Mobile robot navigation using vision and olfaction to search for a gas/odor source. Auton Robot. 2006;20:231–8.
34. Jatmiko W, Sekiyama K, Fukuda T. A PSO-based mobile robot for odor source localization in dynamic advection-diffusion with obstacles environment: theory, simulation and measurement. IEEE Comput Intell Mag. 2007;2:37–51.
35. Oh EH, Song HS, Park TH. Recent advances in electronic and bioelectronic noses and their biomedical applications. Enzym Microb Technol. 2011;48:427–37.
36. Wilson AD. Future applications of electronic-nose technologies in healthcare and biomedicine. In: Akyar I, editor. Wide spectra of quality control. Rijeka: InTech; 2011.
37. Gong DW, Zhang Y, Qi CL. Localising odour source using multi-robot and anemotaxis-based particle swarm optimisation. IET Control Theory Appl. 2012;6:1661–70.
38. Airado-Rodriguez D, Hoy M, Skaret J, Wold JP. From multispectral imaging of autofluorescence to chemical and sensory images of lipid oxidation in cod caviar paste. Talanta. 2014;122:70–9.
39. Matsuo H, Furusawa Y, Imanishi M, Uchida S, Hayashi K. Optical odor imaging by fluorescence probes. J Rob Mechatronics. 2012;24:47–54.
40. Liu C, Yokoyama R, Uchida S, Nakano K, Hayashi K. Odor spatial distribution visualized by a fluorescent imaging sensor. In: Proceedings of 2013 IEEE Sensors; 2013. pp. 1506–1509.
41. Chen B, Mokume M, Liu C, Hayashi K. Structure and localized surface plasmon tuning of sputtered Au nano-islands through thermal annealing. Vacuum. 2014;110:94–101.
42. Chen B, Ota M, Mokume M, Liu C, Hayashi K. High-speed gas sensing using localized surface plasmon resonance of sputtered noble metal nanoparticles. IEEE Trans Sens Micromach. 2013;133:90–5.
43. Chen B, Liu C, Ota M, Hayashi K. Terpene detection based on localized surface plasma resonance of thiolate-modified Au nanoparticles. IEEE Sens J. 2013;13:1307–14.
44. Chen B, Liu C, Watanabe M, Hayashi K. Layer-by-layer structured AuNP sensors for terpene vapor detection. IEEE Sens J. 2013;13:4212–9.
45. Chen B, Liu C, Sun X, Hayashi K. Molecularly imprinted polymer coated Au nanoparticle sensor for α-pinene vapor detection. In: Proceedings of 2013 IEEE Sensors; 2013. pp. 117–120.
46. Prada P, Furton K. Human scent detection: a review of its developments and forensic applications. Revista de Ciencias Forenses. 2008;1:81–7.
47. Moser E, McCulloch M. Canine scent detection of human cancers: a review of methods and accuracy. J Vet Behav Clin Appl Res. 2010;5:145–52.
48. Liu CJ, Furusawa Y, Hayashi K. Development of a fluorescent imaging sensor for the detection of human body sweat odor. Sens Actuators B. 2013;183:117–23.
49. Liu C, Hayashi K. Visualization of controlled fragrance release from cyclodextrin inclusion complexes by fluorescence imaging. Flavour Fragrance J. 2014;29:356–63.

Part IV
Energy Harvesting

Energy Harvesting with Supercapacitor-Based Energy Storage

Sehwan Kim and Pai H. Chou

Abstract Harvesting energy from the environment is a desirable and increasingly important capability in several emerging applications of smart sensing systems. Due to the low-power characteristics of many smart-sensor systems, their energy harvesting systems (EHS) can achieve high efficiency by emphasizing low overhead in maximum power point tracking (MPPT) and the use of supercapacitors as a promising type of energy storage elements (ESE). Considerations in designing efficient charging circuitry for supercapacitors include leakage, residual energy, topology, energy density, and charge redistribution. This chapter first reviews ambient energy sources and their energy transducers for harvesting, followed by descriptions harvesters with low-overhead efficient charging circuitry and supercapacitor-based storage.

1 Introduction

Energy-harvesting smart sensing systems have been receiving growing attention in recent years. Smart sensing systems are those with autonomous control, communication, computation, and storage capabilities and are now used in a wide range of applications from wearable to environmental monitoring. Miniaturization, wireless communication, and high-capacity data storage capabilities open up new application domains by enabling a complete system to be mounted on or implanted inside many more physical objects than ever before. However, batteries need to be replaced or recharged, and they are often the most expensive part of the system. Although wireless communication makes it more flexible to deploy smart sensing systems at scale and can save expensive wiring cost, battery replacement can be even more

S. Kim
Dankook University, 119, Dandae-ro, Dongnam-gu, Cheonan-si,
Chungnam, Republic of Korea
e-mail: paul.kim@dankook.ac.kr

P.H. Chou (✉)
University of California, Irvine CA, 92697-2625 USA,

National Tsing Hua University, Hsinchu, Taiwan
e-mail: phchou@uci.edu

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_10

costly if not prohibitive if the sensing nodes are deeply embedded. Utility power is not readily available at many deployment sites or remote locations, and energy harvesting is therefore mandatory in such cases.

However, anecdotes have often been about how costly and bulky these energy harvesters may be, how they fail to sustain days of poor weather, and how their batteries still fail after a year or two. The cost, size, and poor weather sustainability can be addressed by incorporating energy-harvesting circuitry that can extract the maximum amount of power from an energy transducer such as a solar panel over a wide range of supply conditions with low overhead. Several recent features that distinguish such harvesters from their utility-grade, larger counterparts include emphasis on *low overhead* in maximum power point tracking (MPPT) or maximum power transfer tracking (MPTT), and the use of *supercapacitors* as a potential type of energy storage elements (ESE) to address the problem of battery aging [4, 6, 14, 17, 37, 42].

Supercapacitors, also known as ultracapacitors or electrochemical double layer capacitors (EDLCs), have long life cycles and have been identified as a promising type of ESE for smart sensor nodes. In particular, supercapacitors and photovoltaic (PV) modules make an excellent combination for energy harvesters. This has motivated researchers to design efficient charging circuits for supercapacitors in their sensing systems. Supercapacitors have lower *energy density* than batteries do by an order of magnitude but much higher *power density*, which enables their use in applications that require short-term high power draw, such as electric vehicles and medical equipment [5, 32]. In particular, despite the lower energy density, their very long life cycles make them suitable for use as ESE for energy harvesting systems (EHS) [4, 6, 16, 37]. Such a system usually consists of the following four components: the energy transducers, (e.g., solar, wind, vibration etc.), energy-harvesting circuitry, energy storage subsystem, and target load, as shown in Fig. 1.

The main issues with EHSs for smart sensing systems are constraints on the form factor, harvesting efficiency, low-overhead harvesting circuitry, scalability to multiple reservoirs, and cold booting control. To solve these issues, each of the subsystems may be optimized in isolation, but together they can be jointly optimized with interesting trade-offs. Therefore, it is critical to devise (1) charging circuits to maximize harvesting efficiency and (2) the circuits to automatically find the maximum power point (MPP). Furthermore, supercapacitors-based energy storage subsystem should consider the nonlinearity of supercapacitors such as leakage, residual energy, topology, energy density, and charge redistribution to charge the supercapacitors efficiently. As a result, supercapacitor-based energy-harvesting smart sensing systems can lead to several benefits including cost effectiveness, small form factor, and long operating lifetime.

The chapter is organized as follows: In Sect. 2, energy transducers are modeled with an examination of their MPP. Section 3 describes techniques for maximizing the efficiency of energy harvesters, and Sect. 4 takes a close look at supercapacitor-based energy storage subsystems. Finally, Sect. 5 presents a summary of this chapter.

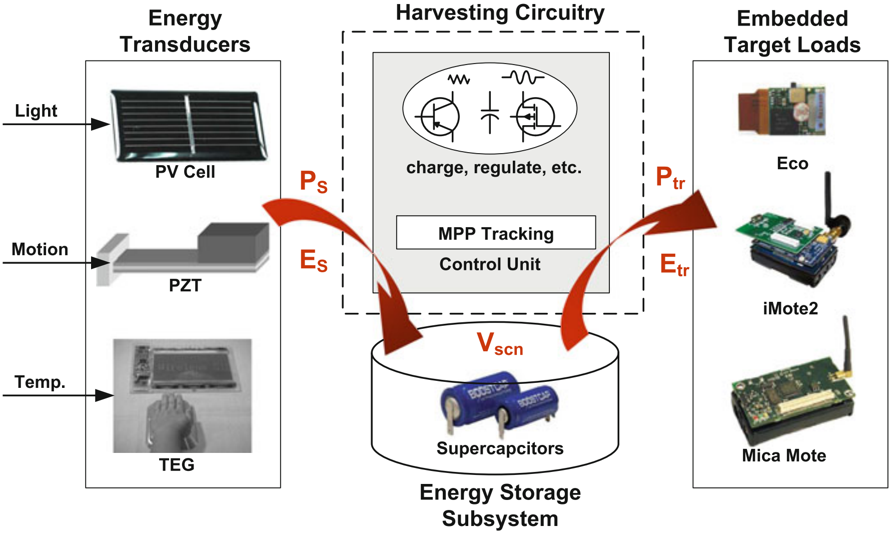


Fig. 1 The block diagram for EHS, powering smart sensing systems: Eco [31], DuraMote [18], and Mini-FDPM [32]

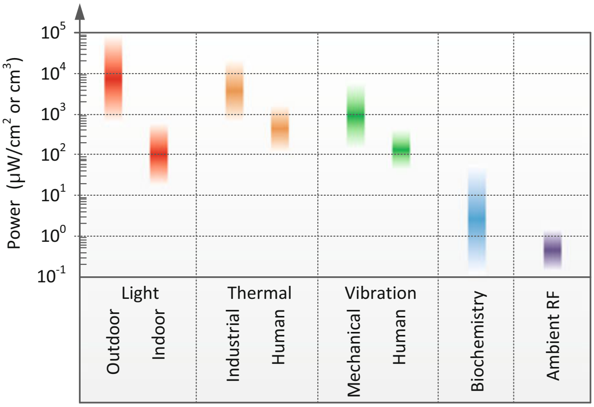


Fig. 2 Ambient sources power densities

2 Energy Transducers

Ambient energy sources are often available in the surroundings of most deployment sites. Examples of such energy sources include mechanical (vibrations, deformations), thermal (temperature gradients or variations), radiant (sun, infrared, RF), and chemical energy (chemistry, biochemistry) sources. They are characterized by different power densities as shown in Fig. 2. Energy harvesting from the sun is the most powerful but is not always available or efficient under low solar irradiation conditions such as poor weather or dark places. Similarly, it is not possible to harvest energy from thermal sources where there is no thermal gradient or to harvest vibration energy where there is no vibration. As a consequence, the source of ambient energy must be chosen according to the deployed environment of the smart sensor nodes.

Each given source of ambient energy can be converted by a different energy transducer that performs conversion to electrical energy. To the system, the key differences of energy transducers are the output power level, alternating current (AC) vs. direct current (DC), the dynamic range, and the impedance model. For instance, Fig. 2 shows that $10\sim100$ μW of available output power level is a good order of magnitude for a 1 cm^2 or a 1 cm^3 energy transducer. Although $10\sim100\mu$W may not be a great amount of power, it can be enough for many applications of smart sensor systems. Of the different types of transducers, windmills [22], magnetic coil generators [24], piezoelectric generators [2, 28, 35], and magnetic induction [7] output AC power, whereas thermal [1] and photovoltaic [14, 34, 37] power sources output DC power.

Maximum power point tracking (MPPT) refers to drawing power from these energy transducers at a level that maximizes the power output. Furthermore, maximum power transfer tracking (MPTT) attempts to find actual MPP by considering the efficiency of harvesting circuitry. Therefore, although there has been extensive research from the device perspective to improve the cost, conversion efficiency, and power density of transducers, it is crucial for system designers to understand their electrical characteristics in-depth in order to analyze their impact on the system being powered. Therefore, in this section, the maximum power point (MPP) of each energy transducer depending on the output power is described with the equivalent circuit models of the energy transducers. The electrical equivalent-circuit model is important for validation by simulation and optimization of the harvester at the system level.

2.1 Direct Current Electricity

This section introduces photovoltaic cells and thermoelectric generators as the representative types of DC output energy transducers by describing its characteristics through the equivalent circuit models.

2.1.1 Photovoltaic Cells

Photovoltaic (PV) cells are the most well-known type of DC-power source for its high power density. Figure 3a shows the equivalent circuit of a solar cell, which can be modeled as a current source with a voltage limiter. Based on the circuit, the output current (I_{pv}) of a solar cell can be expressed as:

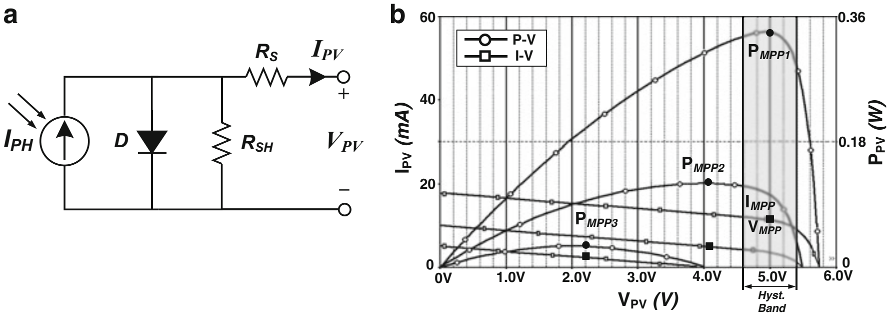


Fig. 3 The characteristics of a solar cell. (**a**) Solar cell circuit model. (**b**) I-V / P-V characteristics and hysteresis window of MPP

$$I_{\mathrm{pv}} = I_{\mathrm{ph}} - I_{\mathrm{o}}\left(e^{q\frac{V_{\mathrm{pv}}+I_{\mathrm{pv}}R_{\mathrm{s}}}{A\,k\,T_{\mathrm{STC}}}} - 1\right) - \frac{V_{\mathrm{pv}} + I_{\mathrm{pv}}\,R_{\mathrm{s}}}{R_{\mathrm{sh}}} \tag{1}$$

where I_{o} is the reverse saturation current, q is the electron charge, A is the diode quality factor, k is the Boltzmann constant, T_{STC} is the operating temperature at Standard Test Conditions (STC), for which the irradiation is 1,000 W/m^2 and the panel temperature (T_C) is 25 °C, and R_{s} and R_{sh} are the panel series resistance and panel shunt resistance, respectively.

Figure 3b shows the simulation result for a solar panel using Eq. (1). The panel parameters were extracted by previous work [36]. The power output of the solar panel is not constant but has a wide dynamic range due to the many intensity levels of sunlight. Moreover, the P_{MPP3} indicates the P-V curve under the low solar irradiation condition in the morning, in the evening, or on cloudy days. In such conditions, the solar cell's voltage may be too low to operate the target system or charge up the ESE, and therefore this energy can be wasted, unless the voltage is boosted up.

2.1.2 Thermoelectric Generators

Thermoelectric generators (TEGs) convert geothermal energy into electrical energy by the Seebeck effect. The TEG has not become widespread yet because of the low conversion efficiency. Despite the low efficiency of the TEGs, the prospect of thermoelectric power generation has rapidly become very promising with the growing public interest in environmental problems in recent years. By using differing combinations of series and parallel connections of the junction pairs, the output voltage and current of the harvester can be adjusted. Typically, a series connection is used to maximize the output voltage, at the expense of current, to reach a usable voltage level at lower temperature gradients.

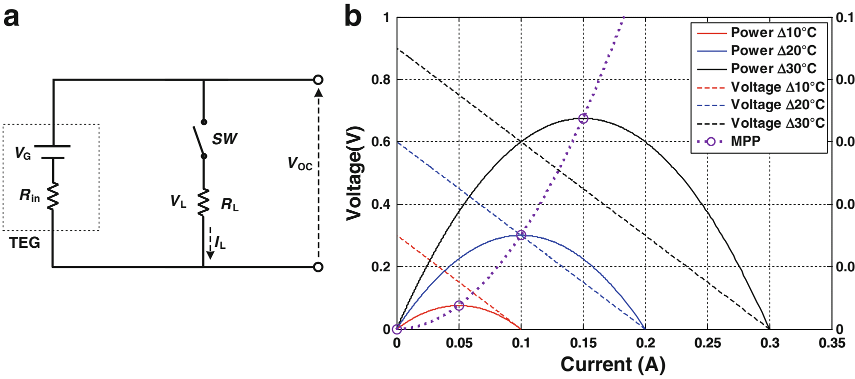


Fig. 4 The characteristics of a TEG. (**a**) TEG circuit model. (**b**) V-I / P-I Characteristics and MPP of a TEG

The basic equivalent circuit of a TEG is illustrated in Fig. 4a, which is modeled the generation voltage source V_G and the internal resistor R_{in}. The open-circuit voltage (V_{oc}) of the TEG is expressed as:

$$V_{oc} = N(\alpha_p - \alpha_n)(T_H - T_C) \tag{2}$$

where α_p and α_n are the Seeback coefficients of n-type and p-type material, respectively. N is the number of thermoelements. ($T_H - T_C$) is temperature difference between hot side and cold side of the TEG. According to Eq. (2), the open circuit voltage (V_{OC}) of the TEG is directly proportional with the number of thermo-elements and the temperature gradients [3].

The output power $P_{out,TEG}$ in this circuit can be given by:

$$P_{out,TEG} = I_L V_L = I_L(\alpha \Delta T - I_L R_{in}) = \alpha^2 \Delta T^2 \frac{R_L}{R_{in} + R_L}, \tag{3}$$

where P_{out} is the output power, I_L is the electric current flowing through the load, V_L is the generated voltage on the load by the TEG, ΔT is the temperature difference between hot side and cold side, R_{in} is the TEG electrical resistance, and R_L is the TEG load resistance.

The power transfer is maximized when the impedance is matched, i.e., the load resistor R_L is equal to the TEG internal resistor R_{in}. This load condition is given by [12, 15]:

$$P_{out,TEG}^{max} = \frac{\alpha^2 \Delta T^2}{4R_{in}}. \tag{4}$$

Using Eqs. (3) and (4), the voltage-current and power characteristics of a TEG can be plotted as shown in Fig. 4b.

When the temperature difference between the surfaces of TEG is changed, the output voltage of the TEG varies accordingly. However, most loads to be powered require a standard supply voltage, which can be produced by DC-DC converters. Otherwise, TEGs are connected in series and in parallel to achieve sufficient power.

2.2 *Alternating Current Electricity*

Mechanical energy is possibly the most prevalent AC-output power source and are found in windmills [22], magnetic coil generators [24], piezoelectric generators [2, 28, 35], and magnetic induction [7], and many more. Among them, in this section, we focus on the vibration-powered generators as AC-output transducers. Vibration energy harvesting is to convert vibrations into electrical power. Actually, turning ambient vibrations into electricity is a two-step conversion process: vibrations are first converted in a relative motion between two elements, thanks to a mass-spring system, that is then converted into electricity by a mechanical-to-electrical converter (e.g., piezoelectric material, magnet-coil, or variable capacitor). As ambient vibrations are generally low in amplitude, when the mass-spring system is in resonance, the relative movement amplitude of the mobile mass is amplified compared to the vibrations amplitude, thereby increasing the harvested power as shown in Fig. 5a. The resonance point can be a maximum power point of the vibration-powered generators.

Figure 5b represents the equivalent model of vibration harvesters. A mass (m) is suspended in a frame by a spring (k) and damped by forces (f_{elec} and f_{mec}). When a vibration occurs $y(t) = Y \sin(\omega t)$, it induces a relative motion of the mobile mass $x(t) = X \sin(\omega t)$ compared to the frame. A part of the kinetic energy of the moving mass is converted into electricity (modeled by an electromechanical force f_{elec}),

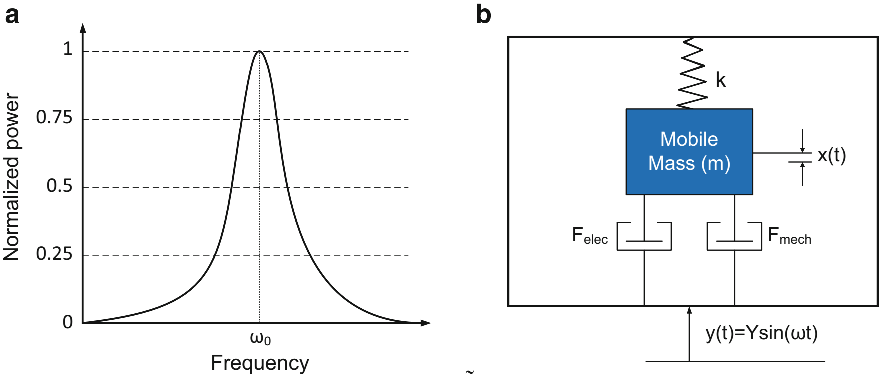


Fig. 5 Vibration energy harvesters. (**a**) Resonance phenomenon. (**b**) Equivalent model

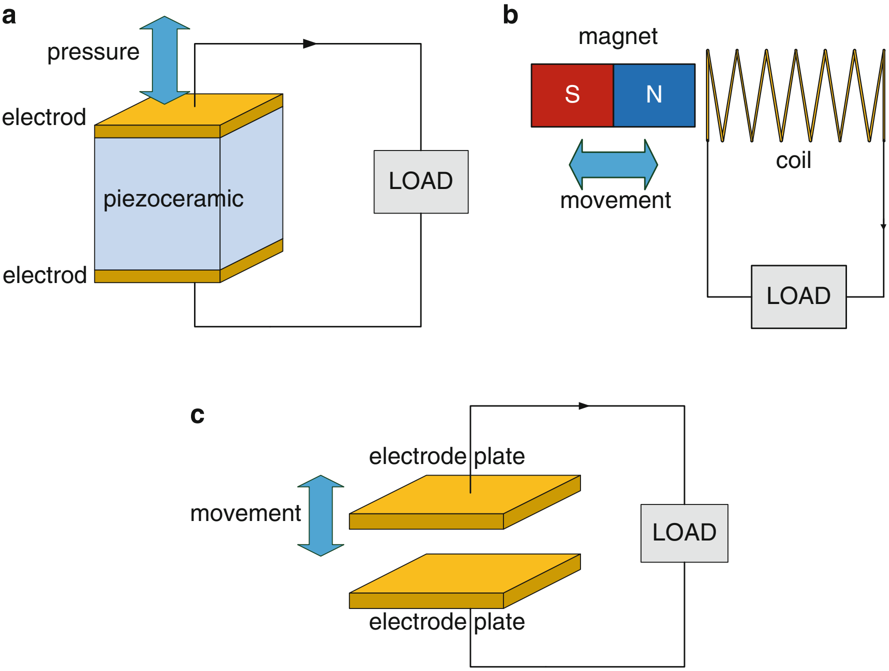


Fig. 6 Mechanical-to-electrical conversion for smart sensor systems. (**a**) Piezoelectric transducer. (**b**) Electromagnetic transducer. (**c**) Electrostatic transducer

while another part is lost in friction forces (modeled by f_{mec}). The maximum output power of a resonant energy harvester submitted to an ambient vibration is reached when the natural angular frequency (ω_0) of the mass-spring system is equal to the angular frequency of ambient vibrations (ω) and when the damping rate $\xi_m = b_m/(2\,m\,\omega_0)$ of the mechanical friction force f_{mec}. The maximum power point can be expressed as

$$P_{\mathrm{out,\,vib}}^{\mathrm{max}} = \frac{m\,Y^2\,\omega_0^3\,Q_m}{8}. \tag{5}$$

To induce this electromechanical force, it is necessary to develop a mechanical-to-electrical converter to extract a part of mechanical energy from the mass and to turn it into electricity [38].

Three main types of devices that convert mechanical energy into electricity are piezoelectric, electromagnetic, and electrostatic generators, as shown in Fig. 6. Piezoelectric ones employ active materials that generate a charge when mechanically stressed or strained. Electromagnetic generators are based on electromagnetic induction arising from the relative motion between a magnetic flux gradient and a conductor. Electrostatic generators use a variable capacitor structure to generate charges from a relative motion between two plates. Detailed explanations for the three main types of devices are as follows.

2.2.1 Piezoelectric Generators

Piezoelectric ceramics have been used for many years to convert mechanical energy into electrical energy. In particular, the use of piezoelectric generators to power human-wearable systems has been extensively studied because of their higher output voltages, high capacitances, and no need to control any gap. Human motion can be characterized by large-amplitude movements at low frequencies, and therefore it is difficult to design a miniature resonant generator to work on humans. Coupling by direct straining of, or impacting on, piezoelectric elements have been applied to human-wearable systems. A subsequent device has been developed [23] by mounting an 8-layer stack of PVDF laminated with electrodes on either side of a 2 mm-thick plastic sheet. This stave was used as an insole in a sports training shoe. At a frequency of a footfall of 0.9 Hz, this arrangement produced an average power of 1.3 mW into a 250 kΩ load.

2.2.2 Electromagnetic Generators

Electromagnetic induction is the generation of electric current in a conductor located within a magnetic field. The conductor typically takes the form of a coil and the electricity is generated by either the relative movement of the magnet and coil or by changes in the magnetic field. One of the most effective methods for energy harvesting is to produce electromagnetic induction by means of permanent magnets, a coil, and a resonating cantilever beam.

Several companies specializing in the field of energy harvesting have emerged over recent years. A shake flashlight or faraday flashlight uses the electromagnetic generator. During shaking, a magnet passes back and forth through a coil of wire and creates an electrical current that is then stored in a supercapacitor or battery. When the flashlight is turned on, the capacitor supplies the stored energy to the bulb. Besides flashlights, micro generating systems have also been developed for watches [20].

2.2.3 Electrostatic Generators

Electrostatic generators are capacitive structures made of two plates separated by air, vacuum, or any dielectric materials. A relative movement between the two plates generates a capacitance variation and then electric charges. These electrostatic generators can be divided into electret-free and electret-based ones. The former uses conversion cycles made of charges and discharges of the capacitor while the latter uses electrets, giving them the ability to directly convert mechanical power into electricity.

An electrostatic generator is well-adapted for size reduction, increasing electric fields, and capacitances. It can also offer the possibility to decouple the mechanical structure and the generator. Finally, it enables development of low-cost devices as they do not need any magnet or potentially expensive piezoelectric material.

3 Techniques for Maximizing Efficiency of Harvesters

Energy harvesters for smart sensing systems can achieve efficient operations through tracking the maximum power point. This section first provides the basic principle of MPPT on energy transducers, specifically solar cells, in terms of an equivalent circuit model. Next, it presents approaches to MPPT and describes techniques for maximizing the efficiency of energy harvesters at subwatt scale, followed by a comparison of MPTT issues.

3.1 Maximum Power Point Tracking

The MPP of an energy transducer changes with the strength of the ambient energy source as shown in Fig. 3b. The purpose of *maximum power point tracking* (MPPT) is to track the supply condition and determine the corresponding load that maximizes the transferred power, namely operating at the MPP. The most significant design consideration for MPP tracking in subwatt-harvesters is to ensure that MPP tracking incurs minimal power overhead, as the output power from the energy transducer is very limited.

3.1.1 Graphical Load-Line Analysis

A solar panel consists of a matrix of solar cells, also known as photovoltaic (PV) cells. This section explains how a solar panel works in terms of a circuit model. Figure 7a shows the equivalent circuit model of a solar cell. It can be described as one ideal current source and a voltage limiter, as shown in Fig. 7a, where I_O is proportional to the sunlight intensity. Therefore, one of the most important issues for a solar cell is how to efficiently deliver as much power to the load (represented by R_L) as possible for a given I_O, as determined by a given level of sunlight intensity.

Figure 7b illustrates MPPT by graphical load-line (represented by R_{L}) analysis, with the assumption that I_{O} increases gradually. When I_{O} is small, most I_{O} will flow to R_{L}, because the diode does not turn on before reaching 0.7 V. As I_{O} increases, V_{L} will eventually approach 0.7 V, and the diode turns on. As a result, any additional increase of I_{O} will result in current flowing to the diode instead of the load. Thus, at high I_{O}, V_{L} is approximately 0.7 V and I_{L} is saturated at $0.7/R_{\mathrm{L}}$. Therefore,

$$I_{\mathrm{S}} = I_{\mathrm{O}} - I_{\mathrm{D}} \tag{6}$$

$$I_{\mathrm{L}} = V/R_{\mathrm{L}} \tag{7}$$

The solution for $I_{\mathrm{S}} = I_{\mathrm{L}}$ and V can be found by plotting I_{S} and I_{L} separately vs. V as shown in Fig. 7b. By graphical load-line analysis, the solution for $I_{\mathrm{S}} = I_{\mathrm{L}}$ and V changes from (1) to (2) and (3) as I_{O} increases. After approaching the point

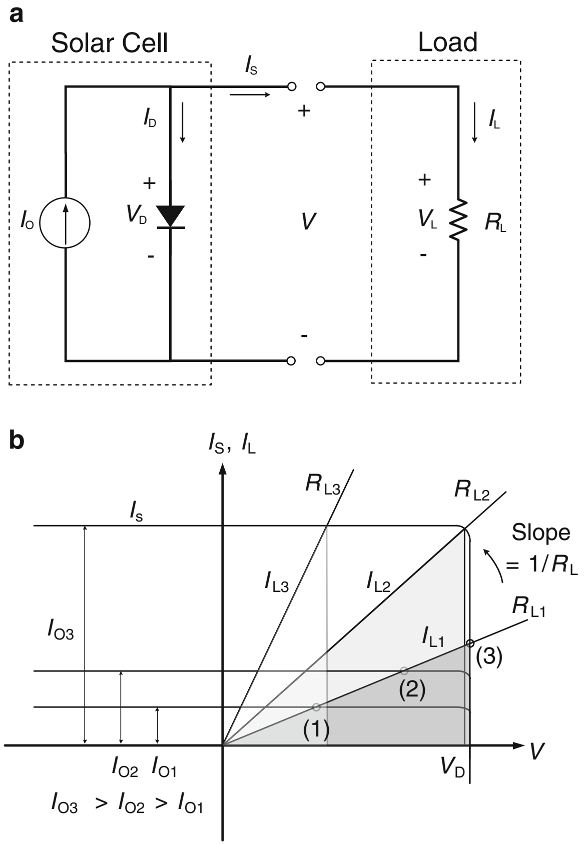


Fig. 7 The basic of MPPT at a solar cell. (**a**) Equivalent circuit model of solar cell. (**b**) I-V curve and load lines of a solar cell

(3), any further increment in I_O will not affect the power conversion efficiency. At this point, one can increase the power conversion efficiency only by lowering R_L, because the slope of the load line is inversely proportional to R_L. In detail, the shaded area of Fig. 7b is equal to harvested power that is transferred to the load. Comparing the three load-resistor values R_{L1}, R_{L2}, and R_{L3}, R_{L2} results in the maximum power conversion when the "diode" is just turned on. This analysis result shows that adjusting the slope of the load line is the pivotal parameter for transferring the maximum power from the solar cell to the load. The saturation voltage V_D can be increased beyond 0.7 V by series and parallel compositions of the solar cells.

Since MPPT overhead is potentially high, many energy harvesters for smart sensor systems have not performed MPPT until recent years. These MPP trackers must consider the net amount of power that can be transferred, i.e., after the MPPT overhead has been subtracted. One common approach is to sacrifice MPPT optimality for significantly reduced overhead. That is, by harvesting within, say, 5–10 % from the MPP, one may cut down on the MPPT overhead significantly, which may result in much higher net power. One way to classify MPPT approaches is consumption side vs. supply side. Consumption side is represented by load matching, while supply side is further divided into sensor-driven and perturbation-based MPPT.

3.1.2 Load Matching

A consumption-side subwatt-solar MPPT approach is called *load matching*, which means to adjust the load directly to maximize the utility of power when available. The load can be adjusted by duty cycling or *dynamic power management* (DPM), among many techniques published in the low-power literature. One reason for maximizing power utility is to minimize power loss due to conversion and energy loss due to storage [9, 29], although one can always store the excess power as yet another form of load.

Actually, load matching is a special case of *load following*, where the duty cycling [11] or DPM [42] tracks the level of available power (e.g., based on a light sensor) without necessarily tracking the MPP (i.e., transferred power). Because load following does not necessarily track the MPP, it can actually lead to system failure if there is no energy storage, because overloading the solar panel will result in lower transferred power than the peak load. Another consideration is that both load matching and load following tend to be application specific.

3.1.3 Sensor-Driven MPPT

With sensor-driven MPPT, a sensor is used to measure the intensity of the ambient power, which is the primary parameter that determines the MPP. For instance, the MPP for a solar panel is primarily determined by the light intensity, and the MPP for a wind generator is primarily determined by the rotational speed of the fan. The sensor value can then be used to determine the load that will result in the MPP. The use of a sensor does not require perturbation to the energy harvesting source and enables very simple circuitry to be built, such as the case with AmbiMax [30]. In fact, AmbiMax can also take a rotational speed sensor for a wind generator. However, a sensor itself may consume power. One alternative that addresses this problem, at least for the solar part, is to use a *pilot cell*, which is a miniature PV cell that outputs its harvested power instead of consuming power [4]. A small pilot cell can be made in about the same size as a photo sensor.

In both cases, however, under partial shading conditions, either the photo sensor or the pilot cell may fail to output a representative value for the solar panel's exposure to solar power. A related problem is aging and other forms of deterioration, where even without partial shading, the photo sensor or the pilot cell's output is no longer a good indicator of the MPP. In the latter cases, the energy harvesting system would need to be re-calibrated.

3.1.4 Perturbation-Based MPPT

Perturbation-based MPPT [13, 25, 39] approaches do not rely on sensors to measure the ambient power level in order to derive the MPP; instead, they test the generator itself to determine the MPP. Such MPPT approaches include open-circuit voltage, short-circuit current, hill climbing, and I-V curve sweeping.

Open-circuit voltage (V_{oc}) and *short-circuit current* (I_{sc}) approaches use V_{oc} and I_{sc} to determine the ambient power level, respectively [27]. This can be viewed as using the entire solar panel as a sensor. However, the price to pay is that it requires the load to be disconnected momentarily while the V_{oc} or I_{sc} is measured. One may approximate the V_{mp} or I_{mp} as a linear function of V_{oc} or I_{sc}, respectively:

$$V_{\mathrm{mp}} = k_1 V_{oc} \tag{8}$$

$$I_{\mathrm{mp}} = k_2 I_{sc} \tag{9}$$

where k_1, k_2 are proportional constants.

Once k_1 and k_2 are known, I_{sc} can be simply measured by shorting the solar cell using periodic switching operation, and V_{oc} also can be measured periodically by momentarily shutting down the power converter. To handle the temporary loss of power while measuring V_{oc} or I_{sc}, some circuitry such as a capacitor is required to power the rest of the system briefly. The accuracy of V_{mp} and I_{mp} depends on the constants k_1, k_2.

A more accurate method is called hill-climbing, where perturbing the duty ratio of the power converter perturbs the solar panel's current and consequently perturbs the solar panel's voltage [21]. On the P-V curve, if the output power increases in response to increasing (or decreasing) duty ratio, then the MPP tracker continues generating perturbation with incremented (or decremented) duty ratio; but if the power decreases, then the subsequent perturbation should be reversed. The system then oscillates about the MPP. The oscillation can be minimized by reducing the perturbation step size. However, a smaller perturbation size slows down the convergence speed of MPPT. In this sense, even though hill-climbing is more accurate, it has the drawback that the convergence speed is unstable depending on the perturbation step size. Hill-climbing tracks two points in order to find the MPP, and thus it consumes more power than V_{oc} or I_{sc} method.

I-V curve sweeping is an even more precise MPPT approach, which measures the I-V characteristic of the solar panel by varying a test load. Note that all other approaches also need to rely on some characterization of the solar panel, also done by sweeping the I-V curve, except that they are done before deployment. The advantages to doing I-V curve sweeping at runtime are that (1) it tracks the exact characteristics of the solar panel over time, even as it ages or becomes dusty, (2) the same MPPT logic can work with a wide range of replacement solar panels automatically, without requiring the user to manually characterize each one and updating the control parameters. As with other perturbation-based MPPT techniques, I-V curve sweeping at runtime also requires the system to be disconnected from the solar panel momentarily. It may incur slightly more cost, but in practice, the cost is only slightly higher than other simpler perturbation-based MPPT approaches.

3.2 Maximum Power Transfer Tracking

The MPP found by conventional trackers that consider only the energy transducer can be quite different from the MPP at the *system level* when taking the efficiency of the charging circuitry into account. To address this problem, maximum power transfer tracking (MPTT) has been proposed.

3.2.1 The Impact of Charging-Circuit Efficiency on MPP

To maximize the amount of energy stored in the supercapacitor, we should maximize $P_{\mathrm{charge}} = \eta \cdot P_{\mathrm{pv}}$. Indeed, conventional MPPT methods maximize P_{pv} with the expectation that P_{charge} is also maximized. However, MPPT without considering the charger efficiency may result in low transferred power, as shown in Fig. 8a. Because the charging circuit efficiency η is varying depending on output current (I_{out}), conventional MPPT techniques cannot guarantee that the maximum P_{ambi} is equal to the maximum P_{scap}, unless the charging circuit's efficiency for the EHS is also taken into account.

Figure 8b shows that MPP_{solar} (marked by squares) is shifted to MPP_{scap} (marked by circles) after power passes through the charging circuit of the harvesting system. The amount of shift is linearly proportional to the output current I_{out}. For instance, at 1,000 $\mathrm{W/m^2}$, we can see a significant drift from MPP_{solar} to MPP_{scap}.

3.2.2 Maximum Power Transfer Tracking

To address this problem, maximum power *transfer* tracking (MPTT) was proposed to consider the efficiency of the charging circuits as a function of the load [19]. To realize MPTT, the efficiency range of charging circuits needs to be considered; thus, the MPPT circuitry should be placed right before the supercapacitors to implement a more accurate MPP tracker. A novel charging circuit with MPTT using a charge pump to charge supercapacitors was proposed in [16]. Its four primary tasks are (1) sensing the current from the ambient power source(s) and selecting the suitable input capacitance C_{in}, (2) sweeping the switching frequency and tracking I_{MPP} using the I-F curve, (3) feeding the microcontroller unit (MCU) the maximum power transfer point (MPTP) and reconfiguring the smart switch array to optimize the input capacitors or connect the reservoir supercapacitors in series or as single cells, and (4) charging the selected supercapacitor with I_{MPP}.

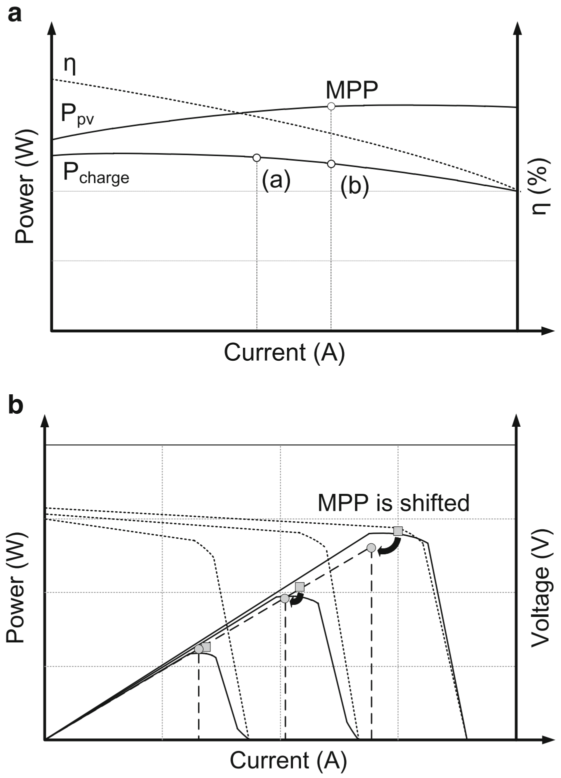


Fig. 8 The shift of MPP at a solar cell depending on solar intensity. (**a**) MPP is shifted due to efficiency (η) varies. $P_{\text{charge}} = \eta \cdot P_{\text{pv}}$. (**b**) MPP_{scap} vs. MPP_{solar}, when $P_{scap} = \eta \cdot P_{abmi}$

3.2.3 MPP Trackers vs. MPTP Trackers

MPPT techniques described above can be realized using commercial off-the-shelf components that supported hysteretic control. For example, one popular component for MPP trackers such as [4, 6, 30] is LTC1440, an ultra-low power comparator from Linear Technology. It supports programmable lower and upper bounds of the hysteresis band as indicated by the shaded region in Fig. 3b. In this way, the actual operating point of the MPP tracker oscillates around the MPP, rather than being a fixed point. By tuning to a narrow hysteresis band, this can lead to higher efficiency of the power converter, but it may not be able to match the dynamic range of the ambient power conditions, making the MPP tracker operate at an inefficient level. Taking PV cells as an example, in the early morning, the solar irradiation intensity can be very low such that the MPP tracker with a narrow tuning hysteresis band fails to properly track the maximum power point as shown in Fig. 3b.

On the other hand, the MPTP tracker in [16] was designed by using a frequency sweeper. Since the frequency sweeping range of a charge pump is directly related to the wide dynamic range of various ambient power sources, a direct digital synthesizer (DDS) was employed to cover the wide dynamic range of the charging circuit. Although it is possible to cover this wide frequency range using multiple analog oscillators, doing so will increase the system complexity and induce higher overall power consumption. For this reason, a DDS can be the best candidate for the charge pump as a substitute for a conventional analog oscillator. A commercially available DDS chip, AD9834, can generate frequencies from 0 to around 40 MHz with power overhead of around 18 mW.

3.2.4 Harvesting Threshold

Although an MPTT charger can convert power efficiently over different levels of ambient power, it is unable to actually charge the ESE when the converted power is below a threshold. Some MPPT and MPTT chargers have relatively high thresholds, because they are unable to cover the wide dynamic range of the ambient power source. Therefore, when below the threshold, the charging efficiency is not the *conversion efficiency* of the circuity, but is 0 % because of zero charging activity. A charger with a higher threshold is said to have a narrower *charging zone* than one with a lower threshold. The overall efficiency of a harvester therefore must consider not only the efficiency of the conversion circuitry but also scale it by the width of its charging zone.

A harvester with high conversion efficiency but a narrow charging zone may be suitable for scenarios with plentiful ambient power. However, it may fail to sustain consecutive days of poor weather or areas with greater seasonal variations of sunlight availability. A common solution to this problem is to *over-design* the system by incorporating a much larger solar panel than necessary, but this can increase the cost significantly. One particular reason for a narrow charging zone in previous designs is the use of buck-type charging circuits [4, 6, 37]. They convert the power from the solar panel down to a lower voltage, and this effectively sets the threshold to a potentially high level (and therefore narrows the charging zone). A buck-boost converter can be used as a general solution, especially for rechargeable batteries. However, they may not be able to handle the wide dynamic range of solar panels. Moreover, both buck and buck-boost converters require the use of an inductor as a low pass filter (LPF), which increases the size of the harvesters, as the inductor tends to be bulky. To address this problem, a charge pump may be a better solution, as it does not require additional inductors and thus can be made smaller, and it also supports a much wider dynamic range [16]. It is also particularly suitable supercapacitor-based storage, which can be charged as long as the charging voltage is higher than the supercapacitor's voltage, rather than requiring a fixed target voltage. This can significantly reduce the harvesting threshold.

4 Energy Storage Subsystems

Batteries are the primary type of power source for smart sensing systems. Among rechargeable batteries, Li-ion and Li-polymer batteries have the highest energy density and high charge-to-discharge efficiency. Charging of a lithium-type battery

Table 1 Comparison between batteries and supercapacitors

	Battery	Supercapacitor
Recharge cycle life time	$< 10^3$ cycles	$> 10^6$ cycles
Self-discharge rate	5 %	30 %
Voltage	3.7 V–4.2 V	0 V–2.7 V
Energy density (Wh/kg)	High (20–150)	Low (0.8–10)
Power density (W/kg)	Low (50–300)	High (500–400)
Fastest charging time	Hours	Sec ∼ min
Fastest discharging time	0.3∼3 h	< a few min
Charging circuit	Complex	Simple

is more complicated and is usually handled by a charging IC. Several works cited this reason and chose nickel metal hydride (NiMH) batteries instead. NiMH is one of the most popular types of energy storage for its relatively high energy density and relatively simple charging method, i.e., trickle charging. Nickel-Cadmium (NiCd) batteries have the advantage of higher discharge rates and can tolerate deeper discharge cycles than lithium batteries can. However, in practice, they can suffer from the *memory effect*, or an apparent loss of capacity if it is recharged before being fully discharged. Rechargeable batteries also have a limited number of recharge cycles on the order of 1,000.

Thanks to long charging-discharging life cycles, supercapacitors have been receiving growing interest as energy storage in addition to or instead of batteries in a new generation of energy harvesters. Table 1 shows a comparison between batteries and supercapacitors. Although its capacity is still much smaller than other types of batteries, a supercapacitor can store enough energy to power many smart sensor systems. In particular, its relatively high maximum recharging cycle life time allows it to be used for long-lifetime applications.

4.1 Supercapacitors in Sub-Watt Energy Harvesters

Supercapacitors have high power density, but they cannot be used as drop-in replacements for batteries without considering their intrinsic characteristics such as the discharge (voltage-vs.-energy) curve, leakage, charge redistribution, residual energy, energy density, topology, and cold booting. Their voltage depends on the amount of stored energy, and they work as a virtual short circuit during charging phase. Moreover, supercapacitors have higher leakage current than rechargeable batteries do.

To address the issues and improve the charging efficiency of supercapacitors in sub-watt energy harvesters, researchers have used buck, boost, or buck-boost dc-dc converters in conjunction with control logic circuitry [4, 37, 41]. The control logic circuitry plays a pivotal role in increasing the charging efficiency of supercapacitors

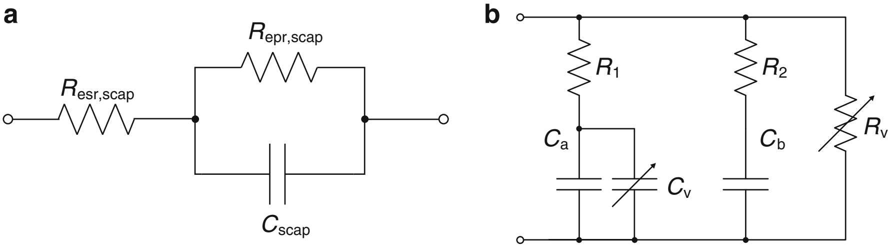


Fig. 9 Equivalent circuit model of supercapacitors. (**a**) R-C circuit model. (**b**) Variable leakage resistance model

by tracking the MPP. Researchers have developed MPP-tracking supercapacitors-charging circuitry using dc-dc converters, and many assume solar sources with high power density under strong sunlight, but it is challenging to charge supercapacitors efficiently with low-overhead MPPT circuitry. Therefore, it is important to first identify the nonideality of supercapacitors and consider it at design time to improve efficiency of supercapacitor charging circuitry.

4.2 *The Characteristics of Supercapacitors*

4.2.1 Equivalent Circuit Model of Supercapacitors

Many circuits-based supercapacitor models have been proposed to simulate the various characteristics of supercapacitors [10, 33, 40]. This section describes two equivalent circuit models: (1) an *R-C equivalent circuit model*, which is suited for the relatively low energy flow and the long time aspects during charging and discharging; (2) the variable leakage resistance model for analysis of charge redistribution of supercapacitors and power management research.

The R-C equivalent circuit model is composed of three components: the equivalent serial resistance $R_{\text{esr, scap}}$, the equivalent parallel resistance $R_{\text{epr, scap}}$, and the capacitor C_{scap}, as shown in Fig. 9a. The $R_{\text{esr, scap}}$ is the internal series resistance, which represents losses in charging or discharging cycles. The $R_{\text{epr, scap}}$ is connected in parallel with the capacitor. The $R_{\text{epr, scap}}$ is used to model the leakage current loss that represents long-term storage characteristics. Differing from the R-C equivalent circuit model, the *variable leakage resistance model* features two resistor-capacitor branches. The capacitor in the first branch includes a constant capacitor C_a and a voltage dependent capacitor C_v. The other branch consists of a constant capacitor C_b. The variable resistor R_v is related to self discharge. The rated voltage V_{nom} denotes the highest voltage to which the supercapacitor can be charged.

4.2.2 Leakage

Given the initial voltage V_0, when the positive and negative terminals of the supercapacitor are opened, the voltage drop due to $R_{\text{epr, scap}}$ is a decay of the initial voltage V_0. Thus,

$$V_{\text{scap}}(t) = V_0 e^{-\frac{t}{R_{\text{epr, scap}} C_{\text{scap}}}} \tag{10}$$

During a particular time, from t_{start} to t_{end}, the leakage energy of the supercapacitor can be expressed as

$$E_{\text{leak, scap}} = 0.5\, C_{\text{scap}}(V^2_{\text{scap}}(t_{\text{start}}) - V^2_{\text{scap}}(t_{\text{end}})) \tag{11}$$

$$= \int_{t_{\text{start}}}^{t_{\text{end}}} \frac{V^2_{\text{scap}}(t)}{R_{\text{epr, scap}}} dt \tag{12}$$

According to Ohm's Law, the leakage current can be written as

$$I_{\text{leak, scap}}(t) = \frac{V_{\text{scap}}(t)}{R_{\text{epr, scap}}} \tag{13}$$

As the capacitance of a supercapacitor increases, its leakage current also increases, while $R_{\text{epr, scap}}$ decreases. In addition, when the voltage of the supercapacitor rises, the leakage current gradually increases; that is, it is proportional to the charged voltage of the supercapacitor. To charge the supercapacitor with a low ambient-power source, the charging power should be higher than the leakage power. Therefore, the power-transfer efficiency of a dc-dc converter and the additional overhead of the control circuit are crucial factors in determining the efficiency of the charger for supercapacitors. Figure 10 shows the characteristic of the leakage current depending on the charged voltage of the supercapacitor. The measured leakage current is marked by asterisks for a 25 F supercapacitor and indicated by squares for an 1 F supercapacitor.

4.2.3 Charge Redistribution

Supercapacitor are made up of two porous electrodes immersed in electrolyte and separated by one porous insulating membrane. Its physical structure increases the farad value as well as the complexity of accurate modeling. They also experience several charge-distribution processes with different time constants, even in isolated and disconnected state. This makes it difficult to identify the process that is responsible for voltage variations. After just being charged for a short period, a disconnected supercapacitor will exhibit a decreasing voltage. This decrease

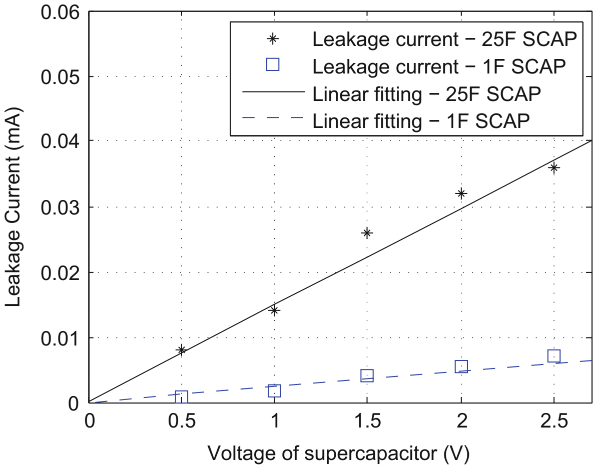


Fig. 10 Leakage current vs. voltage of 25 F and 1 F supercapacitor

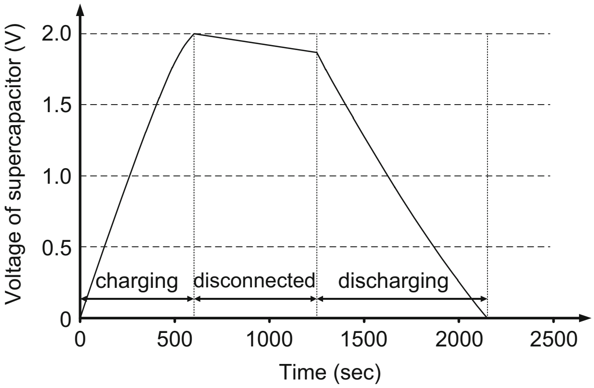


Fig. 11 Voltage decreasing during the delay interval

is mainly caused by a charge distribution within branches. In Fig. 9b, charge redistribution happens when the voltages across C_{sc} (i.e.,$C_{sc} = C_a + C_v$) and C_b do not equal. The dissipated energy due to charge redistribution is mainly consumed by R_2 as long as the charging and discharging current are within a limited range (< 1 A).

$$P_{\text{redist.}} \approx \frac{(V_{C_{sc}} - V_{C_b})}{R_2} \tag{14}$$

From Eq. (14), we can see that the larger difference between $V_{C_{sc}}$ and V_{C_b}, the higher charge redistribution power is. Figure 11 illustrates the internal charge-distribution processes of the supercapacitor. For this, a delay period of 10 min was inserted in between the charging and discharging phases of a 22-F supercapacitor. The capacitor was charged from an initial voltage of 12.5 V, and during the delay, the charged supercapacitor was disconnected. An important phenomenon that accounts for the nonideality of the supercapacitor is the decrease of the supercapacitor voltage during the delay interval. No energy is extracted from the supercapacitor, and still, the voltage decreases. This is mainly caused by the charge-distribution process that, in the short term, is more important than leakage.

4.2.4 Unusable Residual Energy

The unusable energy in the supercapacitor whose voltage is below the minimum dc-dc conversion threshold is called *residual energy*.

E_{residual}: the unusable remaining energy within the supercapacitor.

$$E_{\text{residual}}(t) = \begin{cases} E_{\text{held}}(t) & \text{if } V(t) < V_{\text{conv, min}} \\ \frac{1}{2}C(V_{\text{conv, min}})^2 & \text{if } V(t) \geq V_{\text{conv, min}} \end{cases} \tag{15}$$

where $V_{\text{conv, min}}$ is the minimum conversion voltage of a dc-dc converter or a charge pump. The typical $V_{\text{conv, min}} = 0.7\,\text{V}$, which translates into $E_{\text{residual}} \approx 0.245C$ when the voltage converter is operating.

E_{usable}: the available energy in the supercapacitor for actually driving load.

$$E_{\text{usable}}(t) = E_{\text{held}}(t) - E_{\text{residual}}(t) \tag{16}$$

Most commercially available supercapacitors have a voltage range of 0–2.7 V. For a supercapacitor to supply regulated power (typically 3–5 V) to the load, most subwatt-scale harvesters use a boost-up dc-dc converter. However, the minimum voltage of most commercial off-the-shelf dc-dc converters is 0.7 V (e.g., MAX 1763), below which the converter may work in pass-through mode, but the voltage is still too low to drive the typical load. All supercapacitors can withhold up to $\frac{1}{2} \cdot C \cdot (0.7\text{V})^2$ of residual energy. For example, the residual energy of a harvester using a 300 F supercapacitor is up to 73.5 J. Many subwatt-scale harvesters are designed for wireless sensor nodes that consume 100–150 mW while in active mode: e.g., Mica2 at 3.3 V/16 mA; iMote at 2.5 V/60 mA; Eco node at 3.3 V/30.8 mA. This 73.5 J residual energy can operate Eco node (52.8 mW) for 1392 seconds (23 minutes and 12 seconds) in active mode. This is a considerable amount of unusable residual energy.

4.2.5 Size and Topology

To address residual energy issue, various reconfigurable supercapacitor topologies are attempted for sustainable operation of the target smart sensing systems.

If a single large supercapacitor is employed as the primary energy buffer (reservoir) for sustained operation, then larger capacitance can cause the longer charging time as well as more unusable residual energy. A related problem is *cold booting* [8], the futile cycles of repeated booting and exhaustion while starting a system with little or no *usable* charge (i.e., near the usable threshold) in the supercapacitor, despite the nontrivial *residual* charge in it. To address the issues with the size constraint of the harvester, the topology of ESE needs to be considered at the system level.

Single Supercapacitor Topology

The *single supercapacitor* (SS) topology is the simplest static topology as shown in Fig. 12a. Systems with a *single small supercapacitor* (SSS, 1–5 F) [26, 42] can charge faster but cannot sustain too many days without sunlight. On the other hand, those with a *single large supercapacitor* (SLS, 50–100 F) [4, 37] have a larger amount of storage energy for sustainable operation, but they may have problems with cold booting and take longer charging time. Furthermore, given the same charging-discharging profile, SLS results in a larger amount of residual energy than SSS does.

Reservoir Supercapacitor Array Topology

To address the problems with SS, *reservoir supercapacitors arrays* (RSA) were proposed [6, 30], as shown in Fig. 12b. The purpose of the topology is to shorten the charging time by replacing an SLS with an array of supercapacitors. Since the leakage rate is dependent on the capacitance of supercapacitors, the RSA topology is also helpful in reducing the leakage rate. However, by charging the reservoir supercapacitors sequentially (e.g., [6]) the RSA topology can result in lower energy efficiency of storage. Because the leakage rate of supercapacitors increases rapidly as they approach their maximum rated voltage, one fully charged reservoir supercapacitor would experience the rapid leakage rate while the other reservoir supercapacitor is being charged. One way to address this problem is to keep the supercapacitors voltage-balanced by either *alternating charging* or charging them *in series* [17].

Dynamic Reconfigurable Supercapacitors

According to Fig. 10, the leakage current increases as the terminal voltage approaches the maximum rated voltage of the supercapacitor. Particularly, the leakage power near the maximum rated voltage may be as much as 40 times greater than the lowest leakage power. However, this sharp increase of leakage power near the maximum rated voltage is mitigated along with the lower capacitance of supercapacitors. The SS topology does not have any leakage replenishment

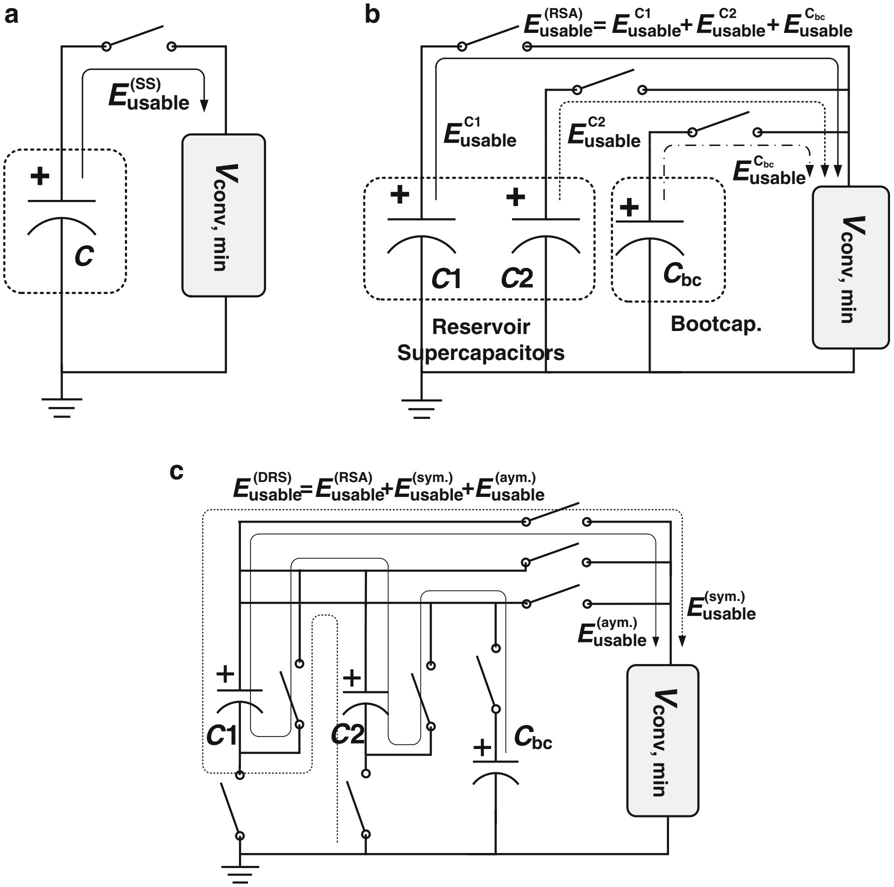


Fig. 12 Supercapacitor topologies for ESE. (**a**) Single supercapacitor topology. (**b**) Reservoir supercapacitor array topology. (**c**) Dynamic reservoir supercapacitor topology

techniques. The RSA topology has better control because it divides the stored energy of an SLS among several supercapacitors, although the leakage sum does not make a great impact on leakage reduction. To address all these issues, the *dynamic reconfigurable supercapacitors* (DRS) scheme was proposed [16]. During charging and discharging phases, the DRS can be configured for different topologies to control the energy leakage, to reduce the unusable residual energy in the supercapacitors, and to improve output-stage efficiency.

4.2.6 Cold Booting

Cold booting, also known as the zero-energy boot-strap problem, is one where the system starts up from nearly fully depleted energy in an energy harvesting system.

A system can enter this state after having been deprived of sunlight for an extended period of time and more sunlight is just becoming available. This is problematic, because if the system starts booting as soon as the harvested power exceeds the usable threshold, it is likely to fail if the harvested power does not increase monotonically at a faster rate than the load. The MCU may boot successfully, but any surge due to RF activities can quickly cause any just harvested stored power to be depleted quickly, too, causing the entire system to fail. Such a system is likely to repeat the futile attempt to boot up until sufficient sunlight is available.

The solutions to cold booting be classified into software control, inhibited start by Schmitt trigger, and bootstrap supercapacitor. Ambimax's [30] feedback regulator used for charging the supercapacitor will respond to inrush current by duty cycling it at a very low rate that is incompatible with MPPT. Everlast [37] addresses the inrush current problem by using a feed-forward regulator instead, but it does not inhibit cold booting. Solar Biscuit [26] and TwinStar [42] use software control and hardware control (Schmitt trigger), respectively, to inhibit futile booting until sufficient energy has been accumulated, but neither performs MPPT, and the inhibiting delay may grow too long for supercapacitors of large capacitance values. DuraCap [6] uses a *bootstrap supercapacitor*, named bootcap (C_{bc}), which has relatively smaller capacitance than the reservoir supercapacitors to solve the cold booting problem by reaching a higher voltage faster with more usable energy for booting. The bootcap has a higher priority to charge, and when it is full, the reservoir ones are charged sequentially (one at a time).

5 Summary

Smart sensing systems have been mainly powered by batteries, but supercapacitors are fast becoming a viable alternative form of energy storage for those smart sensing systems that harvest energy for long-term operation. They overcome the 1–2 year service life of batteries and can deliver high current, but their limitations require solutions at not only the circuit level but also the system level. A complete energy-harvesting system includes an energy transducer, energy harvesting circuitry, an energy storage element, and the smart sensing system as the target load. The state-of-the-art in subwatt-scale harvesters can be characterized by their *low overhead* in maximum power point tracking (MPPT) or maximum power transfer tracking (MPTT), and their use of *supercapacitors* as a potential type of energy storage element. For supercapacitors to replace batteries, the harvester must consider leakage, residual energy, topology, energy density, charge redistribution. Thus, this chapter focuses on describing the impact of the nonlinearities of supercapacitors and the way to compensate or overcome these disadvantages at the system and circuit levels.

References

1. Ahska R, Mamur H. A review: thermoelectric generators in renewable energy. Int J Renew Energ Res (IJRER). 2014;4:128–36.
2. Beeby S, Tudor M, White N. Energy harvesting vibration sources for microsystems applications. J Measure Sci Technol. 2006;17(12):175–96.
3. Bierschenk J. Optimized thermoelectrics for energy harvesting applications. In: Proceedings of the 17th international symposium on the applications of ferroelectrics (ISAF), Santa Re, Feb 23–28, 2008. p. 1–4
4. Brunelli D, Moser C, Thiele L, Benini L. Design of a solar-harvesting circuit for batteryless embedded systems. IEEE Trans Circuits Syst. 2009;56:2519–28.
5. Burke A. Ultracapacitors: why, how, and where is the technology. J Power Sources. 2000;91:37–50.
6. Chen CY, Chou PH. DuraCap: a supercapacitor-based, power-bootstrapping, maximum power point tracking energy-harvesting system. In: Proceedings of the international symposium on low power electronics and design (ISLPED). Austin: ACM; 2010. p. 313–8.
7. Chevalerias O, O'Donnell T, Power D, O'Donovan N, Duffy G, Grant G, O'Mathuna SC. Inductive telemetry of multiple sensor modules. IEEE Pervasive Comput. 2005;4(1):46–52.
8. Chou PH, Kim S. Techniques for maximizing efficiency of solar energy harvesting systems. In: Proceedings of the fifth conference on mobile computing and ubiquitous networking (ICMU 2010), Seattle, WA, USA, 2010. p. 26–8.
9. Chou PH, Li D. Maximizing efficiency of solar-powered systems by load matching. In: Proceedings of the international symposium on low power electronics and design (ISLPED), August 9–11, 2004. p. 162–7.
10. Diab Y, Venet P, Gualous H, Rojat G. Electrical, frequency and thermal measurement and modelling of supercapacitor performance. In: The 3rd european symposium on supercapacitors and applications, Rome, Italy, November 6–7, 2008. p.1066–9
11. Dutta P, Hui J, Jeong J, Kim S, Sharp C, Taneja J, Tolle G, Whitehouse K, Culler D. Trio: Enabling sustainable and scalable outdoor wireless sensor network deployments. In: The fifth international conference on information processing in sensor networks (IPSN/SPOTS), April 19–21, 2006. p. 407–15.
12. Ferrari M, Ferrari V, Guizzetti M, Marioli D, Taroni A. Characterization of thermoelectric modules for powering autonomous sensors. IEEE Trans Instrum Meas. 2009;58:99–107.
13. Hohm D, Ropp M. Comparative study of maximum power point tracking algorithms using an experimental, programmable, maximum power point tracking test bed. In: Conference record of the Twenty-Eighth IEEE photovoltaic specialists conference, September 15–22, 2000. p. 1699–702.
14. Jiang X, Polastre J, Culler D. Perpetual environmentally powered sensor networks. In: Proceedings of fourth international symposium on information processing in sensor networks (ISPN), April 15, 2005. p. 463–8.
15. Kim R, Lai J, York B, Koran A. Analysis and design of maximum power point tracking scheme for thermoelectric battery energy storage system. IEEE Trans Ind Electron. 2009;56: 3709–16.
16. Kim S, Chou PH. Energy harvesting by sweeping voltage-escalated charging of a reconfigurable supercapacitor array. In: Proceedings of the international symposium on low power electronics and design (ISLPED). Fukuoka: ACM; 2011. p. 235–40.
17. Kim S, Chou PH. Size and topology optimization for supercapacitor-based sub-watt energy harvesters. IEEE Trans Power Electron. 2013;28:2068–80.
18. Kim S, Torbol M, Chou PH. Remote structural health monitoring systems for next generation scada. Smart Struct Syst. 2013;11:511–31.

19. Kim Y, Chang N, Wang Y, Pedram M. Maximum power transfer tracking for a photovoltaic-supercapacitor energy system. In: Proceeding of the 16th ACM/IEEE international symposium on low power electronics and design ISLPED. New York: ACM; 2010. p. 307–12.
20. KINETRON: The micro generating system for a watch. http://www.kinetron.eu/wp-content/uploads/2014/04/MGSWatch.pdf
21. Koutroulis E, Kalaitzakis K, Voulgaris, N. Development of a microcontroller-based, photovoltaic maximum power point tracking control system. IEEE Trans Power Electron. 2001;16:46–54.
22. Koutroulis E, Kalaitzakis K. Design of a maximum power tracking system for wind-energy-conversion applications. IEEE Trans Ind Electron. 2006;53(2):486–94.
23. Kymissis J, Kendall C, Paradiso J, Gershenfeld N. Parasitic power harvesting in shoes. In: Proceedings of the 2nd IEEE international conference wearable computing, CA, USA, 1998. p. 132–39.
24. Kymissis J, Kendall C, Paradiso JA, Gershenfeld N. Parasitic power harvesting in shoes. In: Proceedings of the second IEEE international symposium on wearable computers (ISWC). Washington: IEEE Computer Society; 1998. p. 132–39.
25. Lee D, Noh H, Hyun D, Choy I. An improved MPPT converter using current compensation method for small scaled PV-applications. In: The 18th annual IEEE applied power electronics conference and exposition, vol. 1, 2003. p. 540–5.
26. Minami M, Morito T, Morikawa H, Aoyama T. Solar Biscuit: a battery-less wireless sensor network system for environmental monitoring applications. In: The 2nd international workshop on networked sensing systems, 2005.
27. Noguchi T, Togashi S, Nakamoto R. Short-current pulse based adaptive maximum-power-point tracking for photovoltaic power generation system. In: Proceedings of 2000 IEEE international symposium on industrial electronics. vol. 1, 2000. p. 157–62.
28. Ottman GK, Hofmann HF, Bhatt AC, Lesieutre GA. Adaptive piezoelectric energy harvesting circuit for wireless remote power supply. IEEE Trans Power Electron. 2002;17:669–76.
29. Park C, Chou PH. PUMA: Power utility maximization for multiple-supply systems by a load-matching switch. In: Proceedings of international symposium on low power electronic design (ISLPED), August 9–11, 2004. p. 168–73.
30. Park C, Chou PH. AmbiMax: Efficient, autonomous energy harvesting system for multiple-supply wireless sensor nodes. In: Proceedings of 3rd annual IEEE communications society conference on sensor, mesh, and ad hoc communications and networks (SECON), September 25–28, 2006. p. 168–77.
31. Park C, Chou PH. Eco: Ultra-wearable and expandable wireless sensor platform. In: Proceedings of the third international workshop on body sensor networks (BSN 2006). Washington: IEEE Computer Society/Boston: MIT Media Lab; 2006. p. 162–5
32. Park C, No K, Chou PH. TurboCap: Batteryless, supercapacitor-based power supply for Mini-FDPM. In: Proceedings of 3rd european symposium on supercapacitors and applications (ESSCAP), Rome, Italy, November 2008.
33. Petreus D, Moga D, Galatus R, Munteanu RA. Modeling and sizing of supercapacitors. Adv Electr Comput Eng. 2008;8(2):15–22.
34. Raghunathan V, Kansal A, Hsu J, Friedman J, Srivastava M. Design considerations for solar energy harvesting wireless embedded systems. In: Proceedings of the 4th international symposium on information processing in sensor networks (IPSN), April 25–27, 2005. p. 457–62.
35. Roundy S, Wright P. A piezoelectric vibration based generator for wireless electronics. J Smart Mater Struct. 2004;13(5):1131–42.
36. Sera D, Teodorescu R, Rodriguez P. PV panel model based on datasheet values. In: IEEE international symposium on industrial electronics (ISIE), June 4–7, 2007. p. 2392–6.
37. Simjee F, Chou PH. Efficient charging of supercapacitors for extended lifetime of wireless sensor nodes. IEEE Trans Power Electron. 2008;23:1526–36.
38. Williams C, Yates R. Analysis of a micro-electric generator for microsystems. In: Proceedings of eurosensors, 1995. p. 369–72.

39. Xiao W, Dunford W. A modified adaptive hill climbing MPPT method for photovoltaic power systems. In: 2004 35th annual IEEE power electronics specialists conference, vol. 3, June 20–25, 2004. p. 1957–63.
40. Yang H, Zhang Y. Analysis of supercapacitor energy loss for power management in environmentally powered wireless sensor nodes. IEEE Trans Power Electron. 2013;28(11): 5391–403.
41. Zhu GR, Loo KH, Lai YM, Tse CK. Quasi-maximum efficiency point tracking for direct methanol fuel cell in DMFC/supercapacitor hybrid energy system. IEEE Trans Energy Convers. 2012;27(3):561–71.
42. Zhu T, Zhong Z, Gu Y, He T, Zhang ZL. Leakage-aware energy synchronization for wireless sensor networks. In: The 8th annual international conference on mobile systems, applications, and services (MobiSys), June 15–18, 2010. p. 319–32.

Power System Design and Task Scheduling for Photovoltaic Energy Harvesting Based Nonvolatile Sensor Nodes

Yongpan Liu, Huazhong Yang, Yiqun Wang, Cong Wang, Xiao Sheng, Shuangchen Li, Daming Zhang, and Yinan Sun

Abstract This chapter proposes a novel high-efficiency PV power system for nonvolatile sensor nodes. It demonstrates that the storage-less and converter-less system achieves near 90 % energy efficiency by eliminating energy loss from power converters and storage devices. Furthermore, we propose a dual-channel power supply architecture to improve the quality of service when there are time mismatches between harvested energy and workload. A channel controller dynamically selects either direct channel or indirect one to maximize the energy efficiency under failure rate constraints. Both a simulation platform and a prototype are built to validate the architecture. Finally, we presented an intra-task scheduling algorithm for the storage-less and converter-less channel, which leverages neural network training based on solar profiles and task execution. Compared to the inter-task scheduling, it tracks solar variations more quickly and a much lower deadline missing rate is achieved.

1 Introduction

In the world of trillion sensors, battery maintenance becomes a severe problem in both time and costs. There are various energy scavenging technologies to power the sensors from optical, kinetic, thermal, radio frequency energy, and so forth. Among them, photovoltaic (PV) energy is one of the most promising ones, in aspects of power density, efficiency and output voltage and current [1]. Therefore, PV energy harvesting is widely studied for wearable devices [2], body area sensors [3], and structural monitoring [4].

The conventional architecture of PV power systems is shown in Fig. 1. The architecture has three stages; (1) The maximum power point tracking (MPP)

Y. Liu (✉) • H. Yang • Y. Wang • C. Wang • X. Sheng • S. Li • D. Zhang • Y. Sun
Circuit and System Division, Electronic Engineering Department, Tsinghua University,
Rohm Building, Floor 4, 100084 Beijing, P.R. China
e-mail: ypliu@tsinghua.edu.cn

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_11

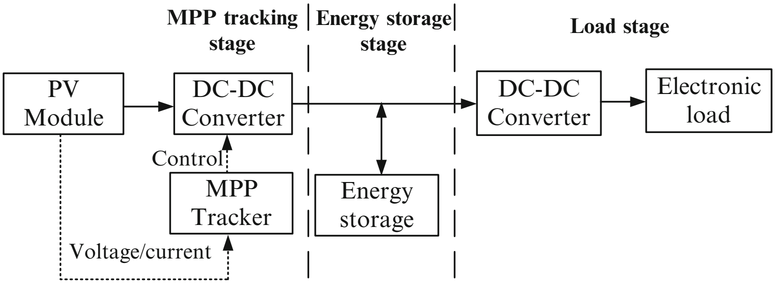


Fig. 1 Typical architecture of a PV power system with MPPT [8]

tracking stage includes an MPP tracker and a DC–DC converter.[1] The MPP tracker tries to extract maximum solar power by monitoring current and voltage of the PV cells and the DC–DC converter is set to either PFM or DPM mode [5, 6]. (2) The second stage performs as a buffer to bridge the energy mismatch between PV cells and workload by capacitor charging or discharging. A supercapacitor is adopted, since it has superior cycle efficiency and lifetime compared to conventional rechargeable batteries [7]. (3) The third stage includes another DC–DC converter as a load regulator to generate a proper output voltage for the load.

However, there are several disadvantages of the architecture, consisting of cascaded DC–DC converters. First, the voltage converters have nontrivial conversion loss. The efficiency of the DC–DC converter varies in a wide rage (40–90 %) when the voltage gap between the PV cell and the supercapacitor changes [9]. In fact, low efficiency frequently occurs because the voltage of the supercapacitor is fluctuant in a wide rage in real applications [10]. Second, the supercapacitor suffers from non-ideal characteristics, such as self-discharge (1 mW for a 10 F supercapacitor at 2 V [11]), charge distribution and non-ideal cycle efficiency [5, 12]. It causes significant energy loss in a milliwatt-level low power sensor node. Third, the inductor based converters and the large size of supercapacitors increase the form factor of the system and lead to the difficulty of on-chip integration and the cost rise. Therefore, a high-efficiency, small volume and low cost PV power system is badly needed for self-powered sensors.

This chapter provides a novel power system with relevant task scheduling for the PV energy harvesting sensor node, which improves both energy efficiency and quality of service (QoS). First, we propose a storage-less and converter-less MPPT architecture. By removing the cascaded DC–DC converters and energy storages in the conventional PV system, the total system efficiency is improved by over 30 % under a wide range of solar irradiance. The MPPT is achieved by performing a fine-grained dynamic power management (DPM) on a nonvolatile microprocessor in the

[1] MPP is used in PV energy harvesting systems to avoid the majority of solar energy dissipated in the PV cells.

sensor node. Furthermore, to extend the availability of the power system to night, we combine the conventional PV system and the storage-less and converter-less PV system as a dual-channel PV power system. An intelligent strategy is proposed to switch the power system between the conventional power channel (indirect channel) and the converter-less channel (direct channel). The dual-channel system achieves both high energy efficiency and low system failure rate. Finally, we implement a QoS-aware task scheduling for the proposed power system by considering the characteristics of both channels.

There are several design challenges to make the proposed PV power system happen: First, the removal of DC–DC converters and energy storage in the PV power system deteriorates the system resilience to power variations. Second, the switching strategy between the direct channel and indirect channel of the power supply is not straightforward. It should maximize the system efficiency and QoS under reasonable switching overhead. Third, task scheduling for the conventional power system is not suitable for the storage-less and converter-less architecture because of the extreme fine-grained DPM.

In the following of the chapter, we overcome above challenges in details. The chapter is organized as follows: Sect. 2 describes the overall architecture of the proposed PV power system and the energy harvesting sensor node. Section 3 presents the storage-less and converter-less MPPT architecture. Section 4 discusses the design of the dual-channel power system and Sect. 5 illustrates the intra-task scheduling for the proposed PV power system. Section 6 summarizes the related works and Sect. 7 concludes the chapter.

2 Overall Architecture

This section describes the overall architecture of the proposed energy harvesting sensor node. The architecture is shown in Fig. 2, which includes two main parts: the power supply system and the nonvolatile sensor node.

The sensor node is constructed by several nonvolatile function units (NVFUs), such as nonvolatile microprocessor, nonvolatile accelerators, etc. Its nonvolatile property makes the extreme fine-grained DPM possible.

The power supply system is a combination of the conventional power channel (indirect channel) and the converter-less channel (direct channel), with several configurable switches. The direct channel connects the solar panel and the sensor node directly via a switch, which is detailed in Sect. 3. The indirect channel consists of an input charger, an output regulator, a supercapacitor, and two switches. When the solar power is sufficient, the indirect channel can store excessive solar energy in the supercapacitor. If the solar power decreases or disappears, the indirect channel provides supplementary energy from the supercapacitor to meet QoS requirements.

The power control is done by the power management unit (PMU). By monitoring the solar power and task executing status, it turns on/off the switches to configure the power channel and the power level of the NV node dynamically. Furthermore,

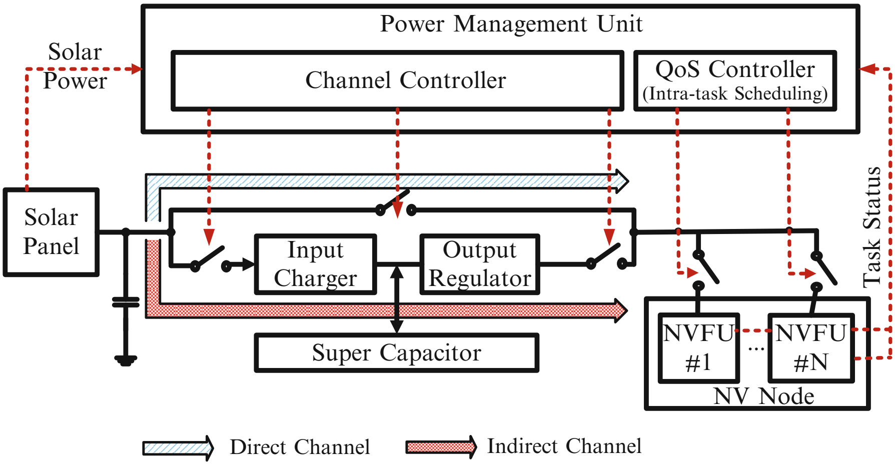


Fig. 2 Architecture of the proposed PV energy harvesting sensor node

the QoS controller executes an intra-task scheduling algorithm to match the load with the solar power under QoS constraints. The detailed algorithms of the channel controller and QoS controller are discussed in Sects. 4 and 5.

3 Converter-Less and Storage-Less Photovoltaic Energy Harvesting

This section provides the converter-less and storage-less photovoltaic power system. Compared with the conventional PV power system, the proposed system removes the DC–DC converters and energy storages to reduce the energy loss and improve the overall system efficiency. Thanks to the development of nonvolatile microprocessors, we achieve a millisecond fine-grained DPM to perform MPPT. The DC–DC converter is replaced by a DPM controlled switch to maintain the MPP of the PV module, and regulate the voltage to the load. Fast DPM makes the output of the PV module swing in a very small range around MPP to achieve high energy efficiency.

3.1 Storage-Less and Converter-Less PV Power System

3.1.1 System Architecture

The architecture of the storage-less and converter-less PV power system is shown in Fig. 3. There are four major components: the PV module, the electronic load, the load switch, and the MPPT controller. The PV module is carefully selected so that

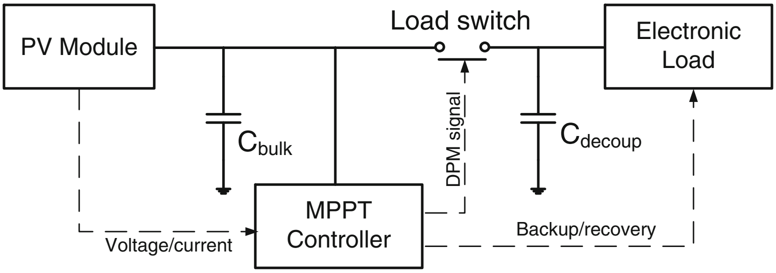


Fig. 3 Proposed converter-less PV system architecture with MPPT

its MPP voltage is close to the legal operating voltage of the electronic load. It is easy to customize MPP voltage by stacking a different number of low voltage PV cells in serial [13]. On the load side, we use a nonvolatile microprocessor based electronic load, which has microseconds sleep and wake-up time and very low energy overheads. The PV module and the electronic load are connected via a load switch. The MPPT controller is responsible to (1) track the MPP of PV module according to its voltage and current by providing DPM signal to the load switch (2) control the backup and recovery of the electronic load. The MPPT controller is implemented with discrete circuit components or micro-controller.

Additionally, there is a bulk capacitor C_{bulk} connected in parallel with the PV module. It extends the time constant of the PV module so that the time constant may match with the feasible DPM period of the load devices. There is a decoupling capacitor C_{decoup} on the load side, which maintains the power integrity of the electronic load. Because the nonvolatile processor based electronic load has fast and energy efficient backup and recovery mechanism, there is no need to store much energy on C_{bulk} and C_{decoup}. Therefore, the size of the capacitors can be small enough to support fine-grained DPM and reduce the self-leakage. In the following part, the working sequence of the proposed system is described in detail.

3.1.2 MPPT with Fine-Grain Dynamic Power Management

The proposed PV power system achieves MPPT with fine-grained dynamic power management (DPM) of the electronic load. The MPPT controller provides a DPM signal to continuously turn on and off the load switch. It forces the PV module output voltage V_{pv} to fluctuate around the MPP voltage V_{mpp} by modulating the DPM signal (see in Fig. 4a). The waveform of V_{pv} in one DPM cycle is illustrated in Fig. 4b. V_{pv} is maintained within a small window, $[V_l, V_h]$, around V_{mpp}. The system working sequence in one DPM cycle is mainly departed to two stages: the "OFF" stage and the "ON" stage. The MPP controller turns off the load switch to shut down the electronic load, once V_{pv} becomes lower than V_l, the system enters the "OFF" stage. After that, the PV module charges the bulk capacitor and V_{pv} begins to rise. After V_{pv} reaches to V_h, the system enters the "ON" stage. The MPPT controller

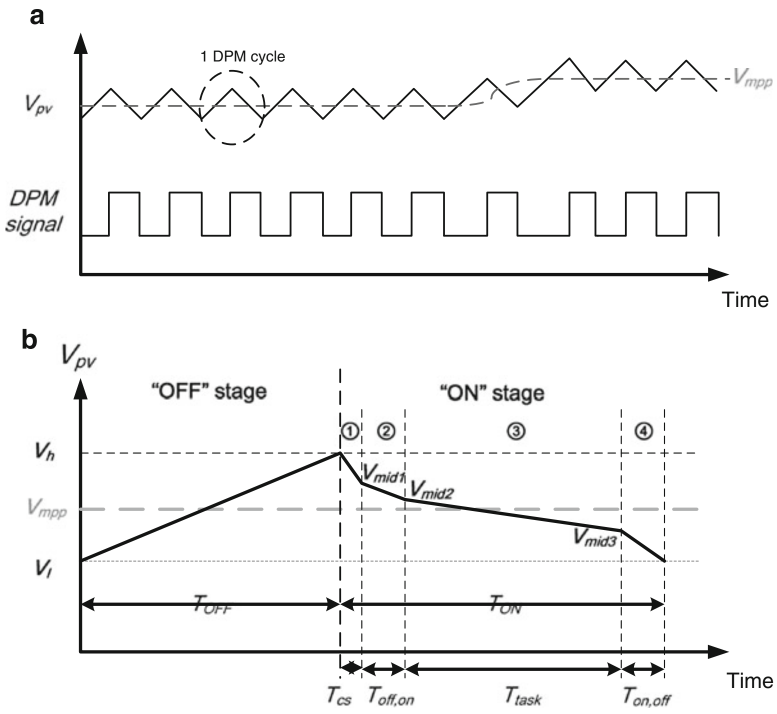


Fig. 4 (**a**) Maximum power point tracking by adjusting the DPM signal; (**b**) PV module output voltage and related system working sequence in one DPM cycle. The "ON" stage contains four processes: ① Charge sharing, ② OFF-to-ON transition, ③ Task execution, and ④ ON-to-OFF transition

turns on the load switch and wake up the electronic load. Then, the PV module output voltage drops as the load is connected to the PV module. (We only consider the condition that the load current is larger than the MPP current, otherwise, the PV module is over-designed.) The "ON" stage can be divided into four processes, which are described in the following:

- **Charge sharing**: After the load switch is turned ON, the bulk capacitor shares the charge with the decoupling capacitor at the output stage, resulting in a voltage drop of the bulk capacitor from V_h to V_{mid1}. The charge sharing process lasts for a short time T_{cs}.
- **OFF-to-ON transition**: After the charge sharing process, the power supply of the electronic load recovers, and the power system goes into OFF-to-ON transition process. It restores the status from nonvolatile memories and prepares to continue the normal task execution. The OFF-to-ON transition time is $T_{off,on}$. When the OFF-to-ON transition finishes, V_{pv} drops to V_{mid2}.
- **Task execution**: Once the OFF-to-ON transition is over, the node goes into normal operation and start executing its tasks. The MPPT controller keeps

monitoring the voltage drop of V_{pv} during the task execution time. When V_{pv} drops to V_{mid3}, which is close to V_l, the task execution process ends and the MPPT controller provides a backup signal to the electronic load. We denote the time of the task execution process as T_{task}.

- **ON-to-OFF transition**: the MPPT controller generates the backup signal and forces the load to backup its states and prepare to be turned off. The ON-to-OFF transition process is reserved for system states backup before turning OFF the load switches. The ON-to-OFF transition time is denoted as $T_{on,off}$. At the end of the ON-to-OFF transition process, V_{pv} drops to V_l.

3.2 System Model

In this section, we provide the models of major components in the proposed PV system, including the PV module and the nonvolatile microprocessor based electronic load. Then we discuss the properties of the proposed system.

3.2.1 PV Module

We use a well-known single diode model [14] (shown in Fig. 5) to simulate the DC output of a PV module. The I-V characteristic of a PV module is represented by (1) and (2).

$$I_{pv} = I_L - I_0(e^{\frac{V_{pv}+I_{pv}R_s}{a}} - 1) - \frac{V_{pv} + I_{pv}R_s}{R_{sh}}, \tag{1}$$

$$a \equiv \frac{N_s n_1 k T_c}{q}, \tag{2}$$

where N_s, n_1, k, T_c and q indicate the number of cells in series, the ideality factor of the diode, the Boltzmann constant, the module temperature, and the electron charge, respectively.

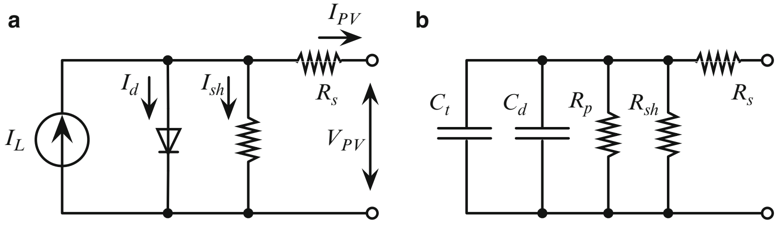


Fig. 5 (**a**) A DC equivalent circuit of a PV module; (**b**) an AC equivalent circuit of a PV module

To determine the relationship between output current I_{pv} and output voltage V_{pv} of a PV module, we need to know the five parameters in Eqs. (1) and (2), including the light current I_L, the diode reverse saturation current I_0, the series resistance R_s, the shunt resistance R_{sh}, as well as the ideality factor n_1. We extract the five parameters (I_L, I_0, n_1, R_s, R_{sh}) with a curve-fitting method. First, we measure a set of voltage-current points under a reference condition (solar irradiance of 200 W/m^2 and temperature of 27 °C). Then we find the best-fit parameters that minimize the deviation between the curve provided by the model and the measured points. For other temperature and solar irradiance condition, we can obtain corresponding parameters using the method in [14].

Table 1 lists the five parameters of a PV module extracted from measurement results. As illustrated in Fig. 6a, the model's value fits well with the actual characteristics of the PV module. To obtain the P_{mpp}, we give the simulation of the P-V curve of the PV module from the I-V characteristics, and it is shown in Fig. 6b. The PV output power P_{pv} is calculated as $P_{pv} = I_{pv} \times V_{pv}$. The MPP is the peak value with zero gradient in the P-V curve, so we have:

$$P_{mpp} = P_{pv}|_{\frac{\mathrm{d}P_{pv}}{\mathrm{d}V_{pv}}=0} \tag{3}$$

3.2.2 Nonvolatile Sensor Node

A nonvolatile processor is highly desired in a very frequent DPM because of its extremely low overheads in system states backup and recovery [15]. A recent work fabricated a nonvolatile microprocessor named THU1010N using a 130 nm

Table 1 Model parameters of the experimental PV module (Test condition: $S = 200\,\mathrm{W/m^2}$, $T_c = 300\,\mathrm{K}$)

I_L (mA)	I_0 (A)	n_1	$R_s(\Omega)$	$R_{sh}(\mathrm{k}\Omega)$	N_s
8.01	1.01^{-16}	1.885	0.17	22.36	2

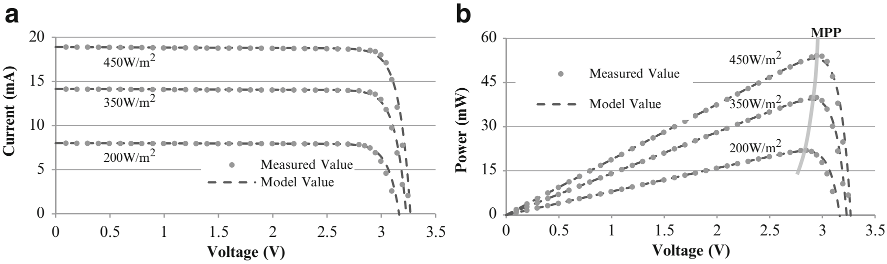


Fig. 6 (**a**) I-V curve; (**b**) P-V curve of a 4.5 × 5.5 cm^2 experimental photovoltaic module under different solar insolation at the temperature of 300 K. An MPPT window is shown on the curve for the irradiance of 200 W/m^2

Table 2 Specifications of the nonvolatile processor

Model	$T_{off,on}$ (μs)	$T_{on,off}$ (μs)	$I_{off,on}$ (mA)	$I_{on,off}$ (mA)	I_{task} (mA)	P_{task} (mW)
THU1010N	3	7	2.2	1.8	2.2	3.29
Electronic load	41	7	9.8	1.8	12.2	40.4

ferroelectric technology from Rohm, where nonvolatile elements are incorporated into each conventional volatile flip-flop [16]. Data movement between flip-flops and their local nonvolatile elements are performed in parallel, thus expediting the backup and recovery processes.

In this paper, we use THU1010N as the nonvolatile microprocessor, and the sensor node prototype described in Sect. 3.4.3 as the electronic load. Table 2 summarizes the power consumption and transition overheads of THU1010N and the sensor node. The statistics of the sensor node take into account the overhead of peripherals of the nonvolatile microprocessor.

3.3 System Efficiency

In this section, we derive the mathematical representation of the overall system efficiency by analyzing each working process. The energy efficiency of the proposed system η_{sys} in one DPM cycle is defined as:

$$\eta_{sys} = \frac{E_{task}}{E_{mpp}} = \frac{P_{task}T_{task}}{P_{mpp}(T_{on} + T_{off})}, \qquad (P_{mpp} < P_{on}), \tag{4}$$

where E_{task} is the effective energy used for task execution, and E_{mpp} and P_{mpp} are the maximum energy and maximum power that can be extracted from the PV cell, respectively. Since P_{task} and P_{mpp} can be easily extracted from component models, we mainly focus on the solving of T_{ON}, T_{OFF} and T_{task}. We use the voltage and current parameters denoted in Figs. 4 and 7, which show their relationships. In the following, T_{OFF}, T_{ON} and T_{task} are solved respectively by using different models in the "OFF" and "ON" stages.

OFF stage: When in the "OFF" stage, the load switch is disconnected and the electronic load is shut down. The PV module output current flows into the MPPT controller and bulk capacitor. With the Kirchhoff's current law (KCL), we have

$$I_{pv} = I_{ctrl} + C_{bulk}\frac{\mathrm{d}V_{pv}}{\mathrm{d}t}. \tag{5}$$

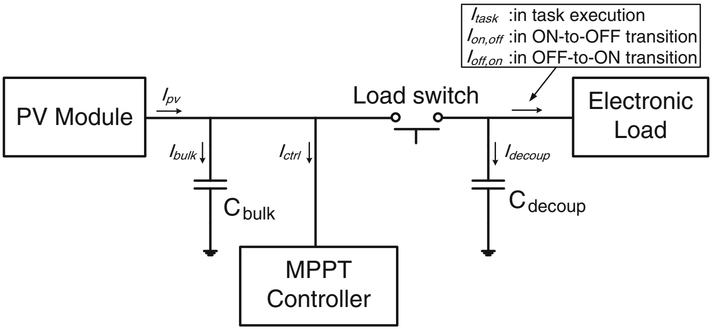


Fig. 7 Current denotation of the proposed PV system

The time to charge the bulk capacitor from V_l to V_h can be derived as follows.

$$T_{OFF} = \int_{V_l}^{V_h} \frac{C_{bulk}}{I_{pv} - I_{ctrl}} \mathrm{d}V_{pv}, \tag{6}$$

where I_{ctrl} is the current consumption of the MPP controller, and C_{bulk} is the size of the bulk capacitor.

Once the bulk capacitor is charged to V_h, the load switch would be turned ON. The "ON stage" consists of the following four processes.

Charge sharing: Immediately after the load switch is turned ON, the charges in the bulk capacitor are shared with the decoupling capacitor C_{decoup} at the output stage, resulting in an instant voltage drop of the bulk capacitor from V_h to V_{mid1}. The value of V_{mid1} is determined according to the law of charge conservation:

$$V_{mid1} = \frac{C_{bulk} V_h}{C_{bulk} + C_{decoup}}. \tag{7}$$

The charge sharing on the capacitor obeys an $e^{-\frac{t}{\tau}}$ pattern, so it takes infinite time for the V_{bulk} to be exactly equal to V_{mid1}. However, we consider that the charge sharing is finished when over 99 % charge is shared. Therefore, the charge sharing time T_{cs} is 5 times of the time constant τ:

$$T_{cs} = 5\tau = \frac{5C_{bulk}C_{decoup}R_{ls}}{C_{bulk} + C_{decoup}}, \tag{8}$$

where R_{ls} denotes the equivalent resistor of load switch. We consider that both V_{bulk} and V_{decoup} reach to V_{mid1} at the end of charge sharing, ignoring the 1 % error.

Task execution: In the task execution process, the V_{pv} drops from V_{mid2} to V_{mid3} (see in Fig. 4). Therefore, T_{task} is considered as the discharging time of bulk capacitor and decoupling capacitor from V_{mid2} to V_{mid3}. According to the current relationships shown in Fig. 7, we have

$$\begin{aligned} T_{task} &= -\int_{V_{mid2}}^{V_{mid3}} \frac{C_{bulk}+C_{decoup}}{I_{bulk}+I_{decoup}} \mathrm{d}V_{pv} \\ &= \int_{V_{mid3}}^{V_{mid2}} \frac{C_{bulk}+C_{decoup}}{I_{ctrl}+I_{task}-I_{pv}} \mathrm{d}V_{pv}, \end{aligned} \tag{9}$$

where V_{mid3} and V_{mid3} can be calculated from the "OFF-to-ON transition" and "ON-to-OFF transition" processes.

OFF-to-ON transition: After the charge sharing process, the node goes into OFF-to-ON transition. The load device restores its previous status and prepare to continue its operation. The OFF-to-ON transition time, $T_{off,on}$, can be obtained from Table 2. Therefore, V_{mid2} can it can be calculated as:

$$V_{mid2} = V_{mid1} - \int_{0}^{T_{off,on}} \frac{I_{ctrl}+I_{off,on}-I_{pv}}{C_{bulk}+C_{decoup}} \mathrm{d}t, \tag{10}$$

where V_{mid1} can be obtained from Eq. (7), $I_{off,on}$ is the current consumption of load during OFF-to-ON transition process.

ON-to-OFF transition: In the ON-to-OFF transition, the V_{pv} drops from V_{mid3} to V_l during the time $T_{on,off}$. Same as Eq. (10), we have:

$$V_{mid3} = V_l + \int_{0}^{T_{on,off}} \frac{I_{ctrl}+I_{on,off}-I_{pv}}{C_{bulk}+C_{decoup}} \mathrm{d}t. \tag{11}$$

Then the time of "ON stage" can be calculated as:

$$T_{ON} = T_{cs} + T_{on,off} + T_{off,on} + T_{task}, \tag{12}$$

where T_{task} and T_{cs} can be obtained from Eqs. (8) and (9).

Finally, by merging Eqs. (6), (12) and (9) into Eq. (4), we obtain the expression of the system efficiency in Eq. (13), in which all the parameters can be extracted from the components' models directly.

$$\begin{cases} \eta_{sys} = \dfrac{P_{task}\int_{V_{mid3}}^{V_{mid2}} \frac{C_{bulk}+C_{decoup}}{I_{ctrl}+I_{task}-I_{pv}} \mathrm{d}V_{pv}}{P_{mpp}\left(\int_{V_l}^{V_h} \frac{C_{bulk}}{I_{pv}-I_{ctrl}} \mathrm{d}V_{pv} + \frac{5C_{bulk}C_{decoup}R_{ls}}{C_{bulk}+C_{decoup}} + \int_{V_{mid3}}^{V_{mid2}} \frac{C_{bulk}+C_{decoup}}{I_{ctrl}+I_{task}-I_{pv}} \mathrm{d}V_{pv} + T_{on,off} + T_{off,on}\right)} \\ V_{mid2} = \dfrac{C_{bulk}V_h}{C_{bulk}+C_{decoup}} - \displaystyle\int_{0}^{T_{off,on}} \frac{I_{ctrl}+I_{off,on}-I_{pv}}{C_{bulk}+C_{decoup}} \mathrm{d}t \\ V_{mid3} = V_l + \displaystyle\int_{0}^{T_{on,off}} \frac{I_{ctrl}+I_{on,off}-I_{pv}}{C_{bulk}+C_{decoup}} \mathrm{d}t. \end{cases} \tag{13}$$

Noted that Eq. (13) holds under the following constraint, which guarantees the DPM based working mechanism:

$$I_{ctrl} < I_{pv} < I_{ctrl} + \min\{I_{on,off}, I_{off,on}, I_{task}\}. \tag{14}$$

3.4 Experiments

3.4.1 Experimental Setup

In this part, we set up two baseline systems to show the advantages of the proposed PV system due to: (1) nonvolatile microprocessor; (2) converter-less MPPT. The first baseline is called the *volatile microprocessor baseline*. It replaces the nonvolatile microprocessor in the proposed system with a conventional FRAM based volatile microprocessor [17]. The backup and recovery time overheads of this volatile microprocessor are 300 μs and 200 μs, respectively. The second baseline system, called the *conventional MPPT baseline*, is the MPPT system with two cascaded converters and a supercapacitor [9].

The MPP tracking for the proposed system and the volatile microprocessor baseline system are based on the constant voltage principle, where the voltage window $[V_l, V_h]$ is set to be [2.75 V,2.90 V]. The bulk capacitor and decoupling capacitor are set to 47 μF and 20 nF, respectively. In the conventional MPPT baseline, the supercapacitor is set to 0.2 F, and the two cascaded converters adopt the models in [9].

All three systems are powered by the same PV module and run the same programs on the microprocessor. The PV module size is set to 15 cm^2, with the nominal MPP voltage of 3 V. Model parameters of the PV module are summarized in Table 1.

3.4.2 Comparison of Energy Efficiency

In this part, we compare the system efficiency and energy breakdown between the proposed system and two baselines during a whole day. Firstly, we give the hourly solar radiation data in a partly cloudy day in Fig. 8, which is chosen from the NSRDB (National Solar Radiation Data Base [18]). The irradiance value shows the average solar irradiance during the last 1 h. We only use the hourly changed solar irradiance data to simplify the simulation, but the proposed method is not restricted to this kind of solar profile. Given the data in Fig. 8, we can compare the overall energy efficiency of each system in a whole day.

Table 3 shows the simulation results of the proposed system and the volatile microprocessor baseline. D_{dpm} is the duty ratio of a DPM cycle, which is defined as $D_{dpm} = \frac{T_{task}}{T_{dpm}}$. By properly choosing the size of PV module, we keep $D_{dpm} < 100\,\%$ to force the system works under DPM mechanism during the whole day. In the last

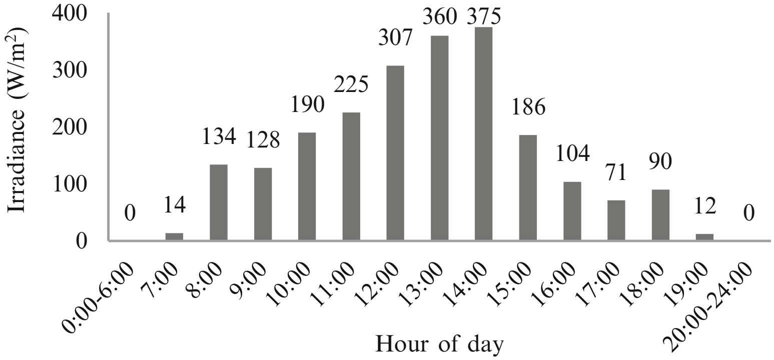


Fig. 8 Hourly solar irradiance during a day

Table 3 System efficiency and DPM duty ratio during a day

Common statistics			Proposed system				Volatile microprocessor baseline			
Time	V_{mpp} (V)	E_{mpp} (J)	Work	E_{task} (J)	D_{dpm} (%)	η_{sys} (%)	Work	E_{task} (J)	D_{dpm} (%)	η_{sys} (%)
7:00	2.52	2.9	No	0	0	0	No	0	0	0
8:00	2.73	30.4	Yes	25.3	20.3	83.3	No	0	0	0
9:00	2.72	29.0	Yes	24.1	19.3	83.0	No	0	0	0
10:00	2.76	43.6	Yes	37.8	30.3	86.5	Yes	2.8	2.2	6.4
11:00	2.78	51.9	Yes	45.7	36.7	87.9	Yes	8.1	6.4	15.5
12:00	2.81	71.6	Yes	65.2	52.4	91.0	Yes	26.6	21.3	37.1
13:00	2.82	84.4	Yes	78.1	62.7	92.5	Yes	42.9	34.4	50.9
14:00	2.82	88.0	Yes	81.7	65.7	92.8	Yes	47.9	38.4	54.4
15:00	2.76	42.7	Yes	36.9	29.6	86.4	Yes	2.5	2.0	5.8
16:00	2.71	23.4	Yes	18.7	15.0	80.0	No	0	0	0
17:00	2.67	15.7	Yes	11.7	9.4	74.4	No	0	0	0
18:00	2.69	20.1	Yes	15.8	12.7	78.4	No	0	0	0
19:00	2.50	2.5	No	N/A	0	0	No	0	0	0
Overall		506.3		440.9	27.3	87.1		130.7	8.1	25.8

row, the overall E_{mpp} and E_{task} are the sum of the values in each hour. The overall η_{sys} is defined as the overall E_{task} dividing the overall E_{mpp}. The overall D_{dpm} is the average of the hourly duty cycles.

Column 4 and 8 in Table 3 indicate whether the system works at the corresponding time. We can see that the proposed system works during 8:00–18:00, while the baseline only works during 10:00–15:00. It means that the proposed system can better tolerate the weak solar irradiance. In Column 7 and 11, the overall system efficiency of the proposed system reaches to 87.1 % with the peak system efficiency of 92.8 %. On the other hand, the volatile microprocessor baseline has a peak system efficiency of only 54.4 %, and the overall efficiency is only 25.8 %. By comparing

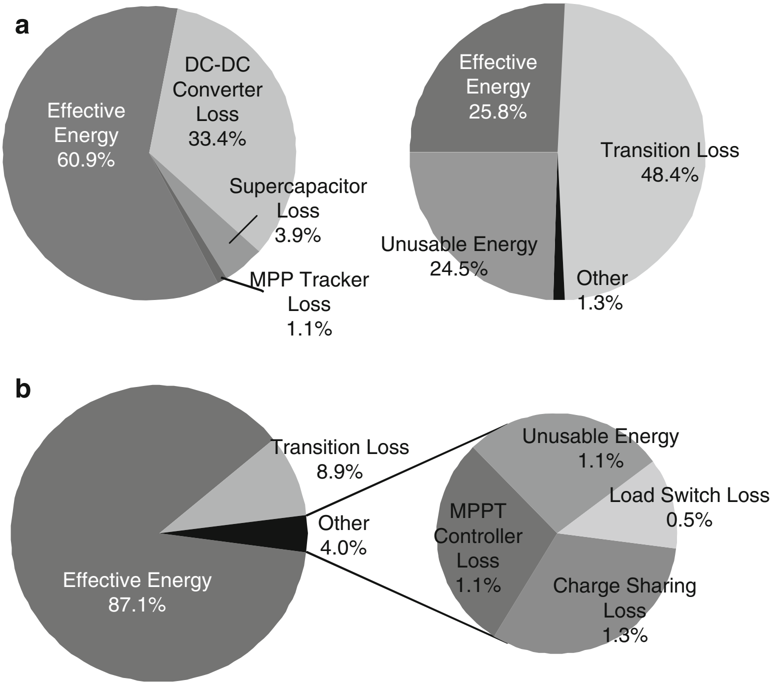


Fig. 9 Energy breakdown pie graphs of the three different PV power systems. (**a**) Conventional MPPT baseline; (**b**) volatile microprocessor baseline; (**c**) proposed system with nonvolatile microprocessor

Column 6 and 10, we find that the overall D_{dpm} of the proposed system is 27.3 %, 3.4× larger than the volatile microprocessor baseline. It means that the proposed system executes 3.4× more tasks during a day than the volatile microprocessor baseline. In sum, the proposed PV power system has better performance on the system efficiency, DPM duty cycle and low solar power tolerance. It is because that the proposed system has smaller transition overheads, which results in more time and energy for task execution.

Figure 9 shows the breakdown of the energy consumption of each system. The effective energy is the energy for task execution, while the remaining parts are energy loss. Figure 9a shows the breakdown of the conventional MPPT baseline. Its energy loss mainly comes from the converters and the supercapacitor. They contribute 37.3 % energy to the total system. Its overall energy efficiency is 60.9 %. The volatile microprocessor baseline avoids those energy loss by removing the converters and the supercapacitor. However, as Fig. 9b shows, the large transition overheads of the volatile microprocessor contribute 48.4 % energy loss. Moreover, this system can not work under weak solar radiation (see in Table 3), which leads to 24.5 % solar energy wasted as “unusable energy”. Those factors render its system efficiency even worse than the conventional MPPT baseline. Observing

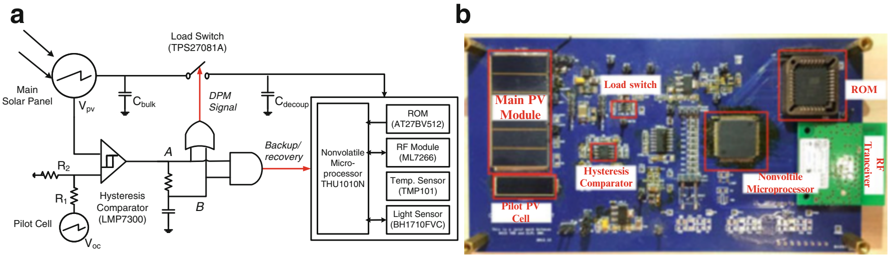


Fig. 10 (**a**) Schematic of the evaluation prototype; (**b**) photograph of the PCB board of the prototype

those challenges, we adopt the nonvolatile microprocessor to improve the transition performance. As Fig. 9c shows, the transition loss is reduced to 8.9 %. Other components, such as charge sharing, MPPT controller, load switch and unusable energy, contribute extra 4 % energy loss in all. The overall energy efficiency is 87.1 %, which is much better than above two baselines. In sum, the proposed system has the least energy loss among the three systems, thanks to its small transition overheads and the storage-less and converter-less architecture.

3.4.3 Prototype Measurements

The schematic of the prototype is show in Fig. 10a. We use the fractional open-circuit voltage (FVoc) technique [19] to implement the MPPT, which is very easy and cheap to implement, and consumes very little energy. The open-circuit voltage and MPP voltage of a PV module has approximately linear relationship shown as: $V_{mpp} = K_{foc}V_{oc}$. Therefore, we use a pilot cell (a small solar cell with the same voltage characteristic as the main solar panel) to provide open-circuit voltage V_{oc}, and set $R_2/(R_1 + R_2) = K_{foc}$. The hysteresis comparator compares the fractional V_{OC} with main solar panel output voltage V_{pv} to clamp V_{pv} around V_{mpp}, and generates the DPM control signal to the load switch. The switch signal and reset signal are output after an RC delay unit and two logic gates, in order to meet the control sequence of the nonvolatile sensor node. The nonvolatile sensor node consists of a nonvolatile processor, a ROM to store the instructions, a low power RF transceiver, and two sensors. The specifications of the system components are shown in Table 4. Figure 10b shows the photograph of the PCB board of the prototype. Some test pins are reserved to output the voltage and current of each component during the system operation. We use a Data Acquisition Card (DAQ) [20] to capture the voltage and current statistics.

We give the measurement comparison between the prototypes of proposed PV system and *conventional MPPT baseline* with 1 F supercapacitor. To give a fair comparison, we use the same PV modules with $S_{pv} = 15\,\text{cm}^2$, expose the prototypes under the same solar irradiance, and operate the same programs on their

Table 4 Specification of components used in the evaluation prototype

Component	Part no.	Active power (mW)	$I_{off,on}$ (mA)	$I_{on,off}$ (mA)	$T_{off,on}$ (μs)	$T_{on,off}$ (μs)
MCU	THU1010N	3.29	2.2	1.8	3	7
ROM	AT27LV010A	16.5	4.8	0	2	0
Load switch	TPS27081	0.03	0	0	0.25	0
RF transceiver	ML7266	16.24	1.7	0	41	0
Sensors	–	4.34	1.1	0	5	0

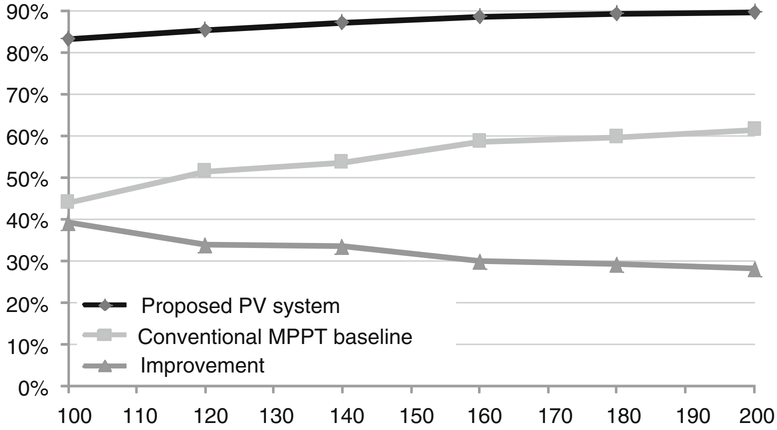


Fig. 11 System efficiency measurements on the proposed PV system and conventional MPPT baseline

microprocessors. Because the supercapacitor is an energy storage, we keep the state-of-charge (SOC) of the supercapacitor the same before and after the testing period. We first use the solar panel to charge the supercapacitor from 1.5 to 3.0 V, then disconnect the solar panel and let supercapacitor drive the electronic load until discharging to 1.5 V. To guarantee the fairness, we set the testing period of the proposed PV system the same as the baseline. Based on above setup, we obtain the system efficiency comparison under different solar irradiance ranged from 100 to 200 $\mathrm{W/m^2}$. The results are shown in Fig. 11.

From Fig. 11, we can see that the system efficiency increases with a larger solar irradiance in both of the two systems. The system efficiency of proposed PV system is 83–90 % with the average of 87.2 %, while the efficiency of the conventional MPPT baseline is 44–61 % with the average of 54.8 %. The system efficiency improvement of the proposed PV system is 28–39 % with the average of 32.4 % on average. However, it decreases with the increase of solar irradiance, which means that the conventional MPPT baseline gains more from larger solar power.

4 High-Efficiency Dual-Channel Photovoltaic Power System

Previous section compares the conventional PV power system in Fig. 12a and a storage-less and converter-less PV power system in Fig. 12b. It demonstrates that the storage-less and converter-less system [10] achieves near 90 % energy efficiency by eliminating energy loss from power converters and storage devices. However, the proposed architecture works in an energy-driven mode with best efforts. It cannot satisfy Quality of Service (QoS) when there are timing mismatches between harvested energy and workloads.

To combine the advantages in both architectures, we propose a novel high-efficiency dual-channel power supply architecture, which contains a storage-less and converter-less direct channel, an optimized "store and supply" indirect channel, as well as channel control circuits. The controller dynamically adjusts channels in different power modes to maximize the energy efficiency under failure rate constraints. Furthermore, we build both a simulation platform and a prototype to validate the architecture.

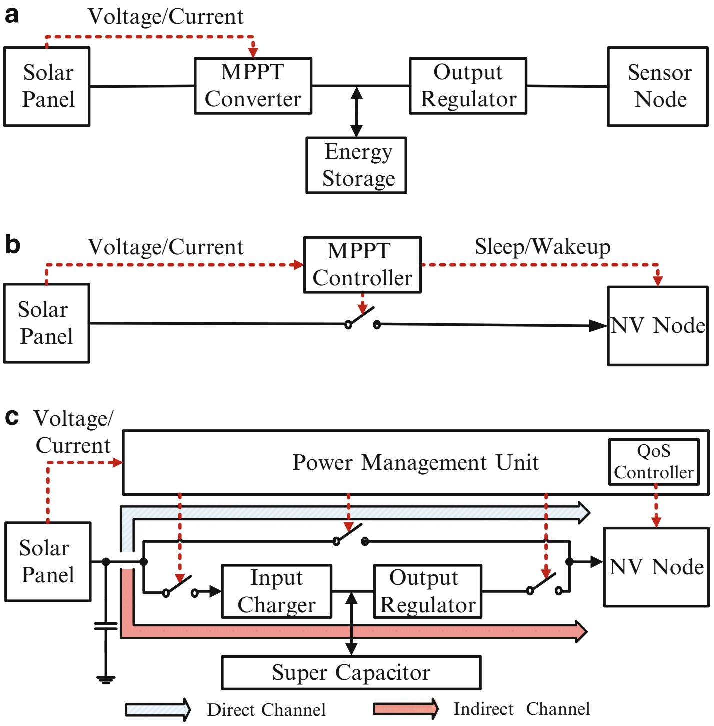


Fig. 12 Architecture comparison between conventional, converter-less and dual-channel. (**a**) Conventional architecture; (**b**) converter-less architecture; (**c**) proposed dual-channel architecture

4.1 A Dual-Channel Power System

The dual-channel architecture is a combination of the conventional power channel (indirect channel) and the converter-less channel (direct channel), with several switches inserted in the middle. Figure 12c shows the proposed architecture. We use a supercapacitor as the energy storage, considering its lifetime is much longer than a battery. The direct channel connects the solar panel to the nonvolatile sensor node via a switch and serves as the main power source when harvested energy is sufficient. It utilizes a small bulk capacitor to buffer solar energy. The voltage of bulk capacitor is kept near the Maximum Power Point (MPP) by controlling the switch. Since the MPP voltage of the selected solar panel is in the range of the sensor node's operating voltage, we can remove both converters and storage units to achieve the highest efficiency. However, this approach puts tough requirements on the sleep/wakeup speed of the sensor node, because unexpected power failures in direct channel can corrupt data and lead to reliability problems. We overcome these challenges by adopting nonvolatile (NV) sensor nodes in the next section.

When solar power decreases or disappears, the indirect channel can provide supplementary energy to satisfy the failure rate requirement. Moreover, excessive harvested energy above sensor's consumption can be stored in the supercapacitor to improve energy efficiency. However, wide variations of supercapacitor voltage will cause larger conversion loss and self-discharge is another important issue in long-term operation.

To deal with all those issues, we develop a power management unit (PMU) to select supply channels and tune the QoS level of the NV sensor node dynamically. The PMU module consists of voltage/current sensing units, switching control units and a QoS controller for the sensor node. The channel control mechanism and QoS tuning algorithm try to maximize energy efficiency while satisfying the failure rate requirements.

4.1.1 Nonvolatile Sensor Node Supporting DQS

In Fig. 13, we propose a nonvolatile sensor node supporting dynamic QoS scaling (DQS). It utilizes the direct channel for high efficiency, while relieving reliability issues and switching overheads. The proposed node contains a sleep/wakeup driver, a nonvolatile processor and several functional modules. We define the QoS level by various combinations of function modules to provide different services. The control signal from QoS controller adjusts working modules according to online solar irradiance. Because nonvolatile sensor nodes switch frequently between different levels, the switching speed and energy are important. Compared with hundreds of milliseconds in conventional sensor nodes, nonvolatile processors provide microseconds on/off switching time and low energy overhead with emerging nonvolatile memory. Therefore, the proposed sensor node supports much higher reliability and fine-grained QoS tuning.

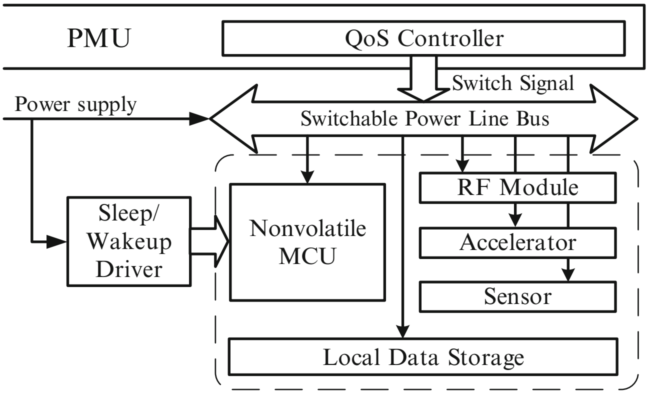


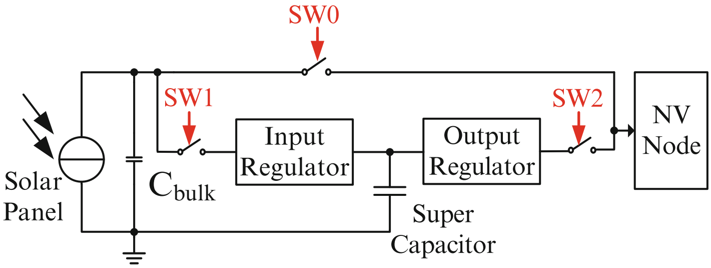
Fig. 13 Nonvolatile sensor node with different QoS levels

The working mechanism is illustrated as below: the sleep/wakeup driver detects power on/off via a sleep/wakeup driver. It generates a reset signal to trigger the sensor to backup data when power off is detected and vice versa. The QoS level can be tuned by controlling the power switches to function units. Generally, there is a basic QoS level to satisfy the fundamental operations in many applications. However, higher QoS level provides better services with larger power consumption. For example, critical structure data should always be sensed and transmitted in bridge health monitoring [21], while environmental and noncritical data can be selectively sensed, stored or transmitted. With DQS features, the proposed architecture can improve energy efficiency significantly under variable solar profiles.

4.1.2 Operation Modes and Switching FSM

In the proposed architecture, the PMU module continuously senses the condition of the scavenged solar energy and controls the work mode of the power system by switches. Meanwhile, it performs an online algorithm to control the QoS of nonvolatile node dynamically via the QoS control unit.

Figure 14 shows the connections and switch statuses of the four operating modes and the corresponding work mode transition is described in the finite state machine in Fig. 15. The input solar energy power P_s and the supercapacitor terminal voltage V_{sc} are the inputs of the FSM. $[P_{cmin}, P_{cmax}]$ is the workload power range that the QoS controller can adjust within, P_{dmin} is the minimal input energy that the node can be powered in direct channel and $[V_{scmin}, V_{scmax}]$ is the operating voltage range of the supercapacitor. As the different components in real cases have different operating ranges and switching overheads, these thresholds can be customized to fit different applications. The details of the operating modes are as follows.



Work Mode	SW0	SW1	SW2
Direct	ON	OFF	OFF
Direct+Charge	ON	DPM	OFF
Direct+Discharge	DPM	OFF	DPM
Discharge	OFF	OFF	ON

Fig. 14 Operation modes of the dual-channel supply system

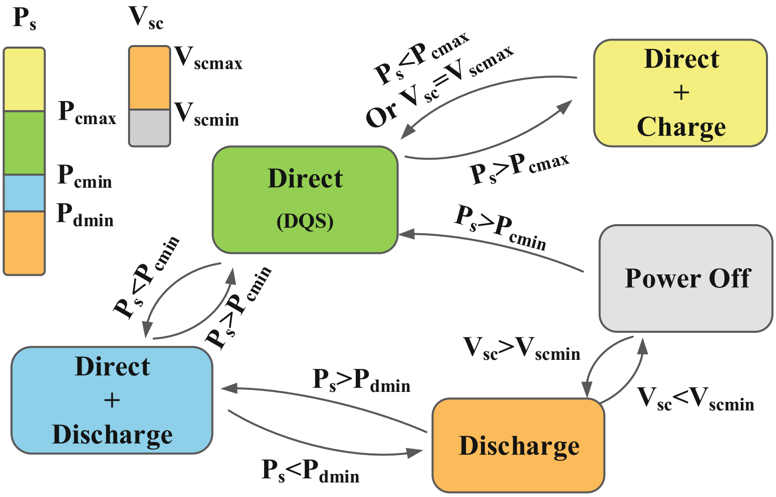


Fig. 15 FSM for operation mode transition

- **Direct Mode**: The switch $SW0$ is connected while $SW1$ and $SW2$ are disconnected. In this mode, the direct channel continuously supplies power to the node and the indirect channel is switched off. As the nonvolatile node supporting DQS can adjust the QoS level within $[P_{cmin}, P_{cmax}]$, the supply system works in this mode when input energy matches the workload consumption to maximize the energy efficiency.
- **Direct + Charge Mode**: The switch SW0 is always on, while SW1 is controlled by DPM signal. This mode is adopted when input energy is larger than P_{cmax} and the supercapacitor is not fully charged simultaneously. The charger stores the excessive solar energy into the supercapacitor. The duty cycle of DPM signal of $SW1$ is tuned to match the average charge power to the excessive energy.

- **Direct + Discharge Mode**: Both $SW0$ and $SW2$ are controlled by two complementary DPM signals, thus supplying power to the node in a time division multiplexing way. It is adopted when the solar irradiance is not strong enough that the solar panel cannot power the node independently, such as in cloudy days. The direct channel is turned off and the output regulator is turned on when the buffer capacitor's voltage drops out of the preset MPP range, so that the supercapacitor compensates the power gap between the input and workload. The next loop repeats when the buffer is recharged again.
- **Discharge Mode**: Both $SW0$ and $SW1$ are turned off and SW2 is turned on in this mode. This mode works when the input power level is lower than P_{dmin} or even absent and the supercapacitor has enough energy to power the load. When the residual energy is depleted, the node is powered off before supercapacitor is recharged or input energy is sufficient to wake it up.

4.2 Evaluation

In this section, we evaluate the performance of proposed dual-channel power system based on our simulation platform. For comparison purpose, the conventional and the converter-less architectures are also implemented as baselines. Moreover, we implemented a prototype board to validate the proposed system.

4.2.1 Experimental Setup

We use the recorded solar light intensity data of California in the year of 2011 and 2012 [22] as the input solar profiles. The time interval between two samples is 1 min. The conversion efficiency of the solar panel used in our experiment is 7 % and the load is a nonvolatile node which has zero power consumption in the sleep mode. It runs periodical solar irradiance sampling and transmitting tasks. Table 5 lists the duty cycle and power consumption in different work modes. The QoS controller tunes the total power consumption of the node by adjusting the duty cycle of the sensor and the RF transmitter. The node falls into sleep mode in the rest time.

Table 5 Application parameters

Task	Sensing	Transmitting	Sleep	Total
Power (mW)	30	100	0	3.1–16
Duty cycle (%)	10–50	0.1–1	49–89.9	–

Table 6 Efficiency comparison between conventional and dual-channel architectures

S (cm^2)	C (F)	Conventional (%)	Proposed (%)	Gain (%)	Failure rate (%)
10	150	74.34	86.01	11.67	37.75
15	250	62.83	81.95	19.12	18.52
20	500	54.22	76.89	22.67	12.93
30	800	39.14	63.15	24.01	1.71
60	1500	19.91	51.78	31.87	0

4.2.2 Comparison of Architectures

Table 6 compares the energy efficiency and failure rate improvements of the dual-channel architecture with conventional one under different pairs of solar panel size and supercapacitor size. Failure rate is the ratio of time when the power system can supply sufficient energy to the sensor node to the total time. We can see that the energy efficiency decreases when larger solar panel and supercapacitor sizes are adopted, but the proposed architecture can achieve up to 31.87 % system efficiency gain over the conventional architecture. Sizing up the sizes of solar panel and supercapacitor reduces the failure rate of the system because more energy can be buffered, but it also brings about more energy waste.

Figure 16 shows the detailed distribution of the conventional and proposed architecture using the setting from the fourth row in Table 6. As we can see, over 31 % energy loss comes from over charge in the conventional architecture and only 38 % energy is supplied to the sensor node. Meanwhile, in the proposed architecture 27 %, total energy is supplied to the node via the direct channel, thus using the over charge energy to enhance the efficiency of the node to 63 %. In real applications, designers can set the sizes of the system according to the load power consumption and failure rate constraints to get the best efficiency.

5 Intra-Task Scheduling on Storage-Less and Converter-Less Architecture

As Fig. 12a has shown, the storage-less and converter-less architecture with non-volatile processing units achieves high energy efficiency. However, it is sensitive to solar variations without energy storage. Traditional inter-task scheduling methods do not work well due to the coarse-grained scheduling. Therefore, we propose an intra-task scheduling algorithm for better quality of service. We first present the motivation and challenges to design an online intra-task scheduling algorithm; Then, we give the system model of the proposed architecture and develop an intra-task scheduling algorithm; Finally, the experimental results are given.

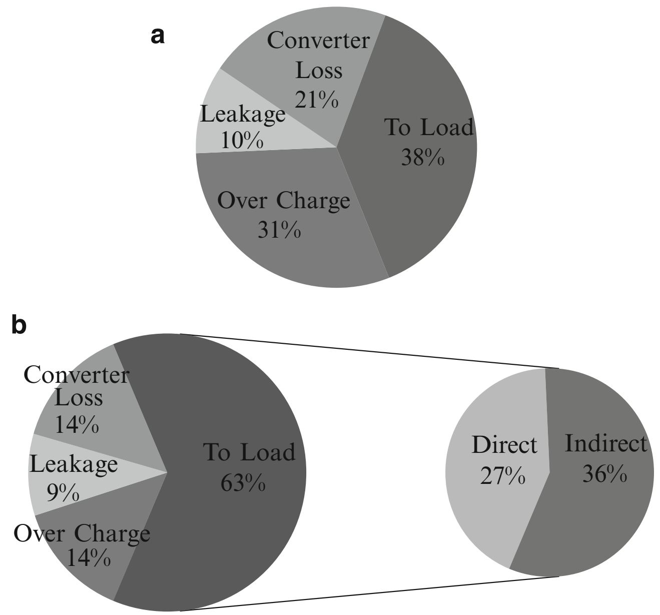


Fig. 16 Energy distribution comparison between conventional and dual-channel architectures ($S = 30\,\text{cm}^2$, $C = 800\,\text{F}$). (**a**) Conventional architecture; (**b**) dual-channel architecture

5.1 Motivation and Challenges

5.1.1 Motivation

Figure 17 shows a motivation example, where an intra-task scheduling algorithm is compared with an inter-task one. The example contains four tasks (τ_1 to τ_4) with their deadlines (D_1 to D_4). Task allocations and dependencies are also presented at the top of Fig. 17. In the figure, the inter-task algorithm schedules tasks only when the NPUs are available. Tasks cannot be interrupted during execution during solar variations. τ_1, τ_2 and the following τ_3 miss their deadlines. On the contrary, tasks can be interrupted and scheduled by the intra-task algorithm based on both solar variation and task status during execution. Thus, energy can be utilized more effectively and more deadlines can be satisfied. Eventually, only τ_1 misses its deadline and the finish time of all the tasks are earlier than those obtained by the inter-task algorithm. It shows that the intra-task scheduling algorithm is much better than the inter-task one on the architecture without energy storage.

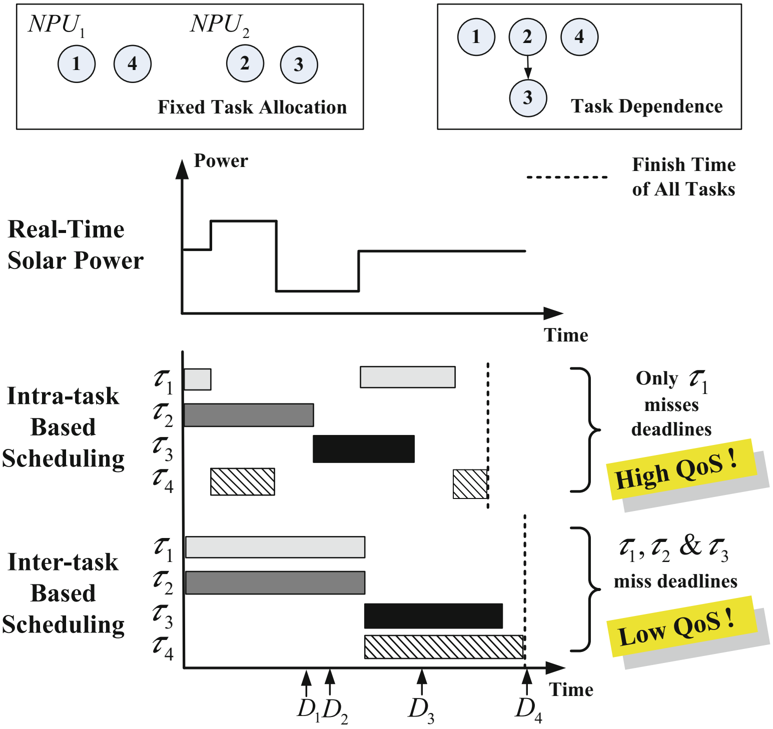


Fig. 17 A motivation example

5.1.2 Challenges

Based on the motivation example presented above, intra-task scheduling can assign any task to be executed or not in any time slot. Thus, the complexity of a brute-force algorithm is $O(2^{N \cdot M})$, where N is the number of tasks and M denotes the number of time slots in a period. Therefore, when and how to schedule tasks are two big challenges for us to deal with.

5.2 *System Modeling and Formulation*

5.2.1 System Modeling

The system model and scheduling variables for online intra-task scheduling are described as follows.

Table 7 Parameters and variables of the system model

	Parameters	Description
Task	D_i	Deadline of τ_i
	L_i	Total execution delay of τ_i
	P_i	Maximal power of τ_i (mW)
	$E_{i,j}$	Task dependence from τ_i to τ_j
System	T	Scheduling period of tasks, $t \in T$
	$P^s(t)$	Solar power at t (mW)
	A_k	Task allocation on NPU_k
Scheduling	$x_i(t)$	$x_i(t) = 1$, if τ_i is executed during the time slot t; otherwise, $x_i(t) = 0$
	$p_i(t)$	Average power consumption of τ_i during the time slot t (mW)
	$l_i^{rem}(t)$	Remaining execution cycles of τ_i on t

Table 7 summarizes the parameters and scheduling variables[2] of the system model. We first present parameters of the system. A directed acyclic graph $G(V, E)$ is used to describe task dependencies. V is the task set and E is the edge set. In V, there are N tasks (denoted as $\{\tau_1, \tau_2, \cdots, \tau_N\}$), which are executed by K nonvolatile processing unit (NPUs). Each task (τ_i) has three parameters: D_i, L_i and P_i. D_i is the deadline of τ_i. L_i and P_i are the number of execution cycles and maximal power of τ_i. In E, $E_{i,j}$ denotes the task dependence from τ_i to τ_j. $E_{i,j} = 1$, if τ_j depends on the results of τ_i; Otherwise, $E_{i,j} = 0$. The system of the solar-powered sensor node contains three parameters. T=$\{1, \cdots, t, \cdots, M\}$ is the set of time slots. $P^s(t)$ denotes the solar power at the time slot t. T is the scheduling period of tasks during which all the tasks must be completed. We assume that tasks can be preempted at any time slot t and M is the number of time slots in the period. A_k is the task allocation on NPU k and meets the NPU resource constraint: each task can only be executed by a certain NPU while an NPU executes one task at the same time.

Then, we give the scheduling variables. $x_i(t)$ is an independent scheduling variable, where $x_i(t) = 1$ if τ_i is executed at t and $x_i(t) = 0$ otherwise. Based on $x_i(t)$, We define two intermediate variables: $p_i(t)$ and $l_i^{rem}(t)$. $p_i(t)$ is the average power consumption of τ_i during the time slot t, and it is calculated as follows,

$$p_i(t) = \begin{cases} 0 & \text{if } x_i(t) = 0 \\ P_i & \text{else if } \sum_{k=1}^{N} x_k(t) \cdot P_k \le P^s(t) \\ P^s(t) \cdot \frac{P_i}{\sum_{k=1}^{N} x_k(t) \cdot P_k} & \text{otherwise.} \end{cases} \tag{15}$$

[2] Note that we use lower case letters for variables and upper case letters for parameters/constants whenever appropriate.

That is, $p_i(t) = 0$, if τ_i is not executed; $p_i(t) = P_i$, if the solar power is sufficient ($\sum_{k=1}^{N} x_k(t) \cdot P_k \le P^s(t)$); $p_i(t) = P_i/[\sum_{k=1}^{N} x_k(t) \cdot P_k]$, which is a fraction of the available solar power, if the solar power is less than the load power ($\sum_{k=1}^{N} x_k(t) \cdot P_k$). $l_i^{rem}(t)$ is the remaining execution cycles of τ_i during the time slot t, which is used to describe its current status. $l_i^{rem}(t)$ is calculated as follows.

$$l_i^{rem}(t) = L_i - [\sum_{k=1}^{t} x_i(k) \cdot p_i]/P_i \tag{16}$$

where $[\sum_{i=1}^{t} x_i(t) \cdot p_i]/P_i$ is the executed cycles of τ_i. $l_i^{rem}(t) = L_i$, if τ_i has never been executed; $l_i^{rem}(t) = 0$, if τ_i is completed.

5.2.2 Problem Formulation

Given the parameters and variables defined above, the scheduling problem can be formulated as an optimization problem. The goal is to find a schedule $\{x_i(t)\}$ for all t $(t \in T)$ that minimizes the deadline miss rate of the tasks executed on the solar-powered sensor node.

$$\min \sum_{i=1}^{N} \theta(l_i^{rem}(D_i))/N \tag{17}$$

where $\theta(\bullet)$ is calculated as follows.

$$\theta(l_i^{rem}(D_i)) = \begin{cases} 1 \text{ if } l_i^{rem}(D_i) > 0 \\ 0 \text{ otherwise} \end{cases}. \tag{18}$$

That is, τ_i misses its deadline, if $l_i^{rem}(D_i) > 0$; Otherwise, τ_i meets its deadline.

5.3 *Proposed Intra-Task Scheduling Algorithm*

5.3.1 Algorithm Framework

An algorithm framework for intra-task scheduling is presented in Fig. 18. It contains three parts. The trigger mechanism is used to choose scheduling points based on variations of both the solar power and task status. Then, task priority calculation

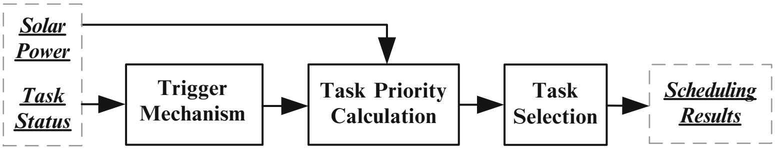


Fig. 18 The algorithm framework

has been done with an ANN trained by an optimal intra-task priority solution. Finally, task selection is performed based on task priorities and scheduling results are generated. We illustrate each part in the following subsections.

5.3.2 Trigger Mechanism

For trigger mechanism design, we find that intra-task scheduling should be done in the following situations: tasks completed, deadline missing and solar variations. Therefore, we use a trigger mechanism to launch the scheduling process, which stores all the situations as triggers.

5.3.3 Task Priority Calculation

To define task priorities, all related metrics of tasks and the solar power should be concerned: task deadlines, task energy and task dependencies. Besides, the solar power also impacts task priorities based on their status, as the supplied power source for task scheduling is changeable. Based on the consideration, the task priority of τ_i (m_i, smaller m_i means the higher priority) is defined as follows,

$$m_i = \sum\nolimits_{k=1}^{3} w_k \cdot M_i^k \tag{19}$$

$\{M_i^k\}$ are metrics of τ_i, which are calculated as follows.

$$M_i^1 = \frac{D_i}{\max\{D_k | k \in [1, N]\}}, M_i^2 = \frac{P_i \cdot L_i}{\max\{P_k \cdot L_k | k \in [1, N]\}}, M_i^3 = (1 + \sum\nolimits_{k=1}^{N} C_{i,k})^{-1} \tag{20}$$

where $C_{i,j}$ is the data dependence from τ_i to τ_j in $G(V, E)$. $C_{i,j} = 1$, if there are data transmission from τ_i to τ_j; Otherwise, $C_{i,j} = 0$. For metrics of τ_i, M_i^1 denotes the impact of the deadline; M_i^2 denotes the impact of the energy consumption; M_i^3 denotes the impact of the task dependency. $\{w_k\}$ are the weights of metrics ($\{M_i^k\}$), which denotes the impact of the solar power based on the current execution status of tasks. Thus, $\{w_k\}$ change during solar variations.

To get optimal $\{w_k\}$, we propose an optimal intra-task priority solution, which is based on both the historical solar power and current task status. First, we get the optimal scheduling results with an integer nonlinear programming (INLP) formulation is shown as follows,

$$objective: \quad \min \sum\nolimits_{i=1}^{N} \theta(l_i^{rem}(D_i))/N, \tag{21}$$

subject to :

$$\text{(a) Task dependency constraint}: \quad f_i < s_j, \quad \forall i, j, \text{if } E_{i,j} = 1, \tag{22}$$

$$\text{(b) Solar energy constraint}: \quad \sum\nolimits_{i=1}^{N} x_i(t) \cdot p_i(t) \leq P^s(t), \quad \forall t \in [t_0, M], \tag{23}$$

$$\text{(c) Task energy constraint}: \quad \sum\nolimits_{t=t_0}^{M} p_i(t) = P_i \cdot l_i^{rem}(t_0), \forall i \in \{i|\tau_i \text{is done}\}, \tag{24}$$

$$\text{(d) NPU resource constraint}: \quad \sum\nolimits_{\forall i \in A_k} x_i(t) \leq 1, A_k = \{i|\tau_i \text{runs on NPU}_\text{k}\}, \\ \forall t \in [t_0, M], \forall k \in [1, K]. \tag{25}$$

where s_i and f_i are τ_i's start and completion time. The objective is to minimize the deadline miss rate based on the optimal scheduling variables ($\{x_i(t)\}$). Besides, there are four constrains in the formulation. Task dependency constraint (22) means that τ_j starts only after all its depending tasks are completed; Solar energy constraint (23) means that the load power consumption is no more than the solar power supply; Task energy constraint (24) means that the total energy consumption of τ_i is a constant; NPU resource constraint (25) means an NPU can only run one task at the same time. The formulation can be solved by a nonlinear programming solver (like LINGO).

Based on the optimized scheduling results ($\{x_i(t)\}$), a reference task priority set $\{m_i^{ref}\}$ can be easily get. And optimal $\{w_k\}$ are achieved based on $\{m_i^{ref}(t)\}$ by solving an integer linear programming (ILP) formulation as follows. In this way, optimal $\{w_k\}$ are gained on the scheduling point t_0.

$$\min \sum\nolimits_{i=1}^{N} |m_i^{ref} - \sum\nolimits_{k=1}^{3} w_k \cdot M_i^k| \tag{26}$$

However, the formulations above are NP-hard problems, which have exponential complexities. It can not be directly used for online task priority calculation. To simplify the formulations, we choose a back propagation (BP) ANN model with single hidden layer for offline training and use it to calculate the optimal $\{w_k\}$ online.

We present the architecture of BP ANN in Fig. 19. It contains three layers: input layer, hidden layer and output layer. The input layer contains $2 + N$ elements [information of solar power ($P^s(t)$) and task status $\{l_i^{rem}(t_0)\}$], where N is the number of task status. The hidden layer contains H neurons, where H is defined by users. For each neuron, $\{W_{h,i}\}$ and $\{B_h\}$ are the weights and thresholds of the hth neurons in the layer. The output layer contains three neurons, which is the number of $\{w_k\}$. For each neuron, $\{T_{j,h}\}$ and $\{Q_k\}$ are the weights and thresholds of the kth neurons in the layer. Then, we train the ANN offline to get the parameters (weights and thresholds of all neurons) with samples (the inputs and outputs above). In this way, task priorities ($\{m_i\}$) for online intra-task scheduling can be achieved effectively by calculating the trained ANN.

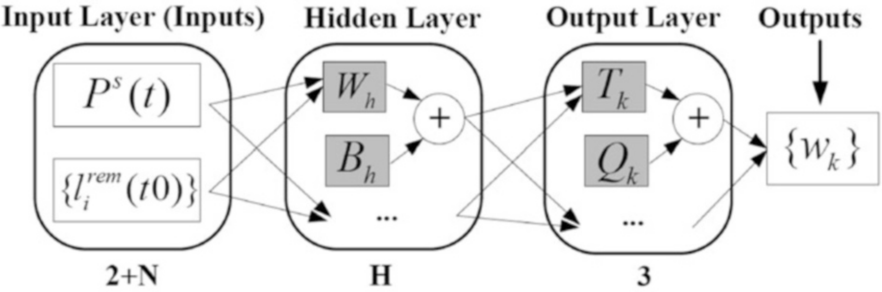


Fig. 19 The architecture of the BP ANN

5.3.4 Task Selection

Based on the calculated task priorities, the algorithm of task selection at t_0 is presented as follows. The inputs are $\{P_i\}$, $P^s(t_0)$ (in Table 7) and calculated task priorities $\{m_i\}$. The outputs are the scheduling results $(\widetilde{x}_i(t_0))$ at t_0.

Algorithm 1: Task selection at t_0

```
   input : $\{P_i\}$,$P^s(t_0)$, $\{m_i\}$
   output: online scheduling results $\{\widetilde{x}_i(t_0)\}$
 1 Initial $\{\widetilde{x}_i(t_0)\}$ as zeros, set $n = 1$;
 2 Sort $\{m_i\}$ in a descending order;
 3 while $\sum_{k=1}^{n} \widetilde{x}_k(t_0) \cdot P_k < P^s(t_0)$ & $n \leq N$ do
 4     Get$\tau_n$ with the $n$th priority in $\{m_i\}$;
 5     if $\tau_n$'s corresponding NPU is available &
 6     $over_n = 0$ & $start_n = 1$ then
 7         Set $\widetilde{x}_n(t_0) = 1$ and execute $\tau_n$ on the NPU.
 8     else
 9         The task remains in the idle state.
10     $n = n + 1$.
```

In the algorithm, initialization is done (Line 1) and $\{m_i\}$ are sorted (Line 2). Then, tasks with higher priorities in $\{m_i\}$ are chosen to be executed (Line 4&7), if the load power consumption is less than the solar power supply (Line 4). For those uncompleted tasks, which have higher priorities but cannot be executed due to data dependencies or resource unavailable (Line 5&6), they remain in the idle state (Line 9). In this way, we get the scheduling results $\{\widetilde{x}_i(t_0)\}$ at t_0.

5.4 Evaluations

5.4.1 Experimental Setup

First, we use three random cases and a real case as benchmarks in the experiment. The random cases are constructed by more than ten kinds of tasks (AES encoder/decoder, JPEG encoder/decoder, FIR filter etc. [23, 24]), where HDL files of tasks are achieved by C2RTL tools [23]. The real case is wild animal monitoring [25] (WAM), which contains eight tasks.[3] In the WAM case, the total execution delays of tasks (L_i) are gained by Modelsim simulation while the maximal power (P_i) of tasks is achieved by DC Compiler synthesis. Besides, deadlines (D_i), total periods (T), dependencies ($E_{i,j}$) and allocations (A_k) of tasks in all benchmarks are defined by users based on the applications and hardware NPUs.

Then, a solar light intensity database in 2012 [22] is used in our experiment. The corresponding harvested solar power $\{P^s(t)\}$ in the database is achieved based on our solar panel, which has an area of 24.5 cm^2 and an energy conversion efficiency of 0.06 % [10]. The profiles of the solar power from January to June are used to train the ANN in the proposed algorithm. We use the solar power from July to August for long term evaluation. Besides, the time slot (ΔT) is 1 min and the neuron number (H) in the hidden layer of the ANN is 10; The algorithms in the experiment are implemented and evaluated by running Matlab on an Intel 3 GHz notebook with 4G RAM.

5.4.2 Comparison of Algorithms

We compare the deadline miss rates and energy utilizations of algorithms with the WAM case in the long term. Figure 20a shows the daily deadline miss rates of 2 months. Compared with W-LSA [26], the proposed algorithm reduces the deadline miss rates by 19.7 % on average and it always performs better, where the minimum deadline miss rate reduction is 1.6 %. Moreover, the deadline miss rates of our algorithm are very close to those of the optimal MILP formulation, with an only 4.9 % degradation on average. Figure 20b presents the daily energy utilizations, where the proposed algorithm always achieves higher energy utilizations (19.2 % on average) than W-LSA. Our algorithm, almost as good as the optimal MILP formulation, is much better than W-LSA.

[3] The eight tasks are periodic locating, heart rate sampling, voice recording, audio processing, emergency response, audio compressing, local storing and data transmitting.

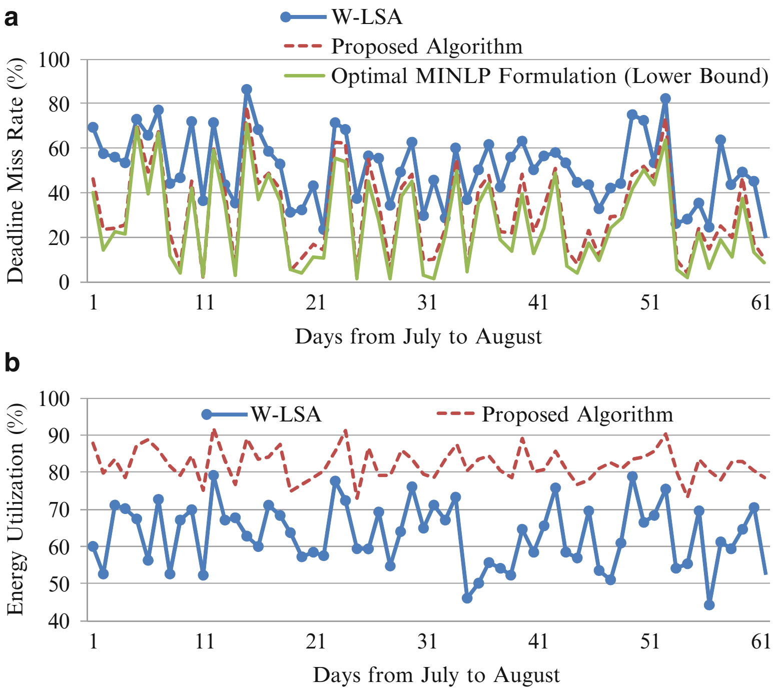


Fig. 20 Comparison of algorithms in the long term. (**a**) Comparison of deadline miss rate (%); (**b**) comparison of energy utilization (%)

5.4.3 Discussions

We analyze the overheads (execution delay and power consumption) of the proposed algorithm. We run the algorithm on a nonvolatile 8051 processor with the WAM case. The execution delay is gained by an oscilloscope and the power consumption is tested by Data Acquisition (DAQ) on the Labview. At each intra-task scheduling point, the execution delay is about 11 ms and power consumption is about 5.75 mW. Considering the numbers of scheduling points in each day (e.g. from Day 1 to Day 5) are 73–515, the energy overheads are 0.01–0.68 %. Therefore, the proposed algorithm is suitable to execute on solar-powered sensor nodes.

6 Related Work

The PV energy harvesting systems for sensor node are widely studied in recent years. Generally, they aim to optimize the energy efficiency and QoS from the PV power systems and task scheduling algorithms. In this part, we mainly summarize these related works from the vision of PV power system and task scheduling.

There are two major factors that impacts the efficiency of a PV power system: (1) the efficiency of MPPT and (2) the efficiency of the power conversion circuit. From the vision of MPPT strategy, the fractional open-circuit voltage (FOCV), which exploits the nearly linear relationship between the PV panel open-circuit voltage and the voltage at maximum power point (MPP), is most preferred for its low computation complexity and low power components [5, 8, 27]. Another option is the perturb and observe (P&O) strategy with more elaborated and precise MPPT. The more complex circuits of power measurement render it appropriate for relatively high power (>50 W) system [28]. From the vision of the conversion efficiency, people prefers energy efficient switching converters with two kinds of control methods. One is DPM control, which is easy to implement with low power circuit components [5]. However, it has poor light-load efficiencies induced by switching loss [29]. The other one is pulse frequency modulation (PFM) control, which has higher light-load efficiency because the switching frequency and related switching loss are scaled down with load current [6]. The dual-mode control methods switching between the DPM and PFM mode by the load current, is desirable to maintain a high efficiency even in a low-power PV energy harvesting system [30].

However, above works mainly suffers from the low energy efficiency from converters and supercapacitors. Pioneering solutions [31, 32] removed the DC–DC converter and replenish the energy storage by a direct connection to the PV cell. However, the direct connection method failed to perform the MPPT and results in a low system efficiency of less than 50 %. Brunelli et al. [33] provided a converter-free PV energy harvesting node with MPPT. However, lacking of efficient backup and recovery mechanism, their system had to adopt large capacitors with hundreds of microfarads to support a coarse-gained DPM, which suffered from large capacitor leakage and long response time from power failures. Multi-path PV power system was also explored in [34]. It improved the global energy efficiency by supplying load via solar panel or a little capacitor buffer. However, the switch strategies it proposed did not consider the power failure rate and QoS, which was critical for the energy harvesting sensor node.

Moreover, many works attempt to improve the system efficiency and QoS from tasking scheduling, including load matching [35], select service levels [26] or scale operating voltage/frequency [36] and cycling [37] according to available harvested energy. They can be generally classified into offline and online approaches. Offline scheduling methods make use of historical solar profiles and assume constant task timing parameters. Kansal et al. [38] proposed an offline scheduling method of duty-cycling between active and low power modes on sensor nodes. Audet et al. [39] presented static schedulers based on the energy consumption of recurring tasks. Although the offline approaches save the scheduling overhead, they are less effective when the solar profiles and task execution status vary in a wide range.

Online approaches have also been studied in a number of papers with higher efficiency on variable solar and task execution profiles. Lazy scheduling algorithm (LSA) is one of the most widely used algorithms. Piorno et al. [40] used a solar prediction method (Weather Conditioned Moving Average, WCMA) for LSA to deal with solar variations. However, none of the LSA based methods considers task execution time variations. Therefore, researchers developed scheduling algorithms with online dynamic voltage and frequency scaling (DVFS). In [41, 42], Liu et al. developed an energy-aware scheduling method and a load-matching adaptive algorithm with online DVFS. Wang et al. [43] proposed a load tuning based scheduling method with distributed online DVFS for solar-powered multicore systems. The method improved solar energy utilization ratio and QoS of the systems.

However, most of above works focused on the conventional "store and use" architecture, and the achievable efficiency is limited by inefficient supply architecture. Therefore, a more efficient power system and an adaptive QoS aware task scheduling are highly required.

7 Conclusion

This chapter proposes a novel high-efficiency PV power system for nonvolatile sensor nodes. By adopting a converter-less and battery-less architecture, energy efficiency of the PV power system is improved by up to 90 %. Furthermore, a dual-channel power system is presented by combining conventional power system with the converter-less one. It provides both high efficiency and system availability when solar power is absent. An intelligent channel controller is developed to adjust supply between two modes under failure rate constraints. Finally, an intra-task scheduling algorithm is presented for the converter-less channel to maximize the quality of service, which leverages neural network training based on solar profiles and task execution.

Acknowledgements This work was supported in part by the NSFC under grant 60976032 and 61204032, High-Tech Research and Development (863) Program under contract 2013AA013201 and National Science and Technology Major Project under contract 2010ZX03006-003-01.

References

1. Raju M, Grazier M. ULP meets energy harvesting: a game-changing combination for design engineers. Texas Instrument; 2008. http://focus.ti.com/lit/wp/slyy018/slyy018.pdf.
2. Toh WY, Tan YK, Koh WS, Siek L. Autonomous wearable sensor nodes with flexible energy harvesting. IEEE Sens J. 2014;14(7):2299–306.
3. Zhang F, Zhang Y, Silver J, et al. A batteryless 19μW MICS/ISM-band energy harvesting body area sensor node SoC. In: ISSCC 2012; 2012. p. 298–300.
4. Magno M, Boyle D, Brunelli D, et al. Extended wireless structural monitoring through intelligent hybrid energy supply. IEEE Trans Ind Electron. 2014;61:1871–81.
5. Brunelli D, Moser C, Thiele L, et al. Design of a solar-harvesting circuit for batteryless embedded systems. IEEE Trans Circuits Syst I Regul Pap. 2009;56(11):2519–28.
6. Lopez-Lapea O, Penella MT, Gasulla M. A new MPPT method for low-power solar energy harvesting. IEEE Trans Ind Electron. 2010;57(9):3129–38.

7. Simjee F, Chou PH. Everlast: long-life, supercapacitor-operated wireless sensor node. In: ISLPED 2006; 2006. p. 197–202.
8. Dondi D, Bertacchini A, Brunelli D, et al. Modeling and optimization of a solar energy harvester system for self-powered wireless sensor networks. IEEE Trans Ind Electron. 2008;55(7):2759–66.
9. Kim Y, Chang N, Wang Y, et al. Maximum power transfer tracking for a photovoltaic-supercapacitor energy system. In: ISLPED 2010; 2010. p. 307–12.
10. Wang C, Chang N, Kim Y, Park S, Liu Y, Lee H, Luo R, Yang H. Storage-less, converter-less maximum power point tracking of photovoltaic cells for a nonvolatile microprocessor. In: 2014 19th Asia, South Pacific Design Automation Conference (ASP-DAC); 2014. p. 379–84.
11. Kim S, Chou PH. Size and topology optimization for supercapacitor-based sub-watt energy harvesters. IEEE Trans Power Electron. 2013;28(4):2068–80.
12. Weddell AS, Merrett GV, Kazmierski TJ, et al. Accurate supercapacitor modeling for energy harvesting wireless sensor nodes. IEEE Trans Circuits Syst II Express Briefs. 2011;58(12):911–5.
13. Tian H, Mancilla-David F, Ellis K, et al. A cell-to-module-to-array detailed model for photovoltaic panels. Sol Energy. 2012;86(9):2695–06.
14. De Soto W, Klein SA, Beckman WA. Improvement and validation of a model for photovoltaic array performance. Sol Energy. 2006;80(1):78–88.
15. Bartling SC, Khanna S, Clinton MP, et al. An 8MHz 75μ A/MHz zero-leakage non-volatile logic-based Cortex-M0 MCU SoC exhibiting 100% digital state retention at V DD= 0V with< 400ns wakeup and sleep transitions. In: ISSCC 2013; 2013. p. 432–3.
16. Wang Y, Liu Y, Li S, et al. A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops. In: ESSCIRC 2012; 2012. p. 149–52.
17. Datasheet of MSP430FR573X mixed signal microcontrollers. Texas Instruments; 2011. p. 1–102. www.ti.com.
18. National Renewable Energy Laboratory, National Solar Radiation Data Base; 2012. http://rredc.nrel.gov/solar/old_data/nsrdb.
19. Lee DY, Noh HJ, Hyun DS, et al. An improved MPPT converter using current compensation method for small scaled PV-applications. In: APEC 2003; 2003. p. 540–5.
20. NI-DAQ family NI USB-6211. National Instrument; 2014. http://www.ni.com/data-acquisition.
21. Liu W, Fei X, Tang T, Wang P, Luo H, Deng B, Yang H. Application specific sensor node architecture optimization-experiences from field deployments. In: 2012 17th Asia and South Pacific Design Automation Conference (ASP-DAC). IEEE; 2012. p. 389–94.
22. Solar radiation intensity in CA, USA. Measurement Instrumentation Data Center (MIDC); 2011–2012. http://www.nrel.gov/midc/.
23. Li S, Liu Y, Hu S, He X, Zhang Y, Zhang P, Yang H. Optimal partition with block-level parallelization in C-to-RTL synthesis for streaming applications. In: ASPDAC 2013; 2013. p. 225–30.
24. Zhang Y. Image engineering (I) Image processing. 2nd ed. Beijing: Tsinghua University Press; 2009.
25. Liu W, Fei X, Tang T, Li R, Wang P, Luo H, Li K, Zhang L, Deng B, Yang H. Design, implementation of a hybrid sensor network for Milu Deer monitoring. In: 2012 14th International Conference on Advanced Communication Technology (ICACT); 2012. p. 52–6.
26. Moser C, Chen J-J, Thiele L. Power management in energy harvesting embedded systems with discrete service levels. In: ISLPED 2009; 2009. p. 413–8.
27. Lee DY, Noh HJ, Hyun DS, et al. An improved MPPT converter using current compensation method for small scaled PV-applications. In: APEC 2003; 2003. p. 540–5.
28. Scarpa VVR, Buso S, Spiazzi G. Low-complexity MPPT technique exploiting the PV module MPP locus characterization. IEEE Trans Ind Electron. 2009;56(5):1531–8.
29. Sahu B, Rincon-Mora GA. An accurate, low-voltage, CMOS switching power supply with adaptive on-time pulse-frequency modulation (PFM) control. IEEE Trans Circuits Syst I Regul Pap. 2007;54(2):312–21.

30. Simjee FI, Chou PH. Efficient charging of supercapacitors for extended lifetime of wireless sensor nodes. IEEE Trans Power Electron. 2008;3(3):1526–36.
31. Jiang X, Polastre J, Culler D. Perpetual environmentally powered sensor networks. In: IPSN 2005; 2005. p. 463–8.
32. Raghunathan V, Kansal A, Hsu J, et al. Design considerations for solar energy harvesting wireless embedded systems. In: IPSN 2005; 2005. p. 52–64.
33. Brunelli D, Benini L. Designing and managing sub-milliwatt energy harvesting nodes: opportunities and challenges. In: VITAE 2009; 2009. p. 11–15.
34. Christmann J-F, Beigné E, Condemine C, Willemin J, Piguet C. Energy harvesting and power management for autonomous sensor nodes. In: DAC 2012; 2012. p. 1049–54.
35. Li D, Chou PH. Maximizing efficiency of solar-powered systems by load matching. In: ISLPED 2004; 2004. p. 162–7.
36. Liu S, Wu Q, Qiu Q. An adaptive scheduling and voltage/frequency selection algorithm for real-time energy harvesting systems. In: DAC 2009; 2009. p. 782–7.
37. Vigorito CM, Ganesan D, Barto AG. Adaptive control of duty cycling in energy-harvesting wireless sensor networks. In: SECON 2007; 2007. p. 21–30.
38. Kansal A, Hsu J, Zahedi S, Srivastava MB. Power management in energy harvesting sensor networks. ACM Trans Embed Comput Syst [(TECS) - Special Section LCTES'05]. 2007;6(4):1–35.
39. Audet D, Oliveira LC, MacMillan N, Marinakis D, Wu K. Scheduling recurring tasks in energy harvesting sensors. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); 2011. p. 277–82.
40. Piorno J, Bergonzini C, Atienza D, Rosing T. HOLLOWS: a power-aware task scheduler for energy harvesting sensor nodes. J Intell Mater Syst Struct. 2010;21(12):1317–35.
41. Liu S, Qiu Q, Wu Q. Energy aware dynamic voltage, frequency selection for real-time systems with energy harvesting. In: Design, Automation, Test in Europe (DATE '08); 2008. p. 236–41.
42. Liu S, Lu J, Wu Q, Qiu Q. Load-matching adaptive task scheduling for energy efficiency in energy harvesting real-time embedded systems. In: 2010 ACM/IEEE International Symposium on Low-Power Electronics, Design (ISLPED); 2010. p. 325–30.
43. Wang Y, Chen R, Shao Z, Li T. SolarTune: real-time scheduling with load tuning for solar energy powered multicore systems. In: 2013 IEEE 19th International Conference on Embedded, Real-Time Computing Systems, Applications (RTCSA); 2013. p. 101–10.

Part V
SoCs and Equipment for Biomedical Sensing

Basic Principle and Practical Implementation of Near-Infrared Spectroscopy (NIRS)

Hyeonmin Bae

Abstract Various brain-imaging techniques, such as CT, fMRI, and EEG, have been introduced with their own strength and weakness. While CT and fMRI systems are anything but portable and thus undermine their use in dynamic conditions, EEG system has poor resolution. NIRS system, however, can be made portable with sufficiently high resolution, enabling its use in dynamic conditions and detecting valuable hemodynamics therefrom. This chapter provides a brief overview of the basic principle on NIRS, the imaging technique of diffuse optical tomography, and the superficial noise reduction method. Then, three different modulations methods for realizing the multi-channel CW NIRS are introduced. Lastly, the implementation of spread-spectrum-code-based CW NIRS is laid out for illustration.

1 Introduction

There are various functional brain imaging techniques, including functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). fMRI detects the hemodynamics of a functional brain, whereas EEG detects electrical signals resulting from neural activity. Currently, fMRI is known to be the best method of visualizing three-dimensional maps of neural activity, despite two major shortcomings: poor temporal resolution and limited portability. Portability would be an important feature for brain imaging systems because it would allow the detection of hemodynamics under dynamic conditions. Portable functional brain imaging systems would enable us to determine which brain regions are associated with certain daily activities [1, 2]. For decades, near-infrared spectroscopy (NIRS) has been widely used to monitor local changes in cerebral hemodynamics, which are caused by blood oxygenation and deoxygenation due to functional brain activities. The difference in the near-infrared absorption spectra of oxyhemoglobin (HbO2) and deoxyhemoglobin (HbR) enables the concentrations of the two molecules to be separately identified. NIRS can be implemented in a relatively small form

H. Bae (✉)
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
e-mail: hmbae@kaist.ac.kr

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_12

factor, and thus it offers greater portability than fMRI [3–5]. In addition, the temporal resolution of NIRS ($\sim$0.01 s) is two orders of magnitude better than that of conventional fMRI ($\sim$1 s). The temporal resolution of NIRS is limited only by the modulation scheme used on the optical sources and by the bandwidth of the receiver. Therefore, NIRS is appropriate for monitoring organic-chemistry phenomena, including hemodynamic activities and cell swelling [6]. However, NIRS has a critical weakness that discourages its widespread use: it has lower spatial resolution than fMRI [7].

Frans F. Jobsis, began developing a device utilizing the principle of NIRS in 1977. F. F. Jobsis was the first to measure the oxygen saturation levels and cerebral blood flow (CBF) levels in a quantitative manner. In 1988, the Hamamatsu Corporation launched its first commercialized NIRS device, NIRO-1000, with support from Delpy and his collaborator Mark Cope; this device had a single detection channel with four lasers used to cover different wavelengths. From 1990 until the present, numerous universities and companies worldwide have been engaged in the research and development of NIRS devices [8]. Recently, Hitachi introduced its EGT-7000 system, equipped with 120 channels, which enables visualization of the entire area of an adult's brain cortex by utilizing 40 pairs of optical sources and detectors. In 2002, Barbour, associated with SUNY, developed the DYNOT (dynamic near-infrared optical tomography) system, which allows 3D tomography while using continuous-wave (CW) NIRS; this system was later commercialized by NIRx Medical Technologies. A research group led by S. Arridge at University College London in Great Britain implemented a time-domain brain-imaging device named MONSTIR. MONSTIR uses an alternating series of 780-nm and 815-nm picosecond laser pulses, which are coupled into the 32-channel optical-fiber switch, and incorporates the time-correlated single-photon counting (TCSPC) technique.

2 NIRS Principle

Neuronal activity is known to be associated with changes in the local cerebral blood flow and local cerebral blood volume in the arteries. Additionally, concurrent variations in venous HbO2 and HbR concentrations occur. These changes are referred to as cerebral hemodynamics and are depicted in Fig. 1. Figure 2 shows the extinction spectra of HbO2 and HbR in the near-infrared range [9]. The light absorption by water, lipids, and cytochrome aa3 in the brain channel at wavelengths of 600–900 nm is negligible. In contrast, HbO2 and HbR are the two substances that absorb most of the light at these wavelengths. Thus, NIRS employs wavelengths of 600–900 nm, which experience very low levels of absorption in tissues, allowing deeper penetration into brain tissue.

The light emitted by a laser source pointed into a human head can be modeled by a spherical diffusive wave, and the photons that reach the detector (which is also located on the exterior of the human head and is separated from the laser by 2–4 cm) propagate along a banana-shaped trajectory through the brain channel, as

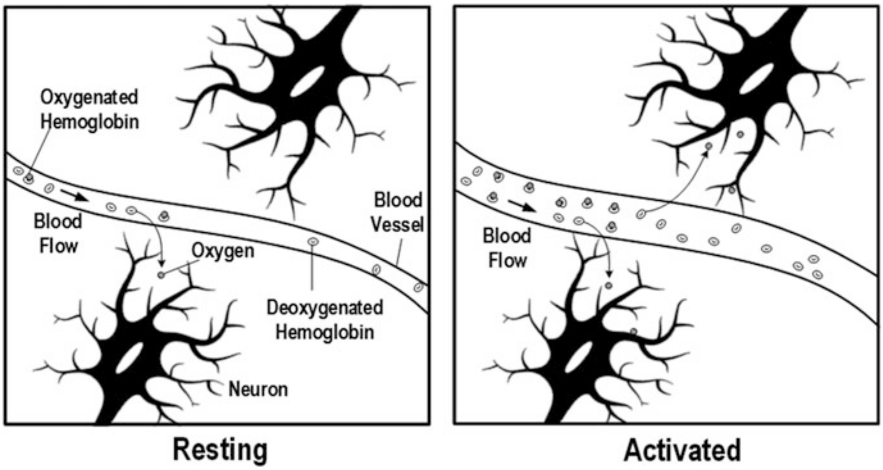


Fig. 1 Principle of NIRS: Neural activation causes oxyhemoglobin and deoxyhemoglobin concentrations to change

shown in Fig. 2. The channel is composed mainly of skin, skull, cerebrospinal fluid, and the cerebral cortex. The intensity of near-infrared light decreases as it penetrates the skull and cerebral layers owing to scattering and absorption. The variations in the absorption spectra are caused mainly by hemodynamic changes in the activated cerebral cortex.

2.1 Implementation Methods of Near-Infrared Spectroscopy (NIRS) (Time Domain, Frequency Domain, and Continuous Wave)

To characterize a subject of interest, time-domain (TD) NIRS utilizes optical pulses that are a few picoseconds in length by employing a solid-state laser. The emitted pulse experiences scattering and absorption, resulting in differences in the received photon distribution that depend on the characteristics of the tissue. The maximum intensity, the decay time, and the average area of the received pulse, all of which are properties of the temporal point spread function (TPSF), allow us to obtain diverse biometric information. In the case of TD NIRS, while its high spatial resolution is a strength, owing to its ability to clearly distinguish the absorption and the scattering effect, its implementation complexity and its cost inefficiency are its weak points (Fig. 3).

Frequency-domain (FD) NIRS measures the intensity attenuation and phase shift (Φ) of the light that passes through a tissue, by using an area detector or a photon-counting detector. Through this process, FD NIRS obtains the TPSF, as referenced

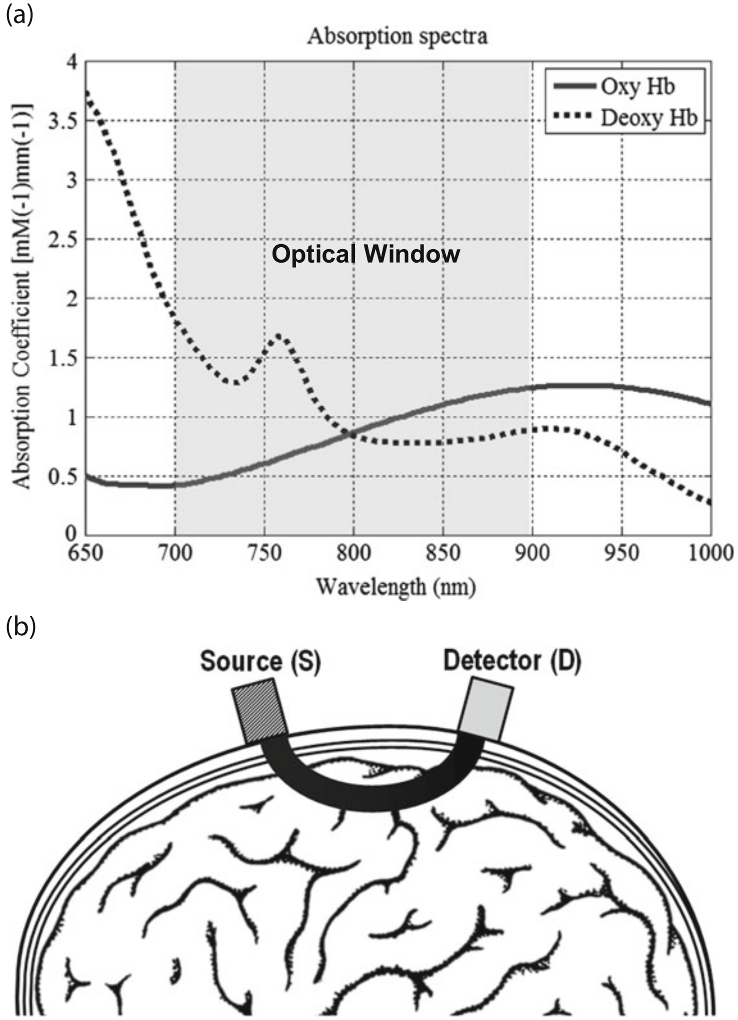


Fig. 2 (**a**) Absorption coefficients for oxyhemoglobin, deoxyhemoglobin, and water and (**b**) trajectories of detected light when propagating through a human head (rig)

in TD NIRS, as well as a variety of biometric information. Just like TD NIRS, FD NIRS is able to accurately analyze the different effects caused by the absorption and the scattering of light. However, the correlation between the path-length factor and the phase shift does not hold for modulation frequencies higher than 200 MHz, which limits the expansion of the number of channels.

CW NIRS uses either a photodiode (PD) or an avalanche photodiode (APD) to measure the attenuation profile of multi-wavelength light that passes through the tissue. Of the three methods introduced in this chapter, continuous-wave (CW) NIRS

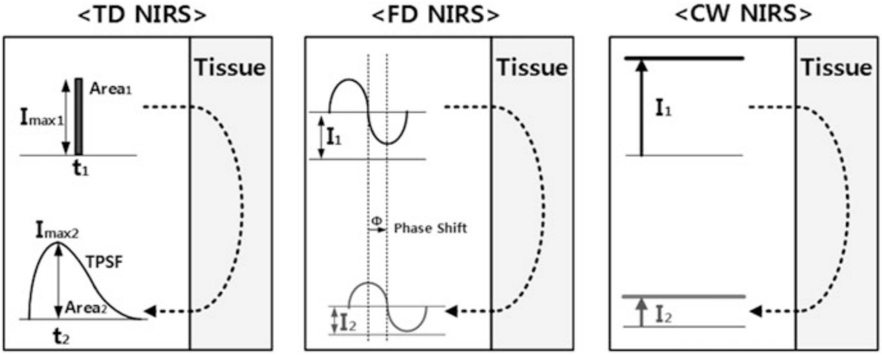


Fig. 3 Time-domain, frequency-domain, and continuous-wave-domain NIRS [10]

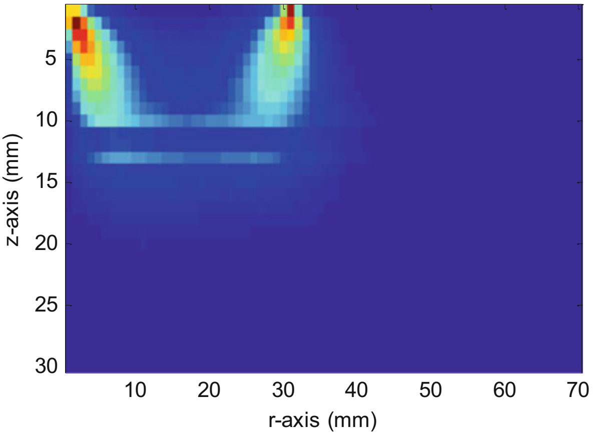
was the first to be developed. Unlike TD NIRS and FD NIRS, the CW NIRS method does not yield the absolute value of the absorption or the scattering effect. CW NIRS can only extract relative values, and thus, it can show the trend in the changing values. The major strength of CW NIRS is that the NIRS system can be simplified in structure, enabling the device to be made small and light as well as inexpensive, which is why many of the commercialized NIRS devices currently available have adopted the CW NIRS method. Section 3 will discuss in detail an implementation example of CW NIRS.

2.2 Light Propagation and the Modified Beer–Lambert Law (MBLL)

2.2.1 Photon Propagation in the Brain and the Modified Beer–Lambert Law (MBLL)

When the surface of the brain is irradiated with photons, the path of the photons within the brain can be described by Brownian motion. As a result of the Brownian motion, the propagating photons either get absorbed and disappear or else scatter to locations within a few centimeters from the point of incidence; some of these latter scattered photons re-appear on the surface. An individual photon may seem to move randomly within the brain, but when the photons are observed as a group from the viewpoint of the detector, the trajectories of the detected photons take a typical path that has a banana-shaped form, as shown by the simulation result in Fig. 4. Although this banana-shaped path (1) allows the source–detector pairs, which are attached to the surface of the brain, to measure the hemodynamics of the brain and (2) is the basis of the functional NIRS (fNIRS) technology, an apparent limitation of this approach is that white matter in the brain acts as a strong scatterer; hence, the photons can only penetrate to the depth of the cerebral cortex and no deeper.

Fig. 4 Result of Monte Carlo simulation of photon transport in the brain



Such photon movements in the brain can be numerically estimated and visualized by Monte Carlo simulation, and they can be analyzed by the radiative transfer equation (RTE) [11, 12]. However, because of the complexity of the RTE model, either a diffusion equation that is similar to the RTE or a modified Beer–Lambert Law (MBLL) that applies the differential path-length factor (DPF) approach is more often used in the real world.

MBLL is the most commonly used equation for NIRS. The emitted light from a source is absorbed and scattered in the brain channel. The variation of channel characteristics such as hemoglobin concentration causes the variation in the detected optical signal power. This can be described by Eq. (1), where ϕ is the optical density; I_o is the incident light intensity; I is the detected light intensity; ε is the extinction coefficient of the molecules; C is the concentration of the molecules; L is the distance between the light source and the detector; P is a DPF, which accounts for increases in the photon path length caused by tissue scattering; and G is a factor which accounts for the measurement geometry of the detector.

$$\phi = -\log \frac{I}{I_o} = \varepsilon CLP + G \tag{1}$$

Changes in the oxy- and deoxy-hemoglobin cause the detected intensity to change. When the concentrations of oxy- and deoxy-hemoglobin change, the extinction coefficient ε and distance L remain constant, and we can also assume that P and G remain constant. Thus, Eq. (1) can be rewritten as

$$\Delta\phi = -\log \frac{I_{\text{detected time}}}{I_{\text{initial time}}} = \varepsilon \Delta CLP, \tag{2}$$

where $\Delta\phi = \phi_{\text{detected time}} - \phi_{\text{iniial time}}$ denotes the logarithmic change in optical density. $I_{\text{detected time}}$ and $I_{\text{initial time}}$ are the measured intensities before and after the experiment, and ΔC is the change in the concentrations. The other molecules of significance in the wavelength range 600–900 nm are water, lipids, and cytochrome aa3. However, we do not need to consider such molecules, as their contributions, in general, are an order of magnitude less significant than that of hemoglobin. Equation (2) can be rewritten to include the contributions of oxy- and deoxy-hemoglobin, as follows

$$\Delta\phi\left(\lambda\right) = -\log\frac{I_{\text{detected time},\lambda}}{I_{\text{initial time},\lambda}} = \left(\varepsilon_{Hbo,\lambda}\Delta C_{Hbo} + \varepsilon_{HbR,\lambda}\Delta C_{HbR}\right)LP_{\lambda}, \tag{3}$$

where λ indicates a particular wavelength. By measuring $\Delta\phi$ at two wavelengths and using the known extinction coefficients of oxy- and deoxy-hemoglobin at those wavelengths, we can then determine the concentration changes of oxy- and deoxy-hemoglobin. Equation (4) presents the rearranged MBLL, which extracts the variations in the oxy- and deoxy-hemoglobin concentrations.

$$\begin{bmatrix}\Delta C_{HbO}\left(r,s;t\right)\\ \Delta C_{HbR}\left(r,s;t\right)\end{bmatrix} = \frac{1}{d(r)l(r)}\begin{bmatrix}\alpha_{HbO}\left(\lambda_1\right) & \alpha_{HbR}\left(\lambda_1\right)\\ \alpha_{HbO}\left(\lambda_2\right) & \alpha_{HbR}\left(\lambda_2\right)\end{bmatrix}^{-1}\begin{bmatrix}\Delta\phi\left(r,s;\lambda_1,t\right)\\ \Delta\phi\left(r,s;\lambda_2,t\right)\end{bmatrix}, \tag{4}$$

where s and r denote position of source and source-detector separation distance.

2.2.2 Diffusion Equation and DPF Calculation

Table 1 lists the terminology for the parameters necessary for analysis of photon behavior. The fluence rate (photon density) of diffused photons, $\phi(r,t)$, satisfies the diffusion equation given by

$$\frac{1}{c}\frac{\partial}{\partial t}\phi\left(r,t\right) - D\nabla^2\phi\left(r,t\right) + \mu_a\phi\left(r,t\right) = S\left(r,t\right), \tag{5}$$

where c is the speed of light in the tissue, D is the diffusion coefficient as given by

$$D = \left\{3\left[\mu_a + \left(1-g\right)\mu_s\right]\right\}^{-1} \tag{6}$$

where μ_a is the linear absorption coefficient, μ_s is the linear scattering coefficient, g is the mean cosine of the scattering angle, and $S(r,t)$ is the photon source.

The calculation of the fluence rate requires a few assumptions: (1) Scattering dominates absorption: $\mu_a \ll \left(1-g\right)\mu_s$. (2) The medium has particular geometric boundaries. The solution of Eq. (5) in an infinite medium is

$$\phi\left(r,t\right) = c(4\pi Dct)^{-3/2}\,\exp\left(-\frac{r^2}{4Dct} - \mu_a ct\right). \tag{7}$$

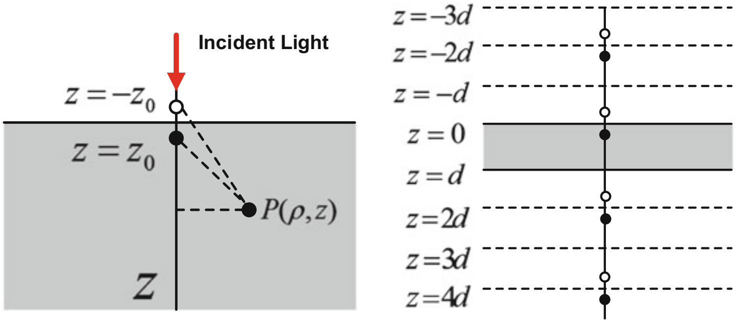


Fig. 5 Geometry for the calculation of the reflectance for a semi-infinite homogeneous medium and for a homogeneous slab

This Green-function form varies depending on how the boundary of the medium is defined. Assuming that all the incident photons are scattered at a depth Z_0, we obtain

$$z_0 = [(1-g)\,\mu_s]^{-1}\ . \tag{8}$$

In order to satisfy the condition $\phi\,(r,t) = 0$ at the physical boundary $z = 0$, a negative image source, as shown in Fig. 5a [13], is included.

Then, the fluence rate per photon incident on a semi-infinite boundary considering the image source is

$$\begin{aligned}\phi\,(\rho,r,t) &= c(r\pi Dct)^{-3/2}\ \exp\left(-\mu_a ct\right)\\ &\times\left\{\exp\left[-\frac{(z-z_0)^2+\rho^2}{4Dct}\right]-\exp\left[-\frac{(z+z_0)^2+\rho^2}{4Dct}\right]\right\}\ .\end{aligned} \tag{9}$$

The number of photons reaching the surface per unit area per unit time, $J(\rho,0,t)$, can be obtained from Fick's law, which is written as

$$J\,(\rho,0,t) = -D\nabla\phi\,(\rho,Z,t)\,|_{z=0}\ , \tag{10}$$

Table 1 Terminology used to describe beams of photons

Parameter	Meaning	Unit
Particle number	The number of particles that are emitted, transferred, or absorbed	N
Particle flux	The variation of the particle number in unit time over unit area. This quantity has a direction	$\frac{d^2N}{dtdA}$, mm^{-2} s^{-1}
Particle fluence	Flux integrated over time = radiative flux	$\frac{N}{dA}$, mm^{-2}
Particle fluence rate	Number of particles crossing through a unit-area cross-section	$\phi = \frac{d^2N}{dtdA}$, mm^{-2} s^{-1}

which leads to the final expression for the reflectance $R(\rho, t)$:

$$\begin{aligned} R(\rho, t) &= |J(\rho, 0, t)| \\ &= (4\pi Dc)^{-3/2} z_0 t^{-5/2} \ \exp(-\mu_a ct) \ \exp\left(-\tfrac{\rho^2+z_0^2}{4Dct}\right) \end{aligned} \tag{11}$$

The difference between the reflectance (flux) and the fluence rate is that the reflectance has a direction, which is orthogonal to the boundary surface (i.e., in the z-direction). The integration of $R(\rho, t)$ over the entire surface to obtain $R(t)$ results in

$$\begin{aligned} R(t) &= \int_0^\infty R(\rho, t)\, 2\pi\rho d\rho \\ &= (4\pi Dc)^{-1/2} z_0 t^{-3/2} \ \exp(-\mu_a ct) \ \exp\left(-\tfrac{z_0^2}{4Dct}\right). \end{aligned} \tag{12}$$

This reflectance can be defined as the probability density of the number of detected photons as a function of the time of flight. The normalized detected light power per unit time t is

$$\frac{P_0}{P_i} = \int_0^\infty R(\rho, t)\, dt \tag{13}$$

where P_o and P_i are the detected and input powers. We can define the absorbance of the medium as $A = \ln\left(\frac{P_o}{P_i}\right)$. Then the DPF can be defined as

$$DPF = \frac{<L>}{r} \cong \frac{v <t>}{r} = \frac{v}{r} \frac{\int_0^\infty tR(\rho, t)\, dt}{\int_0^\infty R(\rho, t)\, dt} = \frac{\partial A}{\partial \mu_a}. \tag{14}$$

The DPF should be an accurate estimate since the average photon path length <L> is much longer than the distance between the source and detector. Detailed descriptions of the procedures used to estimate the DPF are provided in [4, 7, 13–15].

2.3 Diffuse Optical Tomography

Diffuse optical tomography (DOT) is an imaging technique that seeks to improve the accuracy of existing spectroscopy techniques by not making the assumption, upon which existing spectroscopy techniques are based, that the region being measured is homogenous. DOT attempts to identify the spatial distributions of optical properties (μ_a, μ_s) or physiological parameters ($\mathrm{C_{Oxy}}, \mathrm{C_{Deoxy}}$) using the fluence rate measured at the surface of a highly scattering medium. This is an example of an “inverse problem,” which differs from a “forward problem” in which values corresponding to the spatial distribution of optical properties and/or physical parameters are already given; in inverse problems, those values are exactly the unknown parameters that DOT seeks to solve for. Because DOT acknowledges that the medium is heterogeneous, the mathematical basis for solving such a problem would be a heterogeneous diffusion equation.

In order to determine the spatial distribution of the optical properties and/or physiological parameters, the diffusion equation that was simplified under the assumption that the medium was homogeneous needs to be modified to a generalized form. This can be achieved by changing the optical properties to include a position vector **r**, resulting in the following generalized form.

$$\nabla \cdot (D(\mathbf{r}) \nabla U(\mathbf{r})) - (v\mu_a(\mathbf{r}) - i\omega) U(\mathbf{r}) = -vS_{ac}(\mathbf{r}), \tag{15}$$

where S_{ac} is modulated signal with sinusoidal. In order to solve the inverse problem utilizing the linearized method, each of the optical properties is defined as the sum of terms describing the homogenous background medium and the perturbation, as follows:

$$\mu_a(\mathbf{r}) = \mu_{a,0} + \delta\mu_a(\mathbf{r}), \quad D(\mathbf{r}) = D0 + \delta D(\mathbf{r}) \tag{16}$$

The values of $\delta\mu_a(\mathbf{r})$ and $\delta D(\mathbf{r})$ can be found from the fluence rate measured at the surface when the values of $\mu_{a,0}$, D_0 are given, by applying the Born or Rytov approximations. The Born approximation employs Eq. (17) to solve the inverse problem.

$$U(\mathbf{r}) = U_0 + U_{sc}(\mathbf{r}), \tag{17}$$

where U_{sc} indicates photon density of scattered part. By substituting Eq. (16) and Eq. (17) into Eq. (15), we get [16]

$$\begin{aligned} U_{SC}(r_d, r_s) = &-\int \frac{v\delta\mu_a(r)}{D_0} U_0(r, r_s)\, G_0(r_d, r)\, d^3r \\ &+\int \frac{\delta D(r)}{D_0} U_0(r, r_s)\, G_0(r_d, r)\, d^3r, \end{aligned} \tag{18}$$

where $G_0(r_d, r)$, r_s, and r_d indicate Green's function of specific geometry, a position vector expressing source position, and a position vector expressing detector position. Considering the fact that the measurements are made by using a finite number of source–detector pairs, Eq. (18) can be discretized as follows.

$$U_{SC}(r_d, r_s) = \sum_{i=1}^{NM} \sum_{j=1}^{NV} J_{a,ij}\, \delta\mu_a(r_j) + \sum_{i=1}^{NM} \sum_{j=1}^{NV} J_{s,ij}\, \delta D(r_j), \tag{19}$$

where $J_{a,j} = -\frac{v\Delta V}{D_0} G_0\left(r_d, r_j\right) U_0\left(r_j, r_s\right)$, $J_{s,j} = -\frac{\Delta V}{D_0} \nabla G_0\left(r_d, r_j\right) \nabla U_0\left(r_j, r_s\right)$, and NV, ΔV, and NM indicate the number of voxels, the size of unit voxel, and the number of measurements. It can also be written in matrix form as

$$\left[J_{a,ij}, J_{S,ij}\right]\left[\delta\mu_a\left(r_j\right), \delta D\left(r_j\right)\right]^T = \{U_{SC}(r_d, r_S)_i\}, \tag{20}$$

where $[J]$ is equal to $[J_{a,ij}, J_{S,ij}]$

Although it would be ideal to multiply both sides by $[J]^T$ and obtain the values of $\delta\mu_a(\mathbf{r})$ and $\delta D(\mathbf{r})$, due to the fact that $[J]^T[J]$ is either singular or very close to singular, this is not a viable option, and instead, regularization is applied to solve the equation, as shown below.

$$\left([J]^T[J] + R^g[C]^T[C]\right)[\delta\mu_a, \delta D]^T = [J]^T\{U_{SC}\}, \tag{21}$$

where R^g and $[C]$ are regularization factor and regularization operator. The Rytov approximation applies the following equation to solve the inverse problem.

$$U(\mathbf{r}) = U_0 e^{U_{SC}^R(\mathbf{r})} \tag{22}$$

By substituting Eq. (16) and Eq. (22) into Eq. (15), we get the following equation [16].

$$\begin{aligned} \mathrm{U}_{\mathrm{SC}}^{\mathrm{R}}(r_d, r_s) = &-\frac{1}{U_0(r_d, r_s)}\int \frac{v\delta\mu_a(r)}{D_0} U_0(r, r_s)\, G_0(r_d, r)\, d^3r \\ &+\frac{1}{U_0(r_d, r_s)}\int \frac{\delta D(r)}{D_0} U_0(r, r_s)\, G(r_d, r)\, d^3r \end{aligned} \tag{23}$$

The discretized form of Eq. (23) is

$$\mathrm{U}_{\mathrm{SC}}(r_d, r_s) = \sum_{i=1}^{NM}\sum_{j=1}^{NV} J_{a,ij}\delta\mu_a\left(r_j\right) + \sum_{i=1}^{NM}\sum_{j=1}^{NV} J_{s,ij}\delta D\left(r_j\right), \tag{24}$$

where $J_{a,j} = -\frac{v\Delta V}{D_0 U(r_d, r_s)} G_0\left(r_d, r_j\right) U_0\left(r_j, r_s\right)$ and $J_{s,j} = -\frac{\Delta V}{D_0 U(r_d, r_s)} \nabla G_0\left(r_d, r_j\right)$ $\nabla U_0\left(r_j, r_s\right)$. The remaining process is identical to that described for the Born approximation.

The linearized method breaks down when the value of the perturbation term is large [16]. Thus, the linearized method that guesses initial values for the optical properties $\mu_{a,0}$ and D_0 of the background medium and searches for the correct perturbation terms faces the inherent danger of a breakdown. Accordingly, a nonlinear method is used to provide supplementary support for such problems (Fig. 6).

The nonlinear method employs a process in which the spatial distribution of optical properties is repeatedly updated to continuously reduce the errors between

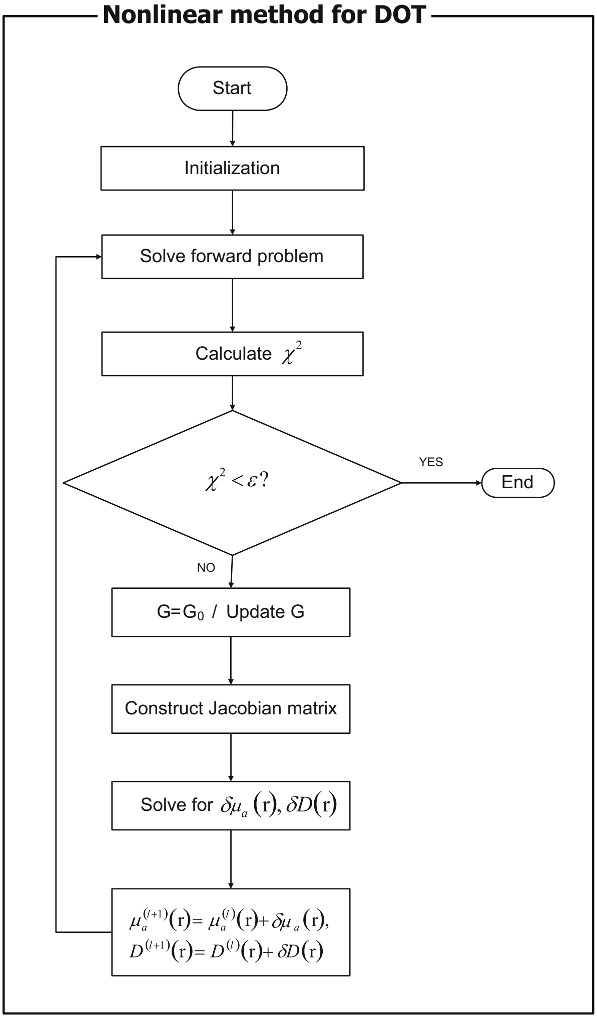

Nonlinear method for DOT
Start
Initialization
Solve forward problem
Calculate χ^2
$\chi^2 < \varepsilon$?
YES
End
NO
G=G$_0$ / Update G
Construct Jacobian matrix
Solve for $\delta\mu_a(\mathrm{r}), \delta D(\mathrm{r})$
$\mu_a^{(l+1)}(\mathrm{r}) = \mu_a^{(l)}(\mathrm{r}) + \delta\mu_a(\mathrm{r}),$
$D^{(l+1)}(\mathrm{r}) = D^{(l)}(\mathrm{r}) + \delta D(\mathrm{r})$

Fig. 6 Flow chart describing the nonlinear method

the measurement and calculated values. Thus, this process requires solving both the forward problem and the inverse problem that were previously discussed. The objective function can be defined by using the Born and Rytov approximations, respectively, as follows:

$$Born : \chi^2 = \sum_{i=1}^{NM} \left| \frac{U_m(r_d, r_s) - U_C(r_d, r_s)}{\sigma_i} \right|^2 \tag{25}$$

$$Rytov : \chi^2 = \sum_{i=1}^{NM} \left| -\frac{1}{\sigma_i} \ln\ U_C\ (r_d, r_s) \right|^2, \tag{26}$$

where σ_i is measurement error. The criteria for when to terminate the iterative calculation of χ^2 must be decided at the outset. Rather than solving the inverse matrix directly, various means such as singular value decomposition (SVD), the simultaneous iterative reconstruction technique (SIRT), and the conjugate gradient method (CGM) can be employed to estimate $\delta\mu_a(\mathbf{r})$ and $\delta D(\mathbf{r})$ in order to efficiently reduce χ^2 [11, 17]. In addition, the updated current state can be applied to the calculation, rather than the initial value of the Green's function G, to increase the convergence speed (Fig. 6).

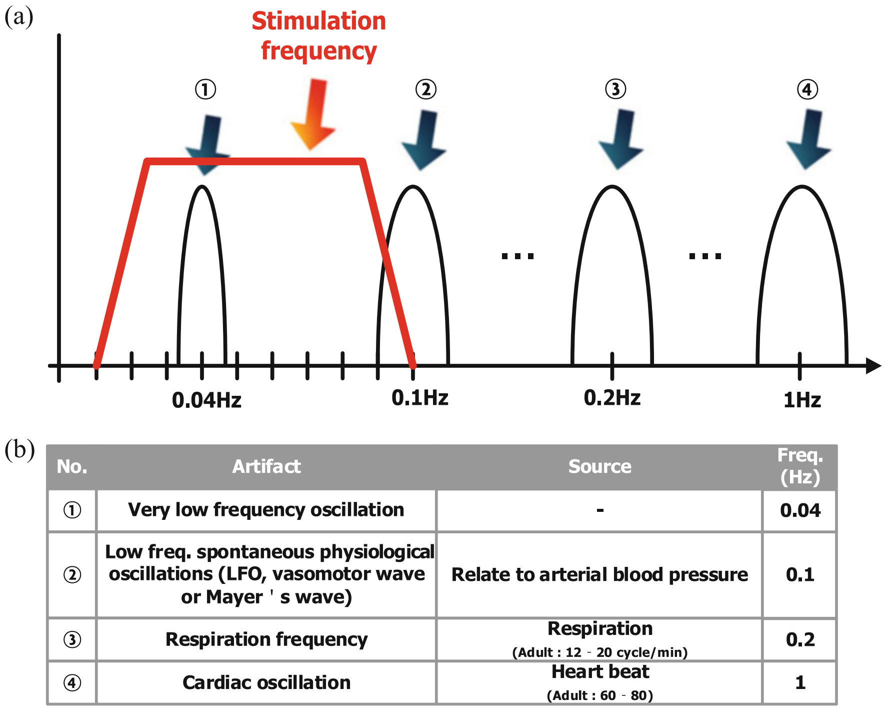
The nonlinear method is preferred over the linear method since it prevents breakdown by repeatedly updating the spatial distributions of the optical properties. However, the nonlinear method has its own weakness in that the calculation using the nonlinear method requires excessive computation time.

2.4 Superficial Contamination

Superficial layer contamination is one of the most critical issues for the development of a robust fNIRS system. Since the fNIRS system injects light into the brain from the outside and then detects (also from the outside of the head) the light that is emitted from the brain, the detected light contains the properties of the cortical layers as well as the superficial layers, such as the scalp and skull. The superficial layers contain diverse physiological artifacts, and therefore the signals measured by fNIRS are contaminated by such unwanted physiological artifacts. Such artifacts not only lower the contrast-to-noise ratio (CNR) of the fNIRS system, but they also obscure the uniqueness associated with each task. Although several solutions have been introduced to address the superficial contamination problem, no single perfect solution exists at this point.

Cardiac and respiratory activities are necessary for metabolism; these activities are the major cause of superficial layer contamination. Auto-regulation by the local tissue in response to changes in blood pressure or in response to the demands of the tissue itself causes various physiological artifacts at the superficial layer. The estimated frequency ranges given in Fig. 7 show a large deviation among individuals.

A variety of methods (summarized in Fig. 8) have been proposed to eliminate the diverse artifacts. Such methods can be classified into two categories: methods that use the reference data extracted from the superficial layer and those that do not utilize the reference data. Methods from the second category remove the cardiac oscillation by using low-pass filtering. However, the non-stationary nature of the cardiac oscillation and the band overlap between the hemodynamics and the contamination signal limit the effectiveness of such methods. In methods from the first category, a reference channel is placed in the vicinity of the light source to



No.	Artifact	Source	Freq. (Hz)
①	Very low frequency oscillation	-	0.04
②	Low freq. spontaneous physiological oscillations (LFO, vasomotor wave or Mayer ' s wave)	Relate to arterial blood pressure	0.1
③	Respiration frequency	Respiration (Adult : 12 - 20 cycle/min)	0.2
④	Cardiac oscillation	Heart beat (Adult : 60 - 80)	1

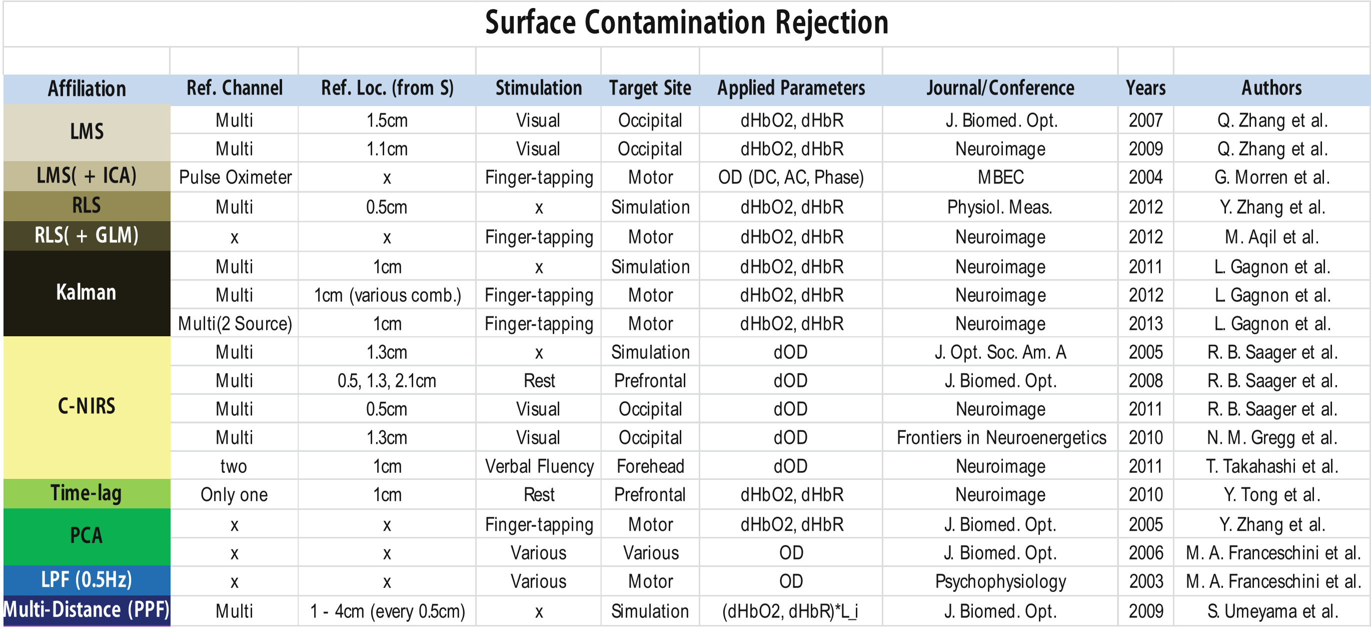
Fig. 7 Characteristics of physiological artifacts in the superficial layer. (**a**) Power spectrum. (**b**) Source and frequency

measure the signal from the superficial layer. However, recent findings from relevant research indicate that this approach has its own limitation in that it is based upon the assumption that physiological artifacts and the hemodynamics of the cortical layer are uncorrelated. Contrary to this assumption, recent findings show that artifacts and stimulation are indeed correlated [18].

2.5 Modulation Method for Multi-Channel CW NIRS

There are three main modulation methods used to distinguish multiple laser signals that are detected by the same detector: time-division multiple access (TDMA), frequency-division multiple access (FDMA), and code-division multiple access (CDMA) (Fig. 9).

The TDMA method allows each channel to operate in a designated time slot. For example, if source1 and source2 are allotted time slots T1 and T2 in the time domain, respectively, source1 emits light during T1 only while source2 remains



Surface Contamination Rejection

Affiliation	Ref. Channel	Ref. Loc. (from S)	Stimulation	Target Site	Applied Parameters	Journal/Conference	Years	Authors
LMS	Multi	1.5cm	Visual	Occipital	dHbO2, dHbR	J. Biomed. Opt.	2007	Q. Zhang et al.
	Multi	1.1cm	Visual	Occipital	dHbO2, dHbR	Neuroimage	2009	Q. Zhang et al.
LMS(+ ICA)	Pulse Oximeter	x	Finger-tapping	Motor	OD (DC, AC, Phase)	MBEC	2004	G. Morren et al.
RLS	Multi	0.5cm	x	Simulation	dHbO2, dHbR	Physiol. Meas.	2012	Y. Zhang et al.
RLS(+ GLM)	x	x	Finger-tapping	Motor	dHbO2, dHbR	Neuroimage	2012	M. Aqil et al.
Kalman	Multi	1cm	x	Simulation	dHbO2, dHbR	Neuroimage	2011	L. Gagnon et al.
	Multi	1cm (various comb.)	Finger-tapping	Motor	dHbO2, dHbR	Neuroimage	2012	L. Gagnon et al.
	Multi(2 Source)	1cm	Finger-tapping	Motor	dHbO2, dHbR	Neuroimage	2013	L. Gagnon et al.
C-NIRS	Multi	1.3cm	x	Simulation	dOD	J. Opt. Soc. Am. A	2005	R. B. Saager et al.
	Multi	0.5, 1.3, 2.1cm	Rest	Prefrontal	dOD	J. Biomed. Opt.	2008	R. B. Saager et al.
	Multi	0.5cm	Visual	Occipital	dOD	Neuroimage	2011	R. B. Saager et al.
	Multi	1.3cm	Visual	Occipital	dOD	Frontiers in Neuroenergetics	2010	N. M. Gregg et al.
	two	1cm	Verbal Fluency	Forehead	dOD	Neuroimage	2011	T. Takahashi et al.
Time-lag	Only one	1cm	Rest	Prefrontal	dHbO2, dHbR	Neuroimage	2010	Y. Tong et al.
PCA	x	x	Finger-tapping	Motor	dHbO2, dHbR	J. Biomed. Opt.	2005	Y. Zhang et al.
	x	x	Various	Various	OD	J. Biomed. Opt.	2006	M. A. Franceschini et al.
LPF (0.5Hz)	x	x	Various	Motor	OD	Psychophysiology	2003	M. A. Franceschini et al.
Multi-Distance (PPF)	Multi	1 - 4cm (every 0.5cm)	x	Simulation	(dHbO2, dHbR)*L_i	J. Biomed. Opt.	2009	S. Umeyama et al.

Fig. 8 Existing techniques for rejecting superficial layer contamination

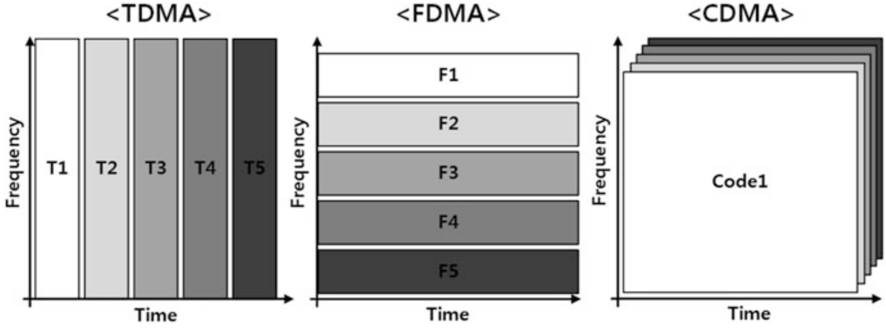


Fig. 9 Schematic representations of TDMA, FDMA, and CDMA

turned off. The opposite would apply during T2. The TDMA method is relatively simple to implement, but the SNR and time-domain resolution are suboptimal.

FDMA is a commonly used method for source separation in conventional NIRS systems. The laser sources are all modulated with different, orthogonal frequencies, and thus the hardware complexity is somewhat problematic. Performance degradation caused by nonlinearities of the laser sources is another technical challenge.

In contrast, the code-division-multiple-access (CDMA) scheme is digital-friendly, and the overall performance is not affected by the analog-domain signal integrity. A digital-domain code generator can adjust the code length, the number of orthogonal codes, and the data rate, all in the software domain. The signal-to-noise ratio in the receiver can be easily maximized by using a simple digital-domain matched filter. The immunity to frequency interference is an additional benefit of the CDMA NIRS.

The exemplary process of hemodynamic extraction through the CDMA NIRS system is illustrated below.

Figure 10 shows the operating principle of the CDMA NIRS system using a Walsh code. Each laser source is driven by an orthogonal Walsh code, and the emitted light signal is attenuated in the brain channel. Each detector receives a combined light signal containing information about the hemodynamics from different regions. The code rate should be made sufficiently high such that the attenuation in the channel remains constant during the code period. A desired laser signal can be separated from the received signal by multiplying the detected signal by the same Walsh code used to modulate the desired laser source and then adding together the code values in one period. Since the Walsh codes are orthogonal to each other, the undesired signal components vanish when the code values are summed together. Figure 10a, b respectively illustrate the cases where each of the laser sources is (a) attenuated by the same amount and (b) attenuated by different amounts. Because the propagation speed of light in the brain is negligibly small compared to the code rate, the detected light signal is almost synchronized with the original Walsh code, and thus no timing recovery circuitry is required.

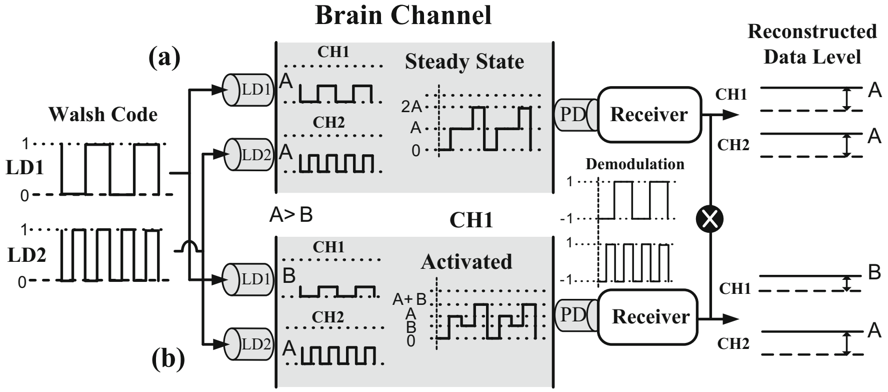


Fig. 10 Schematic of the Walsh-code-based hemodynamics-measuring method for separating each information channel from a single data stream [19]

3 NIRS Implementation Example

3.1 NIRS System Design

Figure 11 shows the implementation example of a CW NIRS system using a CDMA scheme. Vertical-cavity surface-emitting lasers (VCSEL) including two laser sources in the vicinity of 800 nm are utilized. The average power of each laser is adjusted not to violate various regulations. Each laser is modulated with the spread-spectrum code and is driven directly by using a microcontroller (MCU) for source separation on the detector side. Silicon PIN diodes are used as the detectors. The current output from the photodetector is converted to a voltage signal by using a transimpedance amplifier (TIA). The gain and bandwidth of the TIA are >100 dB and <10 kHz, respectively to detect spread-spectrum code without band-limitation while maintaining sufficient sensitivity. A passive bandpass filter with a low 3-dB cut-off frequency of ~0.1 Hz is inserted after the TIA in order to eliminate the DC offset without impairing hemodynamics. A programmable gain amplifier (PGA) adjusts the amplitude of the incoming voltage signal to match the full scale of the analog-to-digital converter (ADC). The gain of the PGA is controlled automatically in the digital domain by using a MCU to overcome inter-personal and locational power variations. Typical dynamic range of the PGA is >50 dB. The resolution and the sampling rate of the ADC are >10 bits and >1 ksamples/s, respectively. The entire analog front-end (AFE) should be designed with fully differential circuits to enhance the immunity to common noise. The detected signals are demodulated in the MCU for the source separation.

Since the size of the head, the thicknesses of the layers composing the head, and the color of the skin all vary depending on the age, race, and gender of the subject, the amount of light absorbed/scattered while traveling through the brain channel

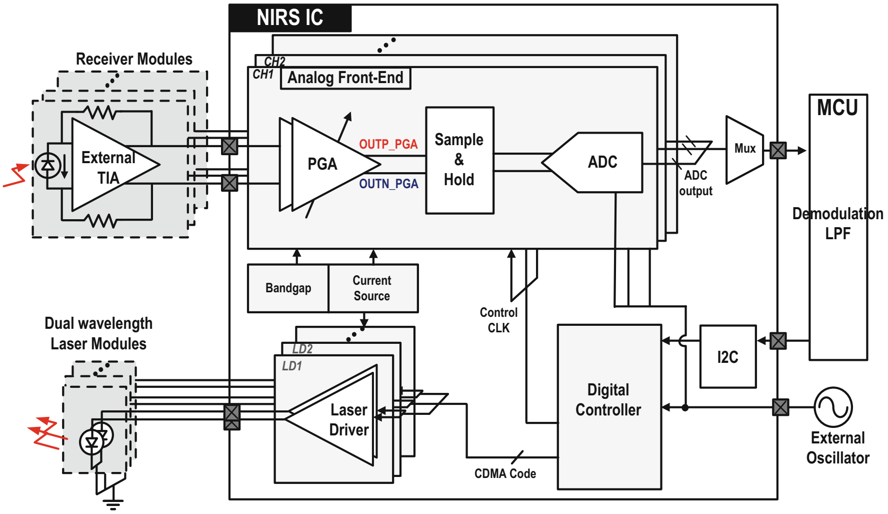


Fig. 11 Block diagram of the proposed NIRS system

varies as well. In addition, even if the fNIRS device is used on the same subject, the power of the received light signal varies significantly depending on the location of the device on the head. Note that the received signal power is strongly coupled to the SNR. A high-power laser source can improve the SNR, but a variety of clinical regulations set the upper bound on the laser power. Therefore, it is crucial to design a highly sensitive receiving circuit that can detect even the slightest signals with sufficient SNR.

3.2 *Transmitter and Receiver Modules*

Figures 12 and 13 show the assemblies of the transmitter and the detector modules, respectively.

In the case of the light transmitter module, the VCSEL laser is placed inside the cylinder with a mounting PCB on top. Light-emitting diodes (LED) can be used as light sources in place of the VCSEL for cost reduction. However, performance degradation is inevitable in this case.

A reference light receiver measures the light signal that traveled through the scalp region without passing through the cerebral cortex. The reference light receiver is placed in the vicinity of the VCSEL laser to measure the physiological artifact/biological changes. In addition, the reference light receiver can also measure the power fluctuation of the VCSEL laser associated with the temperature drift. The spring enables flexible vertical movement of the cylinder within the module to ensure secure contact of the laser with the subject regardless of the surface conditions and curvature of the skull.

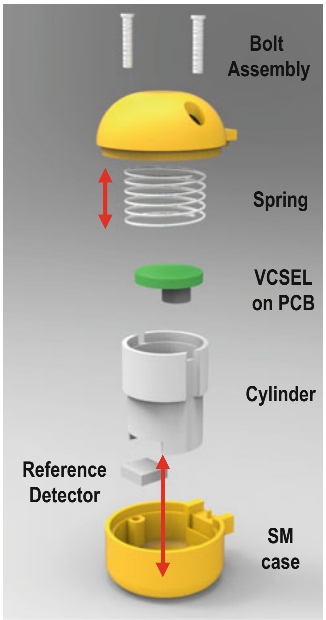


Fig. 12 Assembly of a light transmitter module

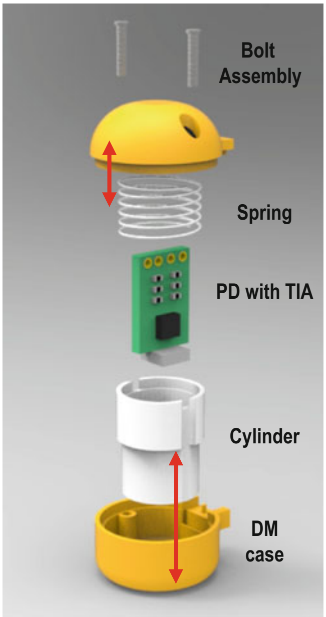


Fig. 13 Assembly of a light receiver module

The light receiver, which consists of a photodetector (PD) together with a trans-impedance amplifier (TIA), extracts the light signal that traveled through the brain channel. The light signal detected by the PD is amplified and converted into a differential voltage signal by the TIA. Just as in the transmitter module, the cylinder is attached to a spring, which ensures secure contact with the subject. Considering the fact that the gain of the TIA is large, prevention of external noise coupling is critical for signal integrity.

3.3 Specifications of the Modulation Code

The change in the concentration of oxy- and deoxy-hemoglobin in the body is a slow process, and most of the associated frequency components exist below 1 Hz. When the CDMA scheme is chosen, the following parameters must be considered.

The clock frequency and the code length must be determined by considering the hemodynamic frequency. The hemodynamics should remain constant during the symbol period for proper extraction. Therefore the symbol rate should be several orders of magnitude higher than the bandwidth of the hemodynamics.

The envelope of the laser light modulated by the spread-spectrum code varies during its passage depending on the hemodynamics of its path. Analytically, this phenomenon is identical to the hemodynamics being modulated by the spread-spectrum code. The frequency-domain representation of the modulated hemodynamics signal is shown in Fig. 14. The modulated hemodynamics signal is located far from the 1/*f* noise in the frequency domain. Then once the received signal is demodulated and low-pass-filtered, the 1/*f* noise can be easily eliminated while retrieving the subject's hemodynamics information. This process is similar to the commonly used frequency-chopping scheme.

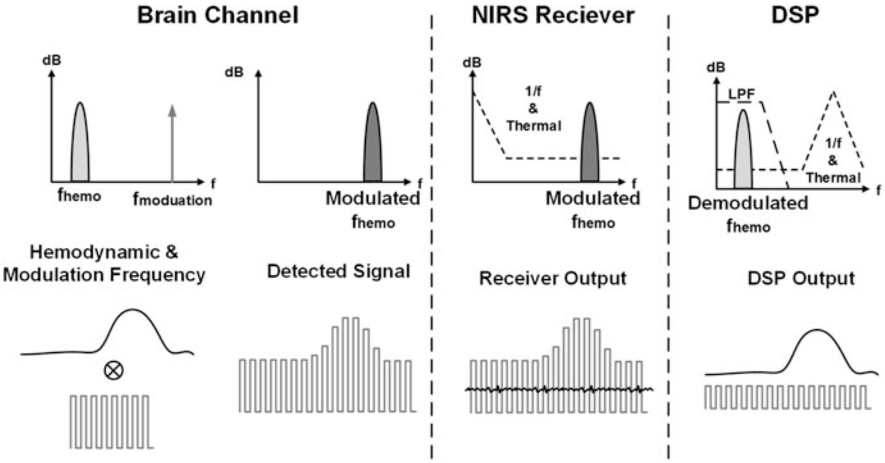


Fig. 14 Frequency spectra of the hemodynamics signal and 1/*f* noise under code-based modulation/demodulation

4 Conclusion

In the near future, an opportunity will arise whereby a functional brain imaging system will be developed, utilizing NIRS, that is portable and able to confirm the proper operation of various brain functions and physical condition of a brain. This would enable medical professionals to suggest more effective means of treatment for those patients suffering from partial or complete brain damage due to brain diseases or accidents. Consequently, such a system would become one of the most important diagnostic tools in the medical field in the future. The number of senior members in our society is currently large and is expected to continue increasing in the future, and the number of patients with brain diseases will increase sharply. A portable brain imaging system will enable those patients and individuals who may potentially be afflicted with some sort of brain disease to monitor their brains and prevent the occurrence of such brain diseases.

Acknowledgements This work is supported by Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Tomioka H, Yamagata B, Takahashi T, Yano M, Isomura AJ, Kobayashi H, Minura M. Detection of hypofrontality in drivers with Alzheimer's disease by near-infrared spectroscopy. Neurosci Lett. 2009;451(3):252–6.
2. Suda M, Takei Y, Aoyama Y, Narita K, Sato T, Fukuda M, Mikuni M. Frontopolar activation during face-to-face conversation: an in situ study using near-infrared spectroscopy. Neuropsychologia. 2010;48(2):441–7.
3. Wolf M, Ferrari M, Quaresima V. Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications. J Biomed Opt. 2007;12(6):062104-062104-14.
4. Atsumori H, Kiguchi M, Obata A, Sato H, Katura T, Funane T, Maki A. Development of wearable optical topography system for mapping the prefrontal cortex activation. Rev Sci Instrum. 2009;80(4):043704.
5. Zhang Q, Yan X, Strangman GE. Development of motion resistant instrumentation for ambulatory near-infrared spectroscopy. J Biomed Opt. 2011;16(8):087008-087008-12. doi:10.1117/1.3615248.
6. Thiagarajah JR, Papadopoulos MC, Verkman AS. Noninvasive early detection of brain edema in mice by near-infrared light scattering. J Neurosci Res. 2005;80(2):293–9.
7. Okada E, Firbank M, Schweiger M, Arridge SR, Cope M, Delpy DT. Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head. Appl Opt. 1997;36(1):21–31.
8. Cope M, Delpy DT. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. Med Biol Eng Comput. 1988;26:289–94.
9. Wray S, Cope M, Delpy DT, Wyatt JS, Reynolds ER. Characterization of the near infrared absorption spectra of cytochrome aa3 and haemoglobin for the non-invasive monitoring of cerebral oxygenation. Biochim Biophys Acta (BBA)—Bioenerg. 1988;933(1):184–92.
10. Delpy DT, Cope M. Near-infrared spectroscopy and imaging of living systems. Philos Trans: Biol Sci. 1997;352(1354):649–59.

11. Fantini S, et al. Semi-infinite-geometry boundary problem for light migration in highly scattering media: a frequency-domain study in the diffusion approximation. J Opt Soc Am B. 1994;11(10):2128.
12. Wang L, et al. MCML—Monte Carlo modeling of light transport in multi-layered tissues. Comput Methods Programs Biomed. 1995;47(2):131–46.
13. Ferrari M, Mottola L, Quaresima V. Principles, techniques, and limitations of near infrared spectroscopy. Can J Appl Phys. 2004;29(4):463–87.
14. Chance B, Leigh JS, Miyake H, Smith DS, Nioka S, Greenfeld R, Finander M, Kaufmann K, Levy W, Young M, Cohen P, Yoshioka H, Boretsky R. Comparison of time-resolved and-unresolved measurements of deoxyhemoglobin in brain. Proc Natl Acad Sci U S A. 1988;85(14):4971–5.
15. Firbank M, Okada E, Delpy DT. A theoretical study of the signal contribution of regions of the adult head to near-infrared spectroscopy studies of visual evoked responses. Neuroimage. 1998;8(1):69–78.
16. O’Leary MA. Imaging with diffuse photon density waves. PhD Dissertation. University of Pennsylvania; 1996.
17. Arridge S, Schweiger M. A gradient-based optimisation scheme for optical tomography. Opt Express. 1998;2(6):213.
18. Kirilina E, et al. The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy. Neuroimage. 2012;61(1):70–81.
19. Choi JK, Choi MG, Bae H-M. An efficient data extraction method for high-temporal-and-spatial-resolution near infrared spectroscopy (NIRS) systems. In: 2012 IEEE International Symposium on Circuits and Systems (ISCAS), vol. 560, no. 563; 2012. p. 20–23.

Wireless CMOS Bio-medical SoCs for DNA/Protein/Glucose Sensing

Shey-Shi Lu and Hsiao-Chin Chen

Abstract The design concepts of cantilever-based DNA sensors, poly-silicon nanowire-based protein/DNA sensors, a hydrogel-based glucose sensor, an ISFET-based pH sensor, and a bandgap-reference-based temperature sensor are discussed. In addition, the fabrication processes for these MEMS biosensors are presented. Sensor interface readout circuits that can deal with voltage, current, capacitive, resistive sensing signals are introduced. Wireless system-on-chips for DNA/protein/glucose sensing are designed and implemented using 0.35-μm CMOS technology. The experiment procedures are described in detail and complete measurement results are provided in this chapter. The cantilever-based DNA sensing system achieves the detectable DNA concentration lower than 1 pM. The detection limit of 10 fM can be reached by the nanowire-based DNA bio-SoC. In vitro test shows a resolution of 40 mM in glucose detection. The temperature sensor shows great linearity from −20 to 120 °C.

1 CMOS Compatible Bio-Sensors

In order to realize sensor system-on-chips (SoCs), bio-sensors need to be fabricated on chip and integrated with the active circuitry. Most importantly, the required post-IC processes should be fully compatible with CMOS technologies to advance the future commercialization of sensor SoCs. In this section, the design, the sensing mechanism, and the process for fabrication of CMOS compatible bio-sensors would be introduced.

S.-S. Lu (✉)
Department of Electrical Engineering, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
e-mail: sslu@ntu.edu.tw

H.-C. Chen
Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_13

1.1 Cantilever-Based DNA Sensor: Design and Fabrication

The cross-section view of the micro-cantilever DNA sensor is depicted in Fig. 1 [1]. Addressed as probe DNAs in Fig. 1, single-stranded DNA molecules with a predefined sequence are immobilized on the top of the micro-cantilever to capture specific DNA molecules. The specific DNA molecule with a sequence that matches the sequence of the probe DNA can be combined with the probe DNA through base pairing to form a single double-stranded molecule in the hybridization process and is called the target DNA.

The sensing mechanism of the DNA sensor is based on the piezoresistor which is embedded in the micro-cantilever as a physical transducer. As shown in Fig. 1, the N^+ polysilicon semiconductor layer in a standard CMOS process is used to realize a piezoresistor whose resistance depends on the applied mechanical stress. The sensing mechanism of the CMOS DNA sensor is illustrated in Fig. 2 [1]. When the sensor is exposed to the target DNA, the hybridization process would take place and change the surface stress. Then the cantilever bends to response to the mechanical stress and hence the resistance of the embedded piezoresistor alters. The variation in the resistance of the piezoresistor can be measured by a resistive readout circuit to achieve the detection of the specific DNA.

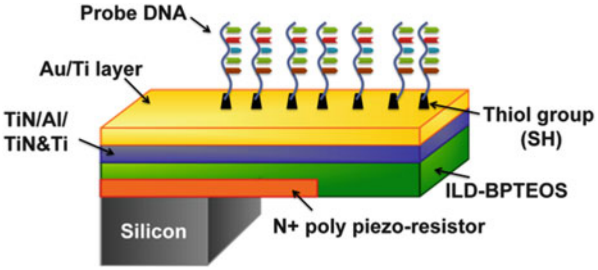


Fig. 1 Cross-section view of the micro-cantilever DNA sensor [1]

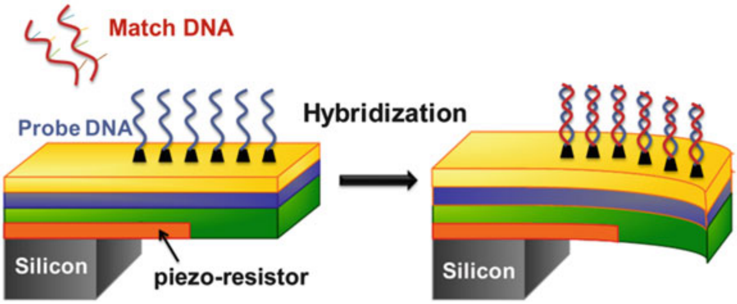


Fig. 2 Sensing mechanism of CMOS DNA cantilever sensor [1]

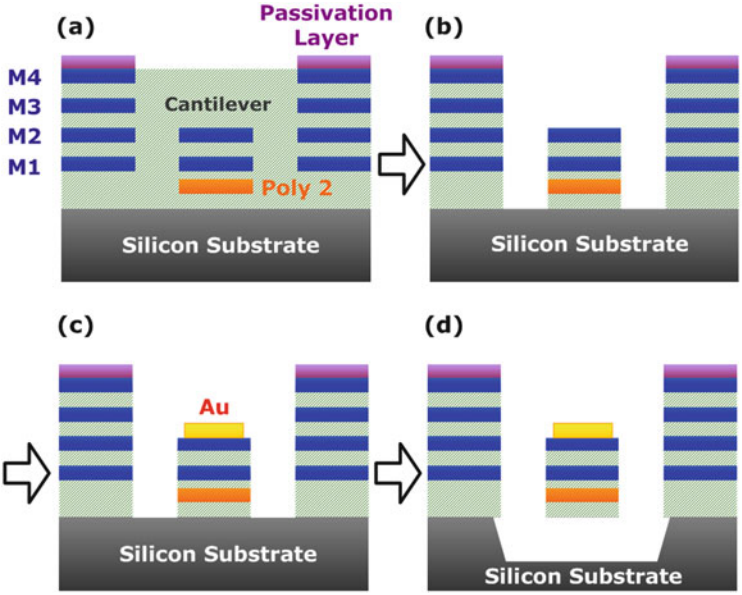


Fig. 3 Post-processing steps [1]

It has been proved that thiol-modified bio-molecules can be reliably bound to a gold surface. As known, gold has excellent biocompatibility. Moreover, gold is quite robust against surrounding changes and can withstand repetitive detection/cleaning cycles when immersed into buffer solution due to its stability. Therefore, a gold layer and thiol-modification are used to immobilize the probe DNAs onto the micro-cantilever. Figure 3 illustrates the post-processing in steps: (a) removal of the passivation layer above the sensor; (b) removing the silicon dioxide (SiO_2, the dielectric material for insulation layer in IC technology) around the defined area of the micro-cantilever by dry etching; (c) depositing a gold (Au) layer upon the micro-cantilever by lift-off method; (d) finalizing the micro-cantilever structure by using dry etching to remove the underneath silicon substrate [1]. During the post-processing, either the first metal layer (M1) or the second metal layer (M2) can be adopted as an etching stop, which determines the thickness (either 1.62 μm or 3.26 μm) of the micro-cantilever.

The DNA sensor is fabricated by using a CMOS compatible MEMS technology which is developed by National Chip Implementation Center of Taiwan. As a well-established CMOS Bio-Microelectromechanical Systems (Bio-MEMS) platform, this technology is performed by combining TSMC 0.35-μm 2P4M CMOS technology with a series of CMOS compatible micromachining post-processing. The scanning electron microscope (SEM) images of CMOS micro-cantilever-based DNA sensors are shown in Fig. 4. To find the optimal micro-cantilever structure, DNA sensors with different types of cantilever structures are realized on the same

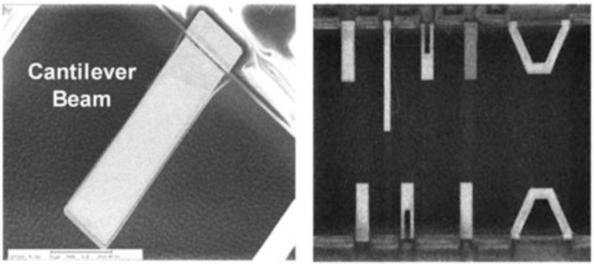


Fig. 4 SEM images of the CMOS cantilever DNA sensors [1]

chip, as shown in Fig. 4. It is worth mentioning that all the cantilevers should exhibit an identical intrinsic resistance for fair comparison of the sensitivity. During the measurement, an analog multiplexer is used to connect one of the DNA sensors with the readout circuit so that each sensor can be evaluated individually. Note that the auxiliary analog multiplexer contributes an on-resistance of 39.5 Ω which affects the measurement results of the DNA sensors. To solve the issue, the readout circuit should perform self-calibration to eliminate the effect of the on-resistance [1].

1.2 Polysilicon Nanowire Based DNA/Protein Sensor

To achieve the goals of mass production and standardization, a poly-silicon nanowire (NW) based DNA/Protein sensor is designed and implemented by using a 0.35-μm 2P4M commercially-available CMOS process, as illustrated in Fig. 5. There are two poly-silicon layers in the 0.35-μm 2P4M CMOS process, where the first poly-silicon layer (poly1) is a heavily doped poly-silicon material designed for the metal gates and the second poly-silicon layer (poly2) with lower N+ doping is designed for on-chip poly-insulator-poly (PIP) capacitors. According to previous studies, the biomolecular sensitivity of a semiconductor-based NW biosensor can be enhanced by lowering the doping concentration of the poly-silicon. Therefore, the poly2 layer of the 0.35-μm 2P4M CMOS process is chosen to implement the poly-Si NW biosensors. The cross-sectional view of a poly-Si NW-based biosensor is depicted in Fig. 6. A meander shape poly-Si NW with certain adequate impedance is designed so that the poly-Si NW sensor can have a good design window of interface circuits.

The Wheatstone bridge architecture is adopted to measure the resistance/conductance variation of poly-Si NWs for the higher sensitivity and the better common-mode rejection ratio (CMRR). Figure 7 shows the micrograph of poly-Si NW sensors in the Wheatstone-bridge. Particularly, the layout of metal layers is designed to minimize the chip area occupied by the poly-Si NW biosensor so that the nearby circuits on the same chip can be well protected during post-IC processing. Moreover, as shown in Fig. 6, all the metal layers and the interconnection layers are

Fig. 5 The illustration of a poly-Si NW-based biosensor

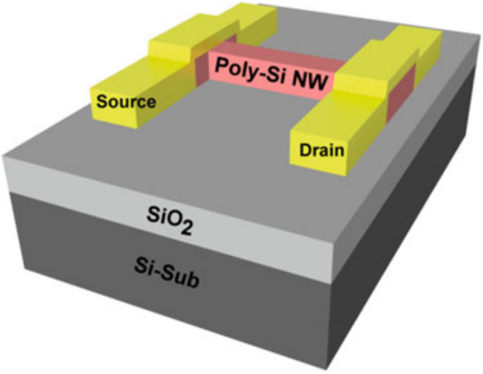


Fig. 6 The cross-sectional view of a poly-Si NW-based biosensor in the 0.35-µm 2P4M CMOS process [2]

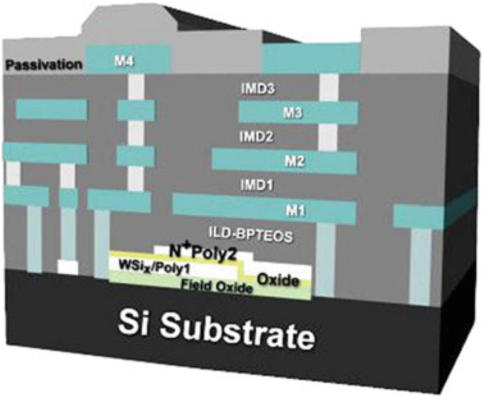
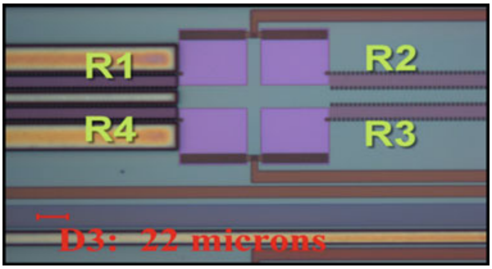


Fig. 7 The micrograph of poly-Si NW sensors in the Wheatstone-bridge [3]



stacked to form an etching stop sidewall surrounding the poly-Si NW biosensors, which prevents lateral etching and is crucial for the yield rate enhancement in the post-etching process.

To enable the poly-Si NW biosensors, it requires a post-etching process to remove most of the dielectric layers above the biosensors after the standard CMOS process so that the sensors can be exposed for DNA/protein detection. Note that the passivation layer (Si_3N_4) of the chip can be selectively removed in the standard CMOS process by a proper bond pad design because the passivation layer over the pad area would be removed to expose the top metal layer (metal4) for wire-bonding.

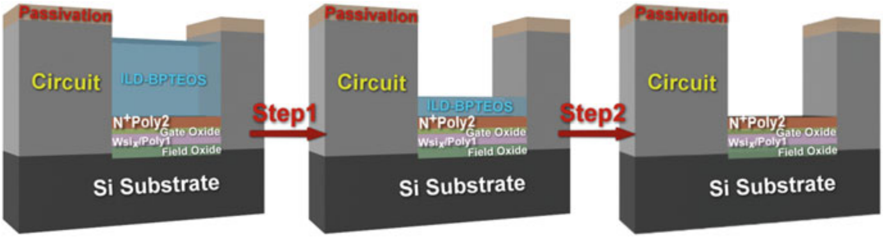


Fig. 8 *Post etching process* (*etch ILD*): Step 1. Dry etch: Reactive Ion Etch (*RIE*), Step 2. Wet etch: Buffered Oxide Etch (*BOE*) [3]

Therefore, the passivation layer above the sensors can be removed in the standard CMOS process to simplify the post-processing. After that, only TEOS oxide is left above the biosensor. To remove most of the TEOS oxide layer above the biosensor quickly and successfully, both dry etching and wet etching methods are utilized. The reactive ion etching (RIE) is first performed by applying trifluoromethane (CHF3) gas to get rid of a great deal of oxide above the sensor. Since both poly-Si NW and on chip circuits may be damaged by long running RIE, wet etching is then performed with buffered hydrofluoric acid (BHF) as the second step of post-etching process to further reduce the oxide thickness to 100 nm. Note that the N+ poly2 exhibits a limited resistance due to its doping concentration so a leakage current through the solid/liquid interface between poly-Si and the aqueous environment could occur and worsen the signal-to-noise ratio (SNR) of the sensing devices. Therefore, a thin oxide layer above the poly-Si NW is essential to reduce the undesired leakage current. The etching process is illustrated in Fig. 8.

1.3 Architecture & Sensing Mechanism

A successful single-chip biosensor system normally relies on a multidisciplinary design covering both sensors and circuits. According to previous works, biosensors made of smaller and thinner low-doping silicon nano-wires can achieve better biomolecular sensitivity. However, the design of silicon nano-wires is not so flexible because most parameters of the devices are restricted by specific process recipes of the manufacturer. To solve the issue, a Wheatstone bridge is adopted as the interface between the sensor and the read-out circuit. Figure 9 illustrates the full-bridge arrangement of the Wheatstone sensor bridge formed by poly-Si NWs. In the full-bridge configuration, two of the poly-Si NWs (R1 and R4) are exposed by post-etching process while the others (R2 and R3) are still covered by the dielectric (oxide) and passivation (nitride) layers. The aminopropyltriethoxysilane (APTES) is used to form amino groups on the surface of the exposed poly-Si NWs so that the probe ssDNA molecules can be immobilized upon the poly-Si NWs by using succinimidyl-4-(N-maleimidomethyl)cyclohexane-1-carboxylate (SMCC) as

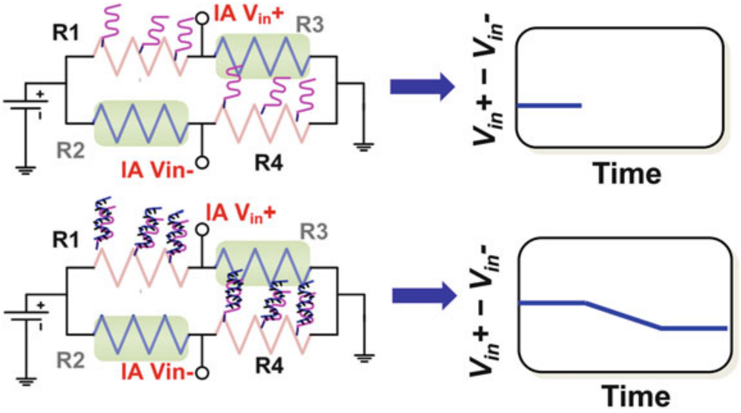


Fig. 9 The full-bridge arrangement of the Wheatstone sensor bridge formed by poly-Si NWs

the linker. Consequently, only the poly-Si NWs R1 and R4 can function as DNA sensors. Since the target ssDNAs carry negative charges in PBS environment, the electrons within the N-type poly-Si NWs would be repelled from the surface of the NW as the target ssDNA hybridizes with the probe ssDNA. Due to the charge repulsion, the hybridization phenomenon decrease the conductance (or increases the resistances) of the poly-Si NWs R1 and R4. Basically, the detection of the complementary target ssDNA that has a specific binding affinity to the HBV probe ssDNA is performed by measuring the variation in the conductance (or the resistance) of the poly-Si NW. The increase in the resistance of R1 and R4 results in the decrease in the output voltage of the Wheatstone bridge ($=\mathrm{IAVin}_{+} - \mathrm{IAVin}_{-}$). This output voltage can be delivered to a readout circuit for further signal processing.

The full-bridge arrangement helps to minimize measurement errors resulted from testing environments and fabrication processes. Moreover, the configuration doubles the differential output signal, as compared with other types of arrangements. To eliminate the measurement errors due to temperature variation, a CMOS temperature sensor can be included to provide temperature calibration for on-chip biomolecular sensing applications. As reported in [2], the temperature sensor shares the same signal path with the poly-Si NW biosensors by using a multiplexer to reduce complexity and power consumption.

1.4 Hydrogel-Based Glucose Sensor

Hydrogel is a cross-linked polymer that can absorb water, as illustrated in Fig. 10a [4]. The most important feature of hydrogels is their ability to swell when put in contact with a thermodynamically compatible solvent like water. The swelling process in hydrogels depends on many factors including the property of the aqueous

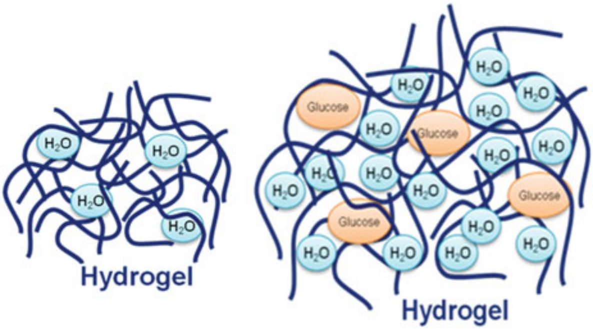


Fig. 10 Absorption and swelling behavior of hydrogel for (**a**) water and (**b**) glucose solution [4]

solution. Consider a glucose-sensitized hydrogel which is absorptive to both glucose and water molecules. When the hydrogel is exposed to a glucose solution to trigger the swelling process, its volume varies with the glucose concentration. The difference in the swelling behavior or the volume change of hydrogels results from that the absorption of glucose molecules helps the polymer chains to relax more so that more H_2O molecules can be filled with to enlarge the volume of the hydrogel, as illustrated in Fig. 10b. Apparently, glucose sensing can be performed by measuring the change in the hydrogel volume, which requires only a drop of the solution. Moreover, the absorption of molecules in hydrogels is actually a diffusion process in which the molecules penetrate into the network of crosslinked polymer chains. The diffusion process is reversible and the hydrogel can conditionally recover from the swelling deformation, which makes the detection mechanism quite suitable for reusable sensors.

The above-mentioned sensing mechanism is applied to a glucose monitoring SoC based on capacitive sensing configuration, as shown in Fig. 11, where its packaging strategy is also depicted [4]. The glucose sensor is implemented with a sandwich structure composed of the glucose-sensitized hydrogel in the middle, an anodic aluminum oxide (AAO) membrane at the top, and a MEMS capacitor at the bottom. As previously mentioned, the volume of the hydrogel varies with the glucose concentration when the hydrogel is exposed to a glucose solution. The volume change then varies the compression force acting on the MEMS capacitor and turns into the change in the air gap of the capacitor. As a consequence, a capacitance variation that depends on the glucose concentration can be observed and delivered to the readout circuit for further processing.

Beneath the hydrogel, the MEMS capacitor is built in a metal-air-metal structure and covered by a polydimethylsilicane (PDMS) material. The top metal plate of the capacitor is a Cr/Au plating deposited on the underside of the PDMS cover, whereas the bottom metal of the capacitor is realized by the top metal layer of the standard CMOS technology. With bisacrylamide introduced as the cross-linker, the glucose-sensitized hydrogel is prepared in situ by a copolymer aqueous solution of methacrylamido phenylboronic acid (20 mol%) and acrylamide

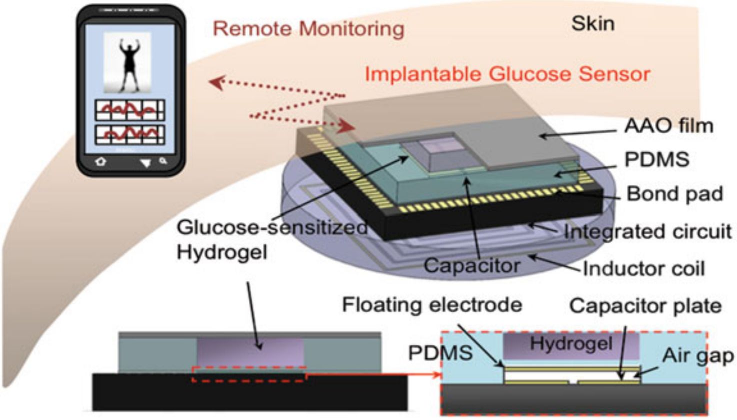


Fig. 11 Schematic of proposed hydrogel-based glucose monitoring SoC and its package [4]

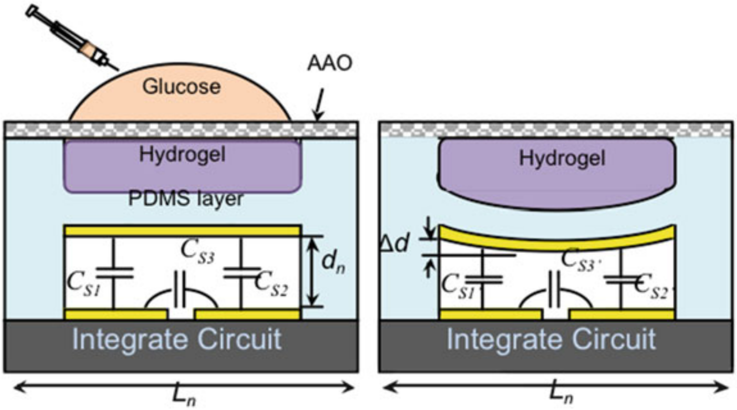


Fig. 12 The principle of hydrogel-based glucose sensor [4]

(80 mol%). A potassium persulfate/tetraethylenediamine redox system is used to initiate polymerization. Allowing only glucose and water molecules to pass through it, the AAO membrane behaves like a filter which blocks bacteria or other impurities to avoid undesired reaction.

The principle of hydrogel-based glucose sensor is illustrated in Fig. 12. Take a close look at this figure, you can find that there are actually two MEMS capacitors inside the sensor and they are connected in series. When the glucose solution seeps through the AAO membrane to react with the hydrogel, the hydrogel starts to swell and squeeze the PDMS layer, which makes the top plate of the capacitors bend and leads to a capacitance variation. The effective capacitance of the device can be calculated as follows. Let C_{S1} and C_{S2} represent the two MEMS capacitors in the sensor. The effective capacitance of each MEMS capacitor before and after applying the glucose solution to the sensor can be expressed as:

$$\mathrm{C}_{S1} = C_{S2} = \frac{\varepsilon_0 \cdot A_t}{2 \cdot (d_n)} \text{ and } \mathrm{C}'_{S1} = C'_{S2} = \frac{\varepsilon_0 \cdot A_t}{2 \cdot (d_n + \Delta d)},$$

where A_t is the effective area of each MEMS capacitor, d_n is the distance between the top and bottom plates and $\mathbf{\Delta}$d is the change in the distance due to the swelling in the hydrogel after applying the glucose solution to the sensor. The effective capacitance of the device before and after applying the glucose solution to the sensor can be expressed as:

$$C_T = \frac{C_{S1} \cdot C_{S2}}{C_{S1} + C_{S2}} + C_{S3} = \frac{\varepsilon_0 \cdot A_t}{4 \cdot (d_n)} + C_{S3} \text{ and } C'_T = \frac{C'_{S1} \cdot C'_{S2}}{C'_{S1} + C'_{S2}} + C'_{S3} = \frac{\varepsilon_0 \cdot A_t}{4 \cdot (d_n + \Delta d)} + C'_{S3},$$

where C_T represents the total capacitance of the device and C_{S3} represents the parasitic capacitance between the bottom plates of the two capacitors. Note that $C_{S3} \sim C_{S3}$'. The change in the distance $\mathbf{\Delta}$d, as well as the change in the effective capacitance of the device, depends on the glucose concentration.

The sensor was applied with glucose solutions of different concentrations from 0 to 240 mM. The experiment results of hydrogel-based glucose sensor are shown in Fig. 13 [4]. As expected, the capacitance of the glucose sensor increases with the glucose concentration. The sensor exhibits the capacitance of 1.1 pF for a concentration of 0 mM and gradually increases as the glucose concentration is increased. The capacitance reaches 2.68 pF for the glucose concentration of 240 mM and seems saturated due to a limit on the volume expansion of the hydrogel.

To conduct a long-term test on the hydrogel-based glucose sensor, a stand-alone sensor is tested every 15 min and applied with a drop of glucose solution whose

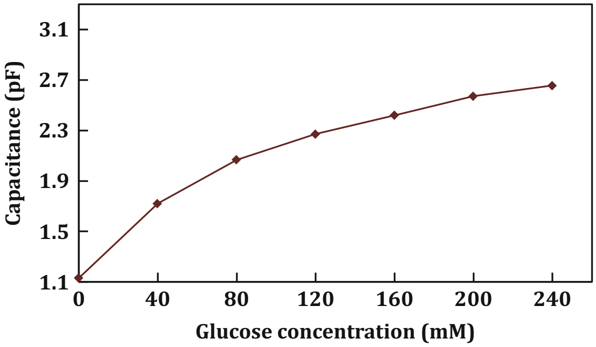


Fig. 13 Measurement result of sensor capacitance vs. glucose concentration [4]

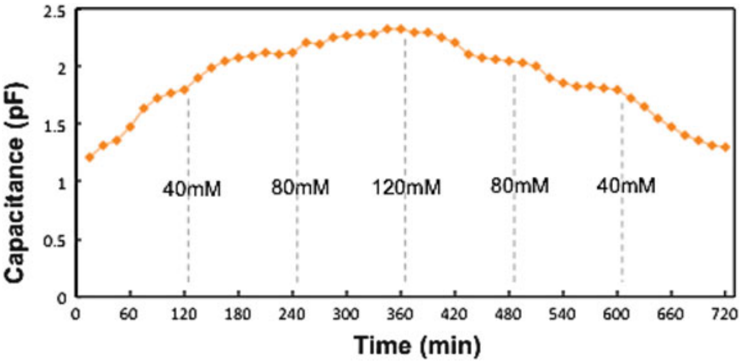


Fig. 14 The measured capacitance of the hydrogel-based sensor varies with the glucose concentration [4]

concentration is gradually increased from 5 to 120 mM and then decreased from 120 mM back to 5 mM. According to the experiment results shown in Fig. 14 [4], the capacitance of the sensor varies with glucose concentration for 12 h. These results prove that the hydrogel-based glucose sensor is reusable and can continuously operate for a long time.

1.5 Ion-Sensitive-FET (ISFET) Based pH Sensor

ISFET based sensors can be realized by using CMOS technologies with either open-gate or floating-gate structure, where open-gate ISFETs can achieve higher sensitivity than floating-gate ISFETs [5]. The penalty of using open-gate ISFETs is an extra post-process required to remove the poly-silicon gate. It becomes quite challenging when the previously mentioned poly-silicon NW-based DNA sensor is also integrated to achieve a multi-sensing system. Note that the open-gate ISFETs would be realized with the first poly-silicon layer (poly1) while the poly-silicon NW-based sensor would be made of the second poly-silicon layer (poly2) by using a CMOS technology with two poly-silicon layers, such as the previously mentioned TSMC 2P4M 0.35-μm CMOS process. Moreover, the post-process needs to be performed without causing severe damage to the thin gate oxide layer under the first poly-silicon layer so that the poly-silicon NW-based sensor under the oxide can be well protected.

Alternatively, the floating-gate ISFET can be developed as a pH sensor to save the trouble mentioned above. The passivation layer made of silicon nitride in a CMOS technology can function as a pH sensitive membrane so that the floating-gate ISFET pH sensor can be realized without any post-IC process. The cross-section view of an floating-gate ISFET pH sensor is depicted in Fig. 15 [5]. Notably, the sensitivity depends on the sensing area which can be increased without enlarging the

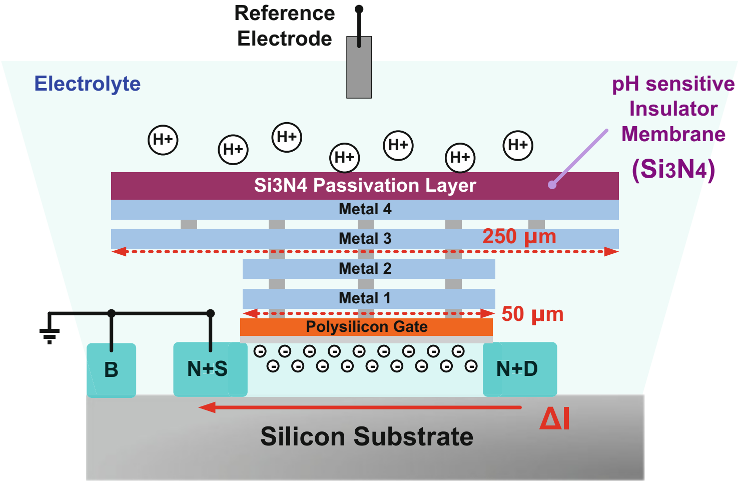


Fig. 15 The cross-section view of the ion-sensitive FET (ISFET) pH sensor [5]

on-chip transistor device (50/50 µm). The top metal area as well as the above Si_3N_4 membrane occupied an area of 250 µm × 250 µm. The sensing mechanism of the open-gate ISFET pH sensor is also illustrated in Fig. 15. The ISFET is configured as a normal NMOS current source with the gate voltage defined by a reference electrode. When the ambient concentration of the hydrogen ion H+ increases (or the pH value falls), the increase in the number of these positively charged ions will decrease the threshold voltage of the ISFET and hence increase its output current.

The pH value of human blood is normally regulated within a narrow range from 7.35 to 7.45. Therefore, a proper range for the pH meter would be from 6 to 8 with a resolution of 0.01 (~8-bit resolution). During the measurement, the current of the ISFET is first adjusted to ~3 nA with a neutral buffer solution (pH = 7). Then the ISFET current would vary with the pH values of solutions under test (from 6 to 8) with the maximum variation of ~3 nA, as shown in Fig. 16.

1.6 Bandgap-Based Temperature Sensor

Bandgap references are widely adopted to provide accurate bias voltage and can be used to realize temperature sensors in standard CMOS process due to its adjustable temperature coefficient. A temperature sensor is realized by using a bandgap reference circuit consisting of vertical pnp BJTs, as shown in Fig. 17 [5]. Conventionally, a temperature-independent bandgap reference voltage (V_{REF}) is generated by combining a PTAT voltage with a base-emitter voltage which exhibits a negative temperature coefficient. PMOSET current mirror pairs $M_{1\text{-}2}$ and $M_{3\text{-}4}$ form a supply independent bias, which ensures that $V_x = V_y$. Both the drain current

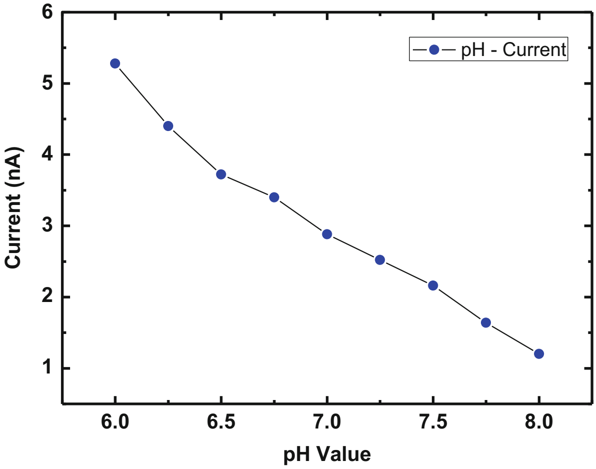


Fig. 16 The ISFET current versus the pH value from 6 to 8

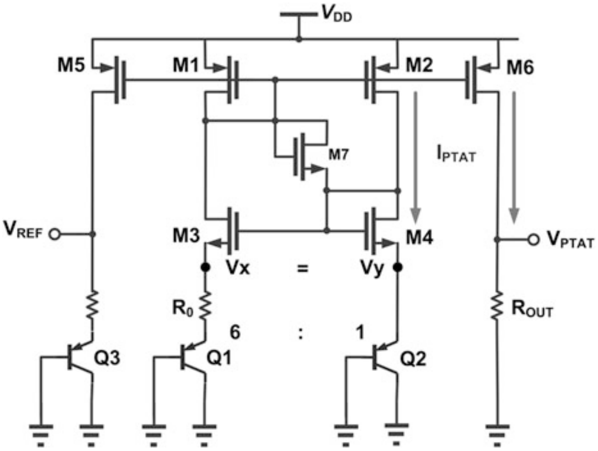


Fig. 17 Schematic of the bandgap reference temperature sensor [5]

of M_1–M_6 (I_{PTAT}) and the bias voltage V_{PTAT} are proportional to the absolute temperature (PTAT) and can be expressed as:

$$I_{PTAT} = I_{D1-4} = I_{D5,6} = \frac{V_{BE2}-V_{BE1}}{R_0} = \frac{V_T \ \ln \ n}{R_0}$$
$$V_{PTAT} = I_{PTAT} \times R_{OUT} = V_T \ \ln \ n \times \frac{R_{OUT}}{R_0} \propto V_T \propto T,$$

where $V_T \ln n$ is the difference between the base-emitter voltages of the two BJTs operating at different current densities. Resistors R_0 and R_{OUT} are implemented with the same type of resistors (poly2 resistor) to avoid unwanted effects resulted from the temperature coefficient or the process variation of resistors.

Since the human body temperature is about 37 °C, the output voltage (V_{PTAT}) at 37 °C is designed to be 0.9 V with a temperature coefficient of 3.78 mV/°C. Note that a simple temperature sensor like this can be used to perform temperature compensation on other sensors. The output of the temperature sensor can be delivered to a readout circuit followed by an ADC to provide temperature information. As presented in [3], the temperature information would be used to reduce errors due to temperature drifts in the output signals of poly-silicon nanowire and ISFET sensors. The required number of bit for ADC and the noise requirement of the readout circuit can be considered as follows. To provide the required dynamic range of an implantable system that operates in a temperature range from 32 to 42 °C, a resolution about 0.05 °C can be achieved with 8 bit. For a limit of detection (LOD) referring to an SNR of 3, the input referred noise of the sensor readout should be no more than 44.5 μV_{rms} so that the minimum output voltage change of 189 μV can be detected at the precision of 0.05 °C. If a sampling frequency of 200 Hz is adopted, the input referred noise of the sensor readout should be no more than 3.15 $\mu V/\sqrt{Hz}$ over 200 Hz. When the sensor is used to monitor the ambient temperature, where the system operates in a wider temperature range from −20 to 120 °C, then it requires a larger number of bits to achieve the reasonable dynamic range and resolution. For example, a resolution of 0.5 °C should be achieved by at least 9 bit. Therefore, a 10-bit ADC and a readout circuit with the programmable gain from 0 to 40 dB are designed to accommodate all the requests with safety margin in these works [3, 5]. For the better accuracy, the bandgap-based temperature sensors can also be improved by using dynamic element matching and auto-calibration, as discussed in [6].

2 Readout Circuits

Most of the time, the sensing results from the sensor would be some tiny little physical quantities accompanied by interfering noise. As the interface between a sensor and the sensor SoC, a readout circuit needs to provide proper amplification for these signals. To achieve that, it should contribute the least noise and distortion while consuming the lowest power and chip area. In this section, the design of readout circuits would be introduced.

2.1 Reconfigurable Multi-Sensor Readout Circuit

In practical applications, it is necessary to monitor several biomedical signals simultaneously for patients with severe heart diseases or lung failure. Physicians need to handle different types of monitoring systems immediately, which is inefficient and could be quite dangerous under those urgent circumstances. An adaptive transducer readout that supports four types of signal acquisition (C, R, I, V) has been proposed for multi-parameter sensing [7]. However, the multi-sensing system in [7] requires four independent interface circuits, resulting in poor hardware efficiency.

A reconfigurable sensor interface which can be shared among four sensors can be realized by using the switched-capacitor topology. Switched-capacitor-based readout circuits have several advantages over other classes of readout circuits [8, 9]. First, all types of signals can be easily transformed into charges stored in the capacitor of the switched-capacitor circuit in the sampling mode, which improves the hardware efficiency since it requires only one sensor interface to cope with many different types of sensors. Second, the fundamental building blocks of the switched-capacitor circuits deal with capacitive loads, instead of resistive loads that require power-hungry output buffers and also generate extra thermal noise.

The correlated double sampling (CDS) technique is utilized to reduce non-ideal effects such as dc offsets and the inherent flicker noise of CMOS amplifiers. Although the chopper stabilization (CHS) technique is adopted in many low-noise sensor interfaces [8], it is not compatible with a sampled system and hence it is not suitable for the switched-capacitor-based readout circuit. Moreover, the CDS technique is superior in several aspects, as compared with the CHS technique. By using the CDS technique, the dc offsets and the flicker noise are directly eliminated at the output of the amplifier where they still exist as modulated noises if the CHS technique is adopted. Consequently, the amplifier that adopts the CDS technique can handle a larger output swing, which is beneficial to a low voltage design. Besides, the CHS technique requires a high-quality low-frequency low-pass filter to suppress the modulated dc offsets and flicker noise, which increases both the power consumption and the die area.

Figure 18 shows the schematic of the reconfigurable switched-capacitor-based readout circuit. There are two stages in the readout circuit. The first stage is a reconfigurable multi-sensor interface consisting of a noninverting switched capacitor amplifier and a multiplexer. Notably, the noninverting switched capacitor amplifier is based on the differential to single architecture. The differential input configuration provides the common-mode rejection of the circuit (CMRR = 164 dB), which helps eliminate the unwanted effects of CMOS switches such as charge injection and clock feed-through. By manipulating the switches ($\varphi_{C,R,I,V}$) in the multiplexer properly, the interface can be reconfigured to cope with any kind of sensor. During cyclic operations, time-division multiplexing (TDM) is adopted to control the switches ($\varphi_{C,R,I,V}$) and this interface will be reconfigured four times every 250 ms to sequentially process four types of signals from the corresponding sensors.

The conversion principles for four types of signals (voltage, current, capacitance and resistance) are illustrated in Fig. 19a–d. In the sampling mode (φ_1), one of the four input signals will be converted into the charges stored in C_1 (10.6 pF), and then it will be transformed into an output voltage through charge redistribution over C_2 (10 pF) in the amplification mode (φ_2). Let ΔQ_1 represent the amount of charges stored in C_1 by the end of the sampling mode. In the amplification mode, the charges gained by C_2 would be equal to the charges previously sampled by C_1, so the output voltage can be expressed as $\Delta Q_1/C_2$. Note that any dc-offset would be differentially stored in capacitors C_2 in the sampling mode and it would be subtracted and cancelled in the amplification mode.

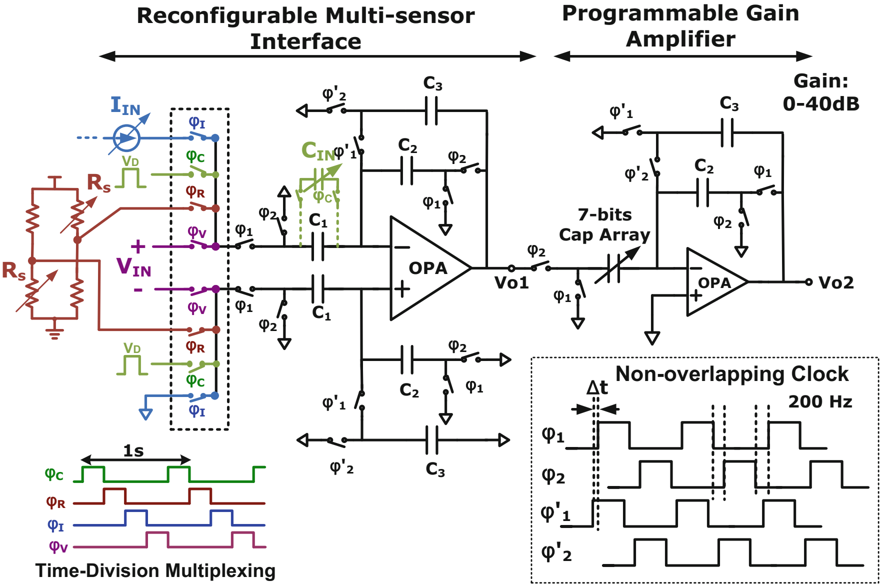


Fig. 18 Schematic of the reconfigurable readout circuit [5]

For voltage signals (V_{IN}), the interface behaves as a conventional noninverting switched capacitor amplifier, and hence the voltage gain of 1.06 V/V can be estimated from the capacitor ratio (C_1/C_2). For current signals (I_{IN}), ΔQ_1 is equal to the product of the input current I_{IN} and the sampling time (T_s), and hence the transresistance gain of 250 MΩ can be estimated for the sampling time of 2.5 ms. It is worth mentioning that the sampling clock rate ($\varphi_{1,2}$) should be set faster than 200 Hz to make a trade-off between the system requirement and the circuit limitation. As the output voltage is inversely proportional to the clock rate and the value of C_2, long clock periods or small capacitors can lead to very large output voltages, which may cause nonlinear characteristics. To cope with an input current of 3 nA with a proper capacitor value ($C_2 = 10$ pF) in the interface, 200 Hz is the minimum clock rate that can prevent the amplifier from saturation. Additionally, the clock rate should exceed the flicker noise corner frequency of the CMOS amplifier in the vicinity of 100 Hz.

To measure the capacitance variation, one of the sampling capacitors (C_1) is replaced by the capacitive sensor (C_{IN}). Then an identical voltage difference ($V_{DD} - V_{cm}$, i.e. 0.9 V) is intentionally fed to both C_{IN} and C_1 for sampling. The difference between C_{IN} and C_1 will cause different amount of charges stored in these sampling capacitors in the sampling mode and hence results in different output voltages in the amplification mode. The conversion gain of 43.7 mV/pF can be estimated from the expression: $(V_{DD} - V_{cm})/(C_1 + C_2)$. Because the unstable post-process of the MEMS capacitive sensor causes inevitable capacitance offsets, a 3-bit capacitor array with the least significant bit capacitance of 0.75 pF is incorporated

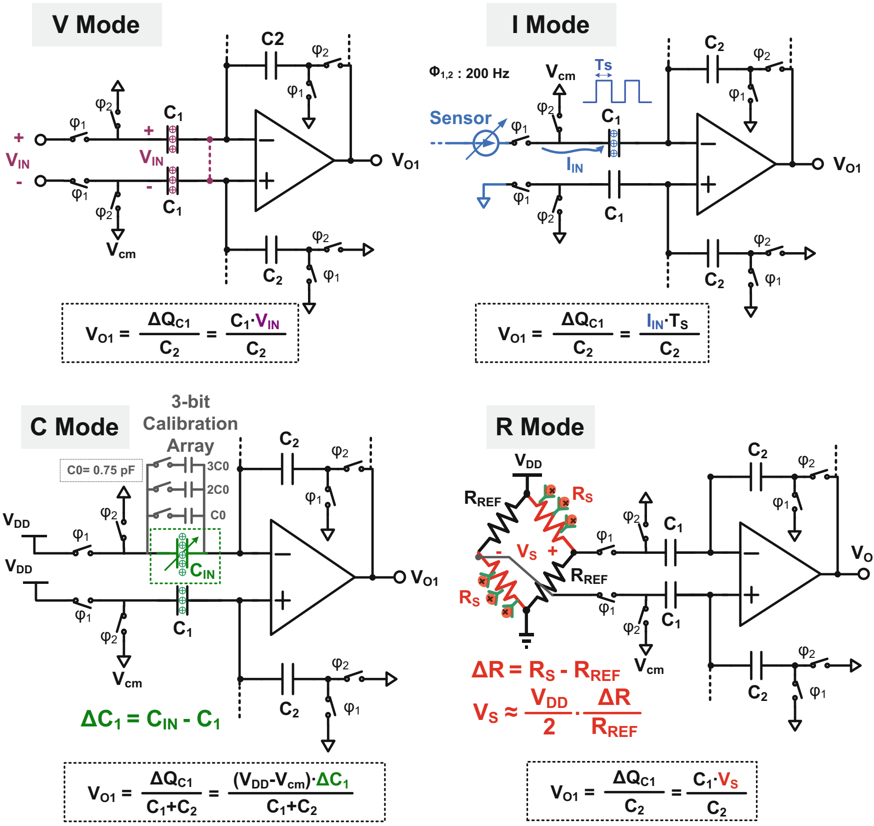


Fig. 19 Conversion principles for (**a**) V-type, (**b**) I-type, (**c**) C-type, and (**d**) R-type inputs [5]

into the sampling capacitor (C_{IN}) to perform calibration. To observe the resistance variation, the resistive sensors (R_S) with reference resistors are configured as a Wheatstone bridge (as mentioned in Sect. 1.1) to convert the resistance variation into a voltage signal ($V_S = V_{DD}/2R_{REF}$) first. Then as a voltage signal, it can be handled by the conventional switched capacitor amplifier with a voltage gain of C_1/C_2. The overall conversion gain of 95.4 mV/kΩ can be estimated from the expression V_SC_1/C_2 for $R_{REF} = 10$ kΩ and $V_{DD} = 1.8$ V.

The second stage of the readout circuit is a programmable-gain switched capacitor amplifier which is also based on the noninverting configuration and employs a 7-bit binary-weighted capacitor array as the input capacitor. With respective conversion gains offered by the first stage, each type of input signal would be converted to a voltage which then can be properly enlarged by the programmable-gain amplifier to enhance the sensitivity. The 7-bit capacitor array is designed with the least significant bit capacitance of 0.1 pF to provide linear gain control. The variable gain can be adjusted from 0 to 40 dB, which extends the input dynamic range of the readout circuit. All the switches are realized as transmission gates to minimize non-ideal effects. Particularly, a non-overlapping clock scheme must be used in this switched-capacitor-based readout circuit to avoid inaccurate charge sharing between capacitors.

The readout circuit suffers from noise folding due to the sampling process in the CDS technique, as other circuits that adopt sampling techniques. The uncorrelated thermal noise would be sampled and hence the thermal noise power at least doubles [10]. Apparently, the thermal noise of the CMOS amplifier needs to be reduced to alleviate the above mentioned problem. A large transconductance (g_m) of the input transistors is preferred. Moreover, large sampling capacitors are advantageous as the integral noise power over the Nyquist interval is limited to KT/C. As previously mentioned, the readout circuit employs the sampling capacitor of 10 pF, and hence the noise power KT/C is 411.4×10^{-12} V^2_{rms}, i.e. 20.3 μV_{rms}. Therefore, the readout circuit can handle the temperature sensor with the resolution of 0.05 °C.

Instead of using reset switches, the capacitor C_3 is connected between the input and output nodes of the op-amps for both stages of the readout circuit in the sampling mode, which prevents dramatic changes in the output voltages and thus relaxes the slew rate requirement of the op-amps. However, the speed of the readout circuit also limits the time-division-multiplexing (TDM) among different sensors. When the readout circuit is reconfigured to handle a different type of signal, it takes at least twelve clock cycles (~60 ms) for the readout circuit to deliver a reliable result, which explains why it takes the interface 250 ms to sequentially process four types of signals from the corresponding sensors. Moreover, the frequency of input signals (<0.1 mHz for the sensors in Sect. 1.1) is much lower than half the sampling frequency (>200 Hz), the voltage at the input terminals of the op amp would not change appreciably from one clock phase to the next due to oversampling. Therefore, the input signal at input terminal of the op amp, as a slowly varying signal, would be nearly cancelled by the CDS switching of C_1 and C_2. For the op amps with a finite gain A, the dc gain of the switched capacitor amplifier can be expressed as:

$$\frac{V_{OUT}}{V_{IN}} = \frac{C_1}{C_2}\left(1 + \frac{C_1 + C_2}{C_2 A^2}\right)^{-1}$$

Notably, the dc gain does not depend on the capacitors C_3, and the gain error is proportional to A^{-2}, indicating that the effect of the finite op-amp gain is reduced. The gain of op amp is designed as large as 110 dB to achieve the negligible gain error for the readout circuit. The op-amps employ PMOS input pairs with long channel ($L \geq 1$ μm) to mitigate the effect of flicker noise. Moreover, all the transistors in the op-amps operate in the sub-threshold region for better current efficiency.

2.2 Capacitive Readout Circuit

The capacitance changes of the glucose sensor mentioned in Sect. 1.1 can be measured by the readout circuit shown in Fig. 20. The readout circuit consists of a current generator, a capacitance-to-frequency converter (C–F converter) [4], and a digital counting amplifier (DCA) processor. The DCA processor, through its control over the current generator, forms a feedback path to improve the detection resolution.

The C–F converter is aimed to convert a capacitance to a periodic pulse signal with a corresponding frequency. Figure 21 shows the schematic of the C–F converter which basically employs a sensing clock loop (SCL) and analog bio-signals from the capacitive sensor would be converted to digital signals in a dynamic comparator by using similar approaches proposed in [11, 12]. As a variable capacitor, the capacitive glucose sensor is connected with a constant on-chip capacitor (C_{int}) in parallel whereas the current generator functions as a dc current source (I_{source}) to charge the capacitors. One terminal of the capacitor is terminated to ground so the charging current is applied to the other terminal from which the output voltage of the capacitor is observed. A reference voltage Vsw is applied to the negative terminal of the comparator.

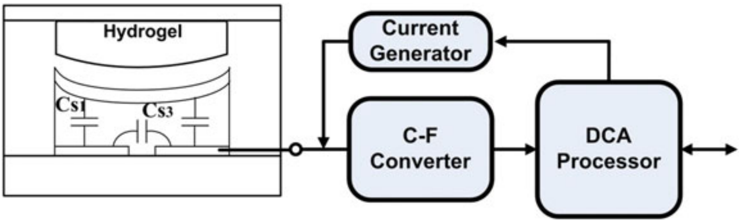


Fig. 20 The readout circuit for capacitive sensors [4]

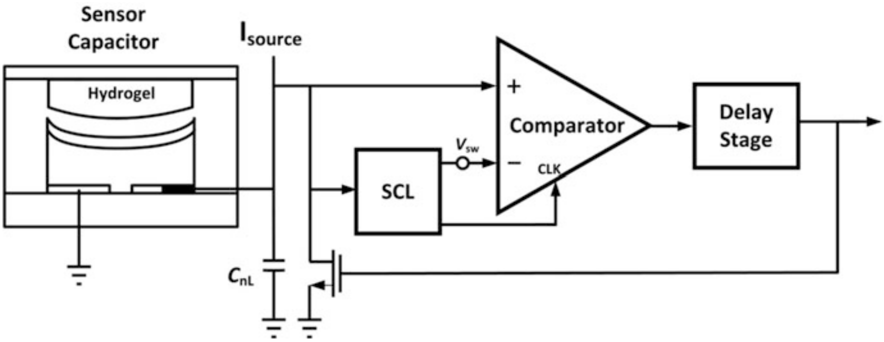


Fig. 21 Schematic of C–F converter [4]

As shown in Fig. 21, the output voltage of the capacitor is delivered to the positive terminal of the comparator. When the output voltage of the capacitor exceeds Vsw, the output state of the comparator will be switched to high. The transition will be delivered to the delay stage and after a while, an NMOS would be turned on by the delayed transition to discharge these capacitors. Due to the time delay, the output voltage is higher than the reference voltage Vsw at the beginning of discharging process. So it also takes a while for the output voltage to fall below the reference voltage and changes the state of the comparator from high to low. Again, the transition would be delayed first and then sent to the NMOS to break the discharging path so that the current source can start to charge the capacitors whose output voltage is below the reference voltage Vsw due to the time delay. Through the processes mentioned above, periodic pulses would be generated at the output of the comparator. Notably, the output voltage of the comparator changes at the rate that depends on the capacitance. Therefore, the pulse width and the interval between two consecutive pulses can be used to measure the effective capacitance.

The measurement results of the C–F converter are shown in Fig. 22. According to the previous discussion, the period/frequency of the pulse waveform depends on the capacitance observed by the C–F converter. When glucose sensor is disconnected from the on-chip capacitor C_{int}, the output frequency of C–F converter is 4.63 Hz. When the capacitive glucose sensor is connected with C_{int} in parallel, the frequency of the CF-converter shifts to 3.76 Hz. Figure 23 shows the operation principle of DCA, where a faster clock is employed to measure the time interval between two consecutive pulses that links to a distinct capacitance. The capacitance resolution of the readout circuit can be improved by adjusting the clock rate.

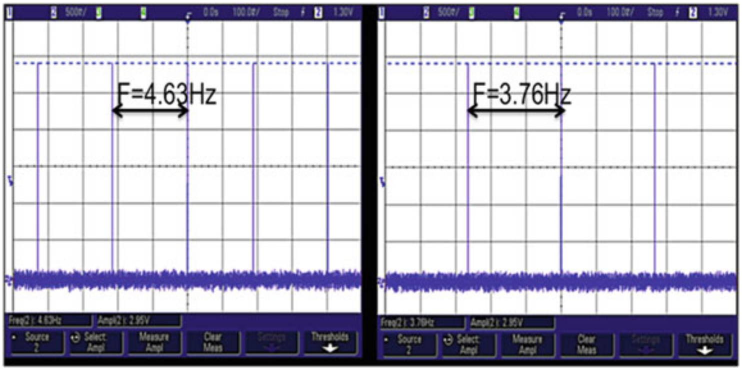
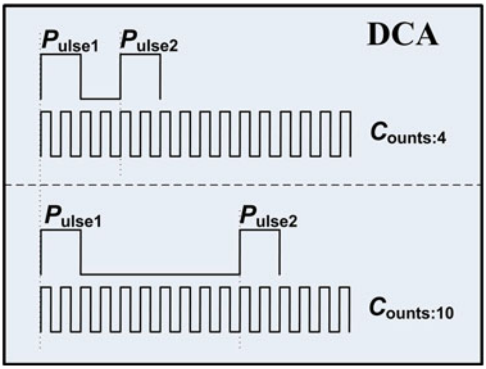


Fig. 22 C–F converter measurement results [4]. *Left*: stand alone on-chip C_{int}. *Right*: C_{int} and the capacitive sensor in parallel

Fig. 23 Operation principle of DCA [4]



2.3 *Oscillator-Based Self-Calibrated Readout Circuit*

Resistive sensors deliver resistance variations as output signals which require readout circuits to perform measurement on resistances. As mentioned in Sect. 1.1, the cantilever DNA sensor requires a readout circuit that recognizes the change in its resistance before and after DNA hybridization. Many approaches have been proposed to measure the resistance variation of sensors [13–16]. As mentioned in Sect. 1.1, the Wheatstone bridge is adopted to convert the resistance variation into a voltage output signal. Unfortunately, the mismatch and the offset problems pose severe design difficulties [17]. By observing the current that flows through the resistance, the resistance measuring method for achieving a wide dynamic range is demonstrated in [18]. However, the precision of the readout circuit is insufficient for a sensor exhibiting tiny resistance variation that is less than 0.02 % of the original resistance.

An oscillator-based readout circuit with self-calibration can be used to measure the tiny resistance variation. As depicted in Fig. 24, the readout circuit is formed from a sensor-merged oscillator, a buffer, a divider, a mixer, a frequency-to-digital (F-to-D) converter, and a calibration controller. The sensor-merged oscillator is a ring oscillator comprised of three different delay cells. The resistive sensor, ex. a piezoresistive DNA sensor, is embedded in one of the delay cells as a variable resistor, and thus the oscillation period would be linear with its resistance. In particular, for small resistance variations, the change in the oscillator frequency (F_{SENSOR}) is approximately linear with the resistance variation of the sensor. The output signal of the oscillator is first passed to a buffer stage to be extended to a rail-to-rail signal. Then the rail-to-rail signal is delivered to a divide-by-4 circuit which provides an output clock (F_{DIV}) with a perfect duty cycle of 50 %. The output clock would be delivered to the mixer while the other input terminal of the mixer is applied with an external clock (F_{CLK}). A simple D flip-flop (DFF) is employed as a mixer to provide frequency down-conversion. The output clock of the divider (F_{DIV}) is multiplied with the external clock (F_{CLK}) by the mixer to increase the relative frequency variation. Due to the simple mixer architecture, the port-to-port

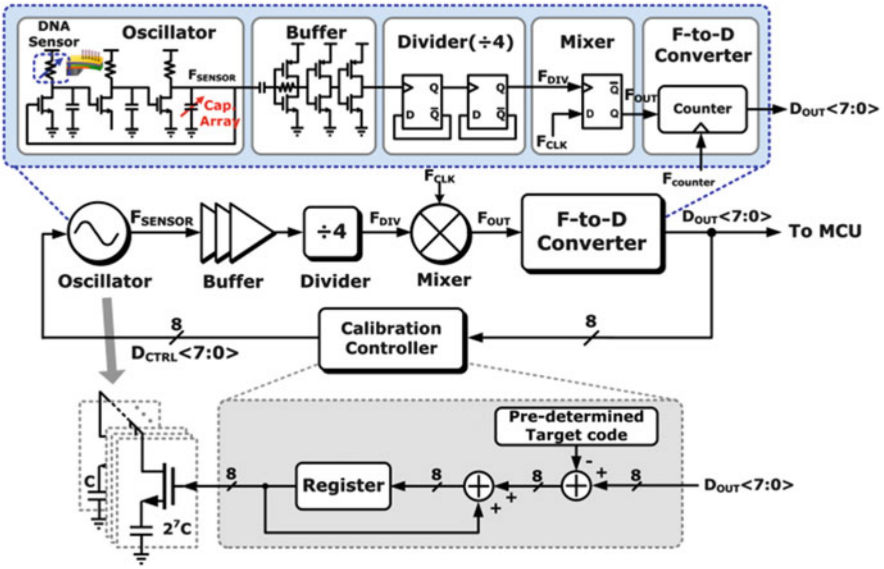


Fig. 24 Architecture and simplified circuits of the oscillator-based self-calibrated readout circuit [1]

isolation of the mixer is very poor. If the frequency of the external clock is close to the frequency of the oscillator, the external clock could cause frequency pulling to the oscillator, which retards the sensing function. In other words, the divider plays an important part in minimizing the frequency pulling in the oscillator.

Simulations are performed with/without the divider to observe the frequency of the down-converted signal versus the frequency difference between the two input signals of the mixer. For comparison, the simulation results are shown in Fig. 25, proving that the accuracy can be effectively improved by incorporating the divider into the readout system.

The mixer is followed by a counter which converts the frequency of the down-converted signal (F_{OUT}) into an 8-bit digital output (D_{OUT}). Let $F_{COUNTER}$ represent the frequency of the counter, then the digital output (D_{OUT}) can be expressed as:

$$D_{OUT} = \frac{|T_{OUT} - T_{0,DN}|}{T_{counter}} = \left| \frac{F_{counter}}{F_{OUT}} - \frac{F_{counter}}{F_{0,DN}} \right| = \left| \frac{F_{counter}}{F_0 + \Delta F} - D_0 \right|, \quad (1)$$

where $F_{0,DN}$ ($T_{0,DN}$) is the initial frequency (period) of the down-converted signal before the sensing process begins, and ΔF is the frequency variation due to the resistance change of the sensor after the sensing process completes. You can tell that FDC actually plays the role of an analog-to-digital converter (ADC) in this oscillator-based readout system. Moreover, the FDC digitizes the absolute frequency change of the down-converted signal before and after the sensing process. Therefore, the readout system can cope with either an increase or a decrease in the resistance of the sensor, as well as in the output frequency.

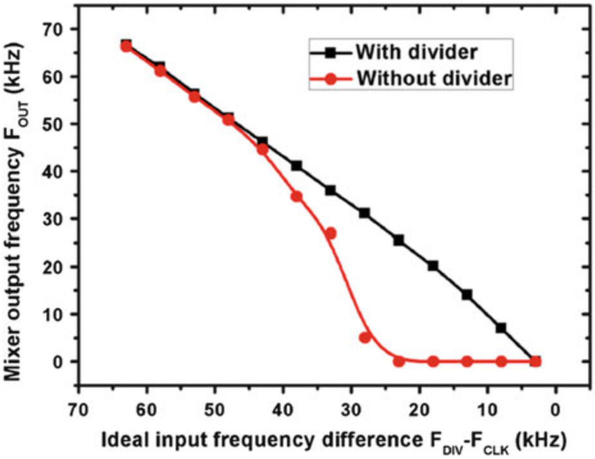


Fig. 25 Simulated results of frequency pulling effect [1]

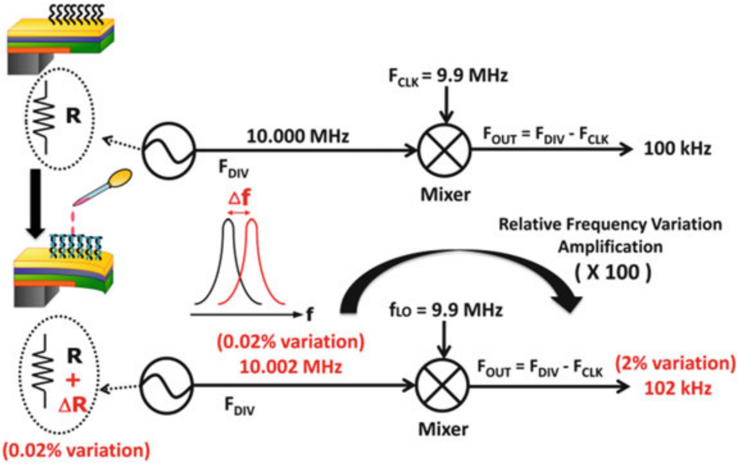


Fig. 26 An example to explain the functionality of the added mixer [1]

As previously mentioned, the mixer transfers the divider output signal (F_{DIV}) down to a lower frequency band (F_{OUTPUT}) to increase the relative frequency variation. An illustration shown in Fig. 26 is used to explain the method in more detail. For simplicity, assume that the divider delivers an output frequency of 10 MHz ($F_{DIV} = 10$ MHz). After the sensing process (DNA hybridization) completes, the piezoresistive sensor exhibits a very small resistance variation of 0.02 %, which leads to a frequency shift of 2 kHz at the divider output (F_{DIV}). Apparently, the frequency shift is much smaller than the original frequency of 10 MHz, which makes the following frequency discrimination quite difficult.

To solve the problem, the frequency band is down-converted to a relatively lower band by using a mixer. Applied with an external clock (F_{CLK}) of 9.9 MHz, the mixer converts the output signal of the divider from 10 MHz to 100 kHz, while the frequency shift due to the sensing process (DNA hybridization) remains unchanged, so that the relative frequency variation is enlarged by 100 times. This method efficiently improves the sensitivity of the system and thus the design constraints in other aspects can be significantly relaxed.

In the previous work [1], the ring oscillator is designed to operate at the frequency around 24 MHz, where the piezoresistive DNA sensor presents the resistance around 3 kΩ. So the divider delivers the frequency F_{DIV} around 6 MHz which is then down-converted to the output frequency F_{OUT} of 200 kHz. As previously mentioned, the output frequency would be translated into the 8-bit digital output by the FDC. In order to support a variety of applications, the clock rate of the counter (FDC) is adjustable so that a wide range of frequency variations can be handled by the readout system. During the measurement, the readout circuit is capable of detecting a small frequency shift of 1.2 kHz, less than 0.02 % in terms of the relative variation. Notably, the tiny resistance variation of 0.6 Ω in the sensor can be estimated from the frequency shift.

A self-calibration mechanism is employed to compensate for the process-voltage-temperature (PVT) variations. As depicted in Fig. 24, a capacitor array is incorporated into one of the delay cells. During calibration, the RC time constant of this delay cell would be configured to adjust the oscillation frequency, as well as the output frequency of the mixer (F_{OUT}). Through altering the control code of the capacitor array D_{CTRL}<7:0>, the FDC output would gradually approach a pre-determined code that corresponds to the condition for $F_{OUT} \sim 200$ kHz, the desired initial condition of the readout circuit after calibration. It is worth mentioning that the temperature effect during the sensing process should be considered because both the piezoresistive sensor and the oscillator-based readout circuit are susceptible to the temperature. When the piezoresistive sensor is embedded in the readout circuit, a relative temperature coefficient of 1,074 ppm/°C can be observed during the simulation, which means only a slight temperature variation of 0.186 °C could yield the same output result as the one from the resistance variation of 0.6 Ω and lead to malfunction. For practical use, a temperature sensor would be required to perform temperature calibration on the readout circuit.

3 Sensor System-on-Chip (SoC)

3.1 Cantilever-Based Label-Free DNA SoC for Hepatitis B Virus Detection

3.1.1 System Introduction

Figure 27 depicts the system block diagram of a wireless DNA detection system. Cantilever-based DNA sensors, an oscillator-based self-calibrated readout circuit, a microcontroller unit (MCU), voltage regulators, and an on–off keying (OOK) transceiver are monolithically integrated. For most of the time, the system operates in the standby mode in which only the OOK receiver and part of the MCU are turned on to listen to commands, therefore, the average power consumption can be significantly reduced. Once meaningful wireless command signals from external mobile devices such as laptops are received, the OOK receiver demodulates the commands and delivers them to the MCU to set up the system parameters and put the system in the readout mode.

In the readout mode, most building blocks of the system are awaked, whereas the receiver is turned off. The change in the resistance of the cantilever-based DNA sensor is first transformed into a frequency shift by the sensor embedded ring oscillator of the readout circuit. Then the frequency shift is converted into an 8-bit digital output signal by the frequency-to-digital (F-to-D) converter. The 8-bit digital output signal is further translated into data with RS232 format by the MCU. Through OOK modulation, the RF carrier is combined with the data in the on-chip transmitter. The modulated signals would be wirelessly delivered to other external mobile devices.

3.1.2 Experiment Results

The DNA SoC was realized in a CMOS Bio-MEMS Process. Figure 28 shows the chip micrograph and the DNA SoC occupies an area of $5 \times 6.08\ \text{mm}^2$. To investigate

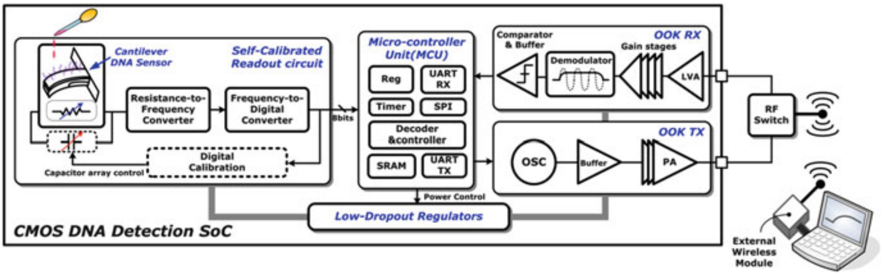


Fig. 27 The system block diagram of a wireless DNA detection system [1]

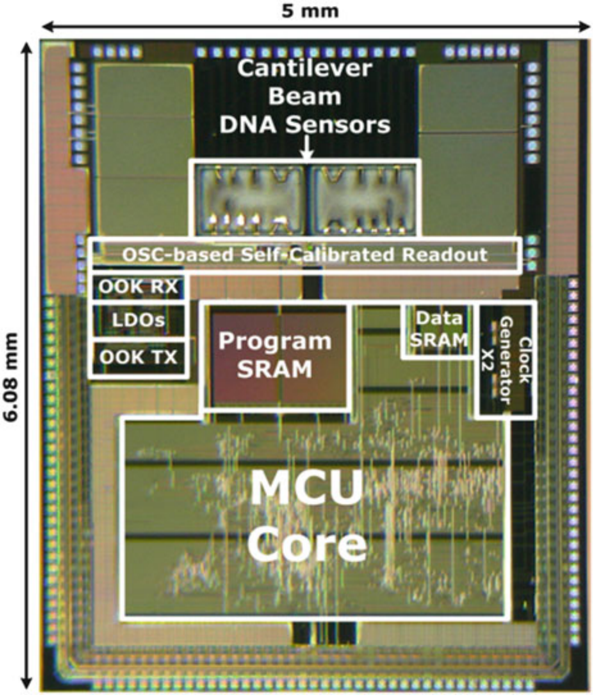


Fig. 28 Chip micrograph [1]

the device sensitivity, DNA sensors were implemented with different structures of micro-cantilevers on this chip. All the circuits need to be well protected during the post process required for the development of the DNA sensors, which ensures the normal function of the SoC during the following experiments. Table 1 summarizes the measured performance of the DNA detection SoC. With high level of integration and the wireless capacity, the system achieves the detectable DNA concentration lower than 1 pM.

The flowchart of the experiment is depicted in Fig. 29. First, the probe DNAs are immobilized on the top of the cantilever to implement the sensor, which takes about 2 h. Then, the DNA sensor is immersed into a solution of phosphate buffered saline (PBS buffer) for experiment initialization. And then, a sample of DNAs is injected to see how many DNAs in the sample can hybridize with the probe DNAs. After waiting for 15 min which is long enough for the DNA hybridization, rinse the sensor with PBS buffer for 10 min to reduce non-specific binding. This step ensures that the stress change of cantilever is mainly caused by DNA hybridization (matched DNAs), rather than non-specific DNA binding (unmatched DNAs). Finally, the sensing chamber is dried for 20 min to obtain steady signals. During the experiment, the temperature was maintained at the room temperature (=25 °C) by a temperature conditioner.

Table 1 Performance summary

Technology	TSMC 0.35 μm Bio-MEMS CMOS
Chip area	30.4 mm^2
DNA sensor	
Structure	Cantilever beam
Dimension (L/W/H)	150/40/3.26 (μm) (selected for experiments)
Resistance	3 kΩ
Resistance variation (DNA hybridization)	<0.02 % (0.6 Ω)
Self-calibrated readout circuit	
Power consumption	1 mW @ 3 V
OSC frequency	24 MHz
Detection sensitivity	2 kHz/Ω
Detection limit (SNR = 3)	1.2 kHz (0.02 %)
Microcontroller unit	
Clock rate	4 MHz
# of instruction	35
Power consumption	Readout mode: 5.7 mW @ 3 VStandby mode: 225 μW @ 3 V
OOK receiver	
Operational frequency	315/402/433.92 MHz
Sensitivity	−62 dBm @ 403 MHz
Power consumption	7.2 mW @ 1.8 V
OOK transmitter	
Carrier frequency	315/402/433.92 MHz
Output power	−6.4 dBm
Power consumption	11.9 mW @ 3 V

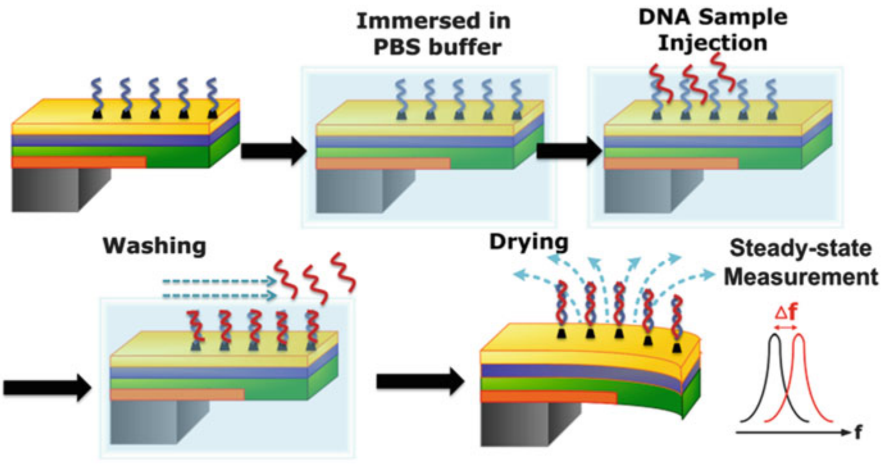


Fig. 29 The flowchart of the experiment [1]

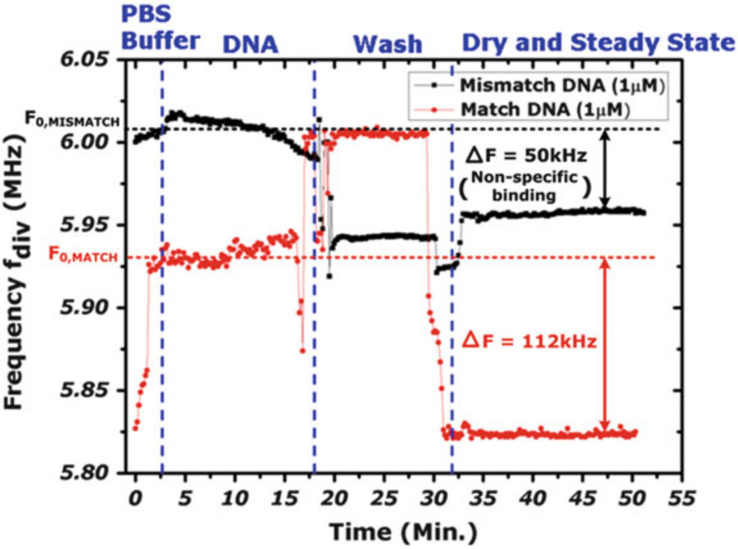


Fig. 30 The divider output frequency is measured over time for experiments on match and mismatch DNA samples [1]

Preliminary Test

To examine the function of the DNA sensor SoC, a preliminary experiment was conducted. In this preliminary experiment, the 5′ thiol-modified DNA (DNA sequence: 5′-HS-ATAGGTCGGTAGGTGAATGG-3′) was chosen as the probe DNA and immobilized on the cantilever. Then a sample of 1 μM all-matched DNAs (5′-CCATTCACCTACCGACCTAT-3′) and a sample of 1 μM all-mismatch DNAs (5′-GGTAAGTGGCGAGTTGGATA-3′) were injected individually, as target DNAs in the experiments. For both experiments, the divider output frequency is recorded over time as shown in Fig. 30, where the black curve (solid square) and the red curve (solid circle) represent mismatch and match DNAs, respectively. Before the DNA samples are injected, the PBS buffer is applied to the sensor SoC as a “no-DNA” control to generate the frequency in the initial state (F_0). In “Wash State”, the sensor is rinsed with PBS buffer for several times, which results in relatively unstable frequency changes due to background fluctuations. Particularly, the frequency increases from the “Initial State” to the “Wash State” for the match DNA and decreases for the mismatch DNA. The phenomenon is probably caused by the following factors. First, the experiment on the cantilever-based DNA sensors is performed in an open environment where the ion concentration of the buffer solution may increase due to evaporation The unstable ion concentration could affect the characteristics of the probe DNAs on the sensors and the target DNA bio-molecules. Second, particles such as DNAs and ions can move in the liquid environment due to electric fields and bump into the cantilever sensor, which also causes measurement uncertainty. Third, the piezoresistor is embedded on the bottom of the cantilever and exposed to the buffer solution. The resistance of the piezoresistor might be affected by the unstable ion concentration of the buffer solution. All the factors mentioned above are attributed to the liquid environment. Therefore, the sensors need to be dried after the "Wash State" so that stable measurement results can be obtained in the "Dry and Steady State".

Generally, the temporally unstable characteristic in the "Wash State" would not be an issue if all the unbound particles are totally removed during the washing procedure and the sensors are perfectly dried in the Steady State. The difference between the divider output frequency in the "Initial State" and that frequency in the stable "Dry and Steady State" can be regarded as the change due to the DNA hybridization. As calculated from Fig. 30, the frequency change is about 112 kHz for match DNAs and is about 50 kHz for mismatch DNAs. Apparently, the match DNAs cause the larger frequency change than the mismatch DNAs, which unequivocally proves the function of the DNA SoC. It should be noted that the smaller frequency change of 50 kHz is resulted from non-specific binding of mismatch DNAs which can be considered as an interfering noise source. Practically, non-specific binding phenomenon seldom occurs in the experiments on match DNAs. Therefore, as long as the frequency change caused by hybridization of match DNAs is larger than that by non-specific binding of mismatch DNAs for the same concentration, the noise can be ignored or eliminated by post signal processing. To avoid degradation in the sensitivity, the undesired effect can also be alleviated by some previously proposed techniques [19].

In the other experiment, two match DNA samples of different concentrations (100 pM and 1 μM) are injected as target DNAs. As shown by the experimental results in Fig. 31, frequency changes (ΔF) of 40 and 16 kHz are induced by the match DNA samples with concentrations of 1 μM and 100 pM, respectively. Namely, the DNA SoC could distinguish concentrations of different DNA samples from 100 pM to 1 μM. It should be noted that the time left for the hybridization process is much shorter than (about one third of) that in the previous experiment. Only 5 min is left for the hybridization process in this experiment to avoid the unwanted effect due to saturation of DNA binding. According to [20, 21], the extent of the hybridization strongly depends on the time for hybridization before saturation of DNA binding occurs. Consequently, the time for hybridization would determine the sensitivity of the DNA sensor, which explains why the frequency change (40 kHz) shown in Fig. 31 is nearly one third of the frequency change (112 kHz) shown in Fig. 30 for the same concentration of 1 μM match DNA.

Furthermore, to evaluate the precision of the system, the short term stability (noise floor) is analyzed according to the steady-state data obtained from this experiment. As depicted in Fig. 32, the short-term Allen deviation σ_y in this experiment reduces to 6.58×10^{-5} at an average time of 160 s, so the frequency deviation in the presence of noise would be equal $F_0 \times \sigma_y = 395$ Hz [22]. This frequency deviation can be used to estimate the minimum recognizable frequency change about 1.2 kHz (0.02 % frequency variation), as the limit of detection (SNR = 3) for the DNA SoC.

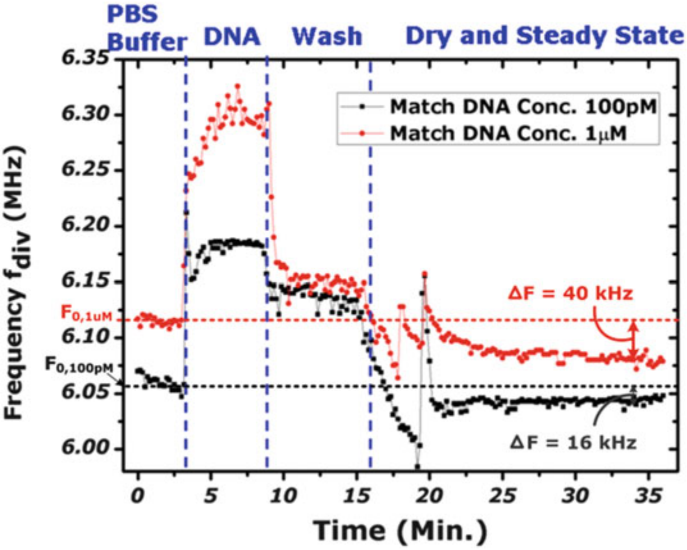


Fig. 31 The divider output frequency is measured over time for experiments on match DNAs with different concentrations [1]

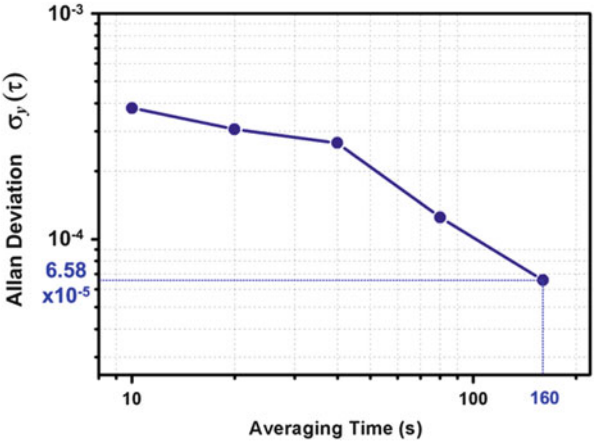


Fig. 32 Short-term Allan deviation according to the experimental data [1]

To investigate the influence of the cantilever structure on sensor characteristics, several sensors with different structures (identified as C1 ∼ 10) were designed and implemented. The experiment results are shown in Fig. 33. For these cantilever-based DNA sensors, the illustration of sensor geometries and a summary table with the detailed information are shown in Fig. 34. Theoretically, a cantilever sensor with a smaller spring constant (larger L/W) is supposed to present a higher sensitivity [23]. However, after the cantilevers are developed and released from

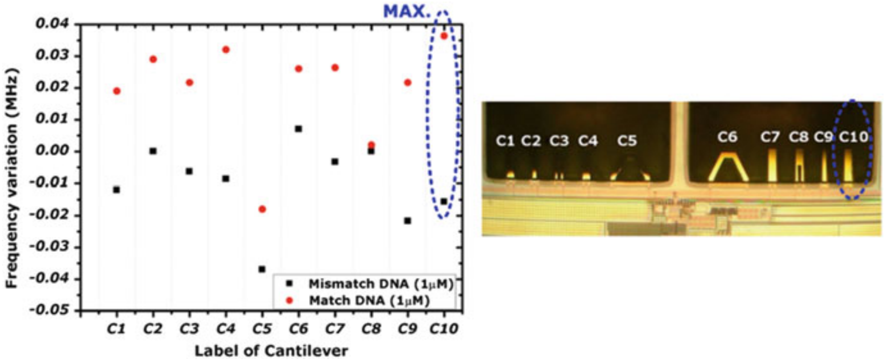


Fig. 33 Frequency variation for different cantilever sensors in this chip [1]

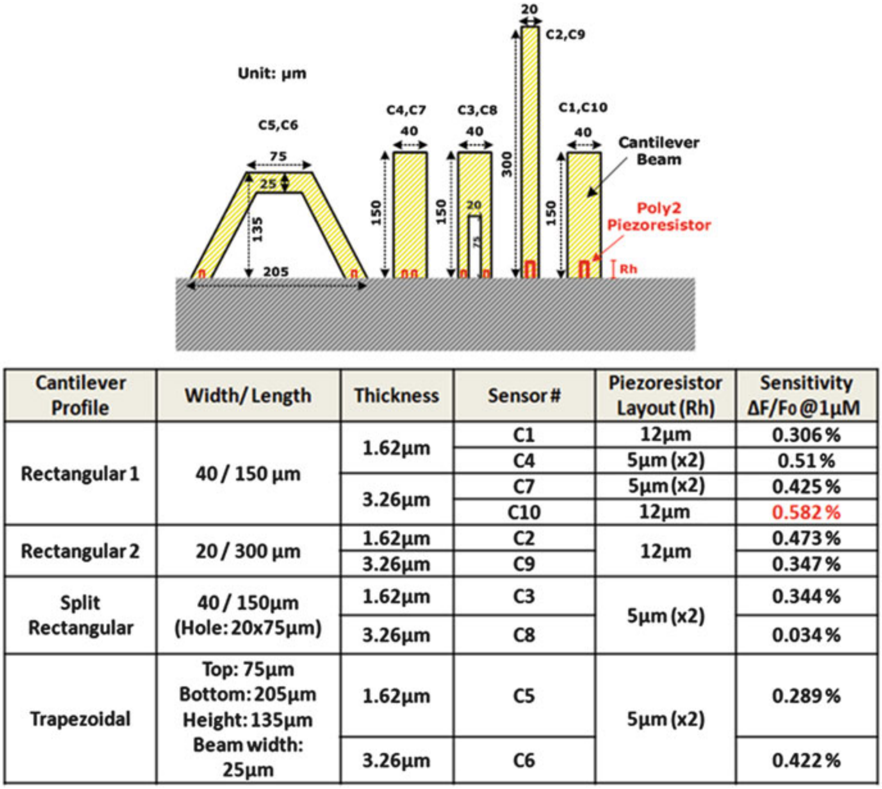
the silicon substrate in the post process, they present different extents of initial bending due to individual residual stresses. This issue has a great impact on the characteristics of the sensors, such as the initial resistance (F_0 frequency) and the sensitivity.

It is found that after the post processing, the sensor with the cantilever structure C10 is usually flatter than sensors with other structures, which means the residual stress in this cantilever C10 is lower than others. This explains why the sensor with the cantilever structure C10 can achieve the best sensitivity among all the sensors, as can be seen in Fig. 34, even though it does not present the smallest spring constant. Notably, the output frequency increases for most of the sensors as the result of DNA hybridization and decreases for a few others, as shown in Fig. 33. It may be due to different residual stresses and/or different sequences of target DNAs [24, 25].

The Fig. 35a shows the measured digital output signal of the DNA SoC for the experiments on match and mismatch DNAs. The digital output results agree with the corresponding frequency changes in Fig. 30. The digital output would be sent to the OOK transmitter which delivers an output power of −6.4 dBm at 402 MHz (in the MICS band), as shown in Fig. 35b. The waveforms of the OOK-modulated signal delivered by the transmitter and the digital output signal recovered from the OOK-modulated signal were measured by an oscilloscope, as shown in Fig. 35c.

Hepatitis B Virus (HBV) DNA Detection

Hepatitis B is the most common serious liver infection which affects around 350 million people worldwide. In order to prove the practical use of the DNA SoC, the detection of hepatitis B virus (HBV) DNA was demonstrated. In this experiment, the HBV (5′-SH-CCGATCCATACTGCGGAAC-3′), as the probe DNA, and several kinds of DNA oligonucleotides, as target DNAs, were used to evaluate the selectivity of the system. All the DNA samples were purchased from Genomics, Taiwan. The sequences of the target DNAs are all-match (5′-GTTCCGCAGTATGGATCGG-3′),



Cantilever Profile	Width/ Length	Thickness	Sensor#	Piezoresistor Layout (Rh)	Sensitivity ΔF/Fo@1μM
Rectangular 1	40 / 150 μm	1.62μm	C1	12μm	0.306 %
			C4	5μm (x2)	0.51 %
		3.26μm	C7	5μm (x2)	0.425 %
			C10	12μm	0.582 %
Rectangular 2	20 / 300 μm	1.62μm	C2	12μm	0.473 %
		3.26μm	C9		0.347 %
Split Rectangular	40 / 150μm (Hole: 20x75μm)	1.62μm	C3	5μm (x2)	0.344 %
		3.26μm	C8		0.034 %
Trapezoidal	Top: 75μm Bottom: 205μm Height: 135μm Beam width: 25μm	1.62μm	C5	5μm (x2)	0.289 %
		3.26μm	C6		0.422 %

Fig. 34 Detailed information of the designed cantilever sensors in this chip [1]

one base pair (1-bp) mismatch (5′-GTTCCGTAGTATGGATCGG-3′), three base pair (3-bp) mismatch (5′-GTTCCGTGATATGGATCGG-3′), and all-mismatch (5′-ACCTTATCTACCTACCTAT-3′), respectively. Figure 36 shows the measured divider output frequency before and after DNA hybridization for different samples including the PBS buffer and the target DNAs with four kinds of sequences as previously mentioned. As expected, the hybridization process involving the all-match DNA sample causes the largest frequency change. According to the experiment results, the DNA SoC is capable to distinguish between the match HBV DNAs and the mismatch DNAs even for those with one base pair mismatch sequence.

The frequency changes are normalized by the initial frequency to obtain the relative frequency change $(F_{steady} - F_0)/F_0$ for the experiments on match HBV DNAs with different concentrations (1 pM, 100 pM, and 10 nM), as shown in Fig. 37. As mentioned previously, the PBS buffer is treated as the "no-DNA" control and applied to the system first to obtain the initial frequency F_0 for each experiment.

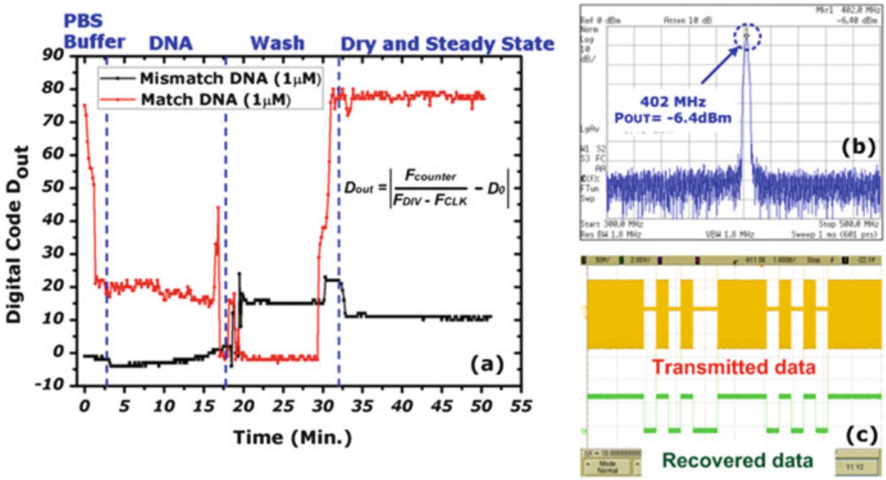
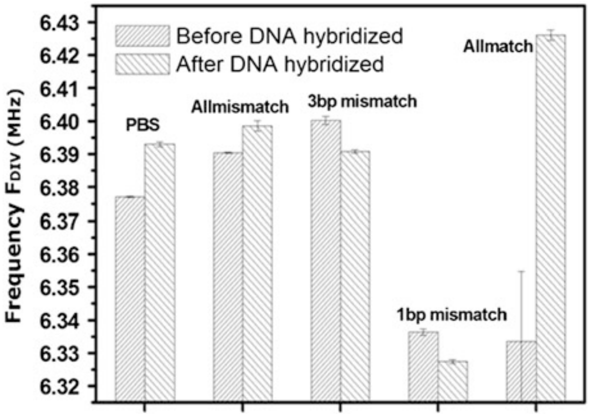


Fig. 35 (**a**) Measured digital data for match and mismatch DNA conditions. (**b**) Spectrum of the transmitter output. (**c**) Waveforms of the transmitted data and the recovered data displayed in the oscilloscope [1]

Fig. 36 Measured divider output frequency for different target DNA sequences before and after DNA hybridization [1]



For sensing of HBV DNA, a concentration range from 1 pM to 10 nM can be provided by the system, as can be seen from Fig. 37 where the relative frequency change is almost linear with the DNA concentration in log scale. For each concentration, measurement results were collected from the experiments on five pieces of sensor samples and used to calculate the standard error of the mean (SEM), as shown by the error bar in Fig. 37. It is noteworthy that the relative frequency change of 0.9 % for the 1 pM HBV DNA sample is much larger than the minimum recognizable frequency change (0.02 %) that is estimated for the limit of detection (SNR = 3). According to the measured results and the previous analysis, the limit of detection of the system would be less than 1 pM in terms of the DNA concentration, indicating that the DNA sensor SoC is suitable for most clinical applications.

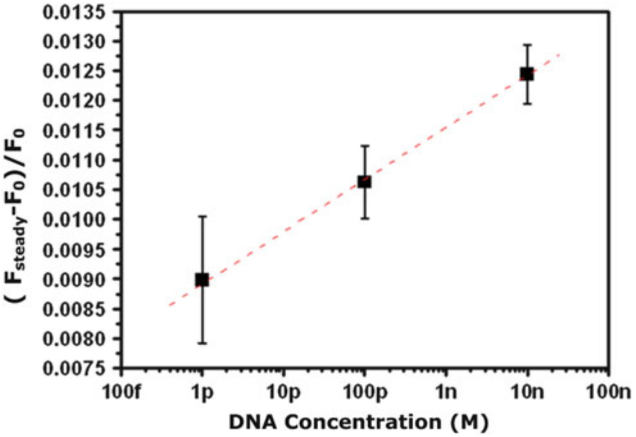


Fig. 37 Relative frequency change for match HBV DNAs with different concentrations (1 pM, 100 pM, and 10 nM) [1]

Practical Issues

The effect of nonspecific binding is a critical problem to practical applications for bio-molecule detection. Actually, it is nearly impossible to create a sensing environment without nonspecific binding in practical applications. As an origin of interfering noise for the detection system, it should be minimized in practical use. Nonspecific binding can be alleviated to enhance the selectivity of the system by several techniques. For example, blocking agents or antifouling agents [19] can be added into the sensor chamber to prevent subsequent nonspecific binding. In addition, more effective washing procedure before the steady-state condition can sometimes improve the selectivity. Apparently, a compromise needs to be made between the selectivity and the complexity of sample preparation.

Moreover, restrictions or notes should be clearly described in user guides for such type of devices in the future. The required information can be obtained as follows. The DNA SoC should be tested under more different conditions including the worst case. It should be tested on DNAs with different sequences, for different lengths of time, with different DNA concentrations or buffer solutions, and within a wider temperature range, etc. Then useful information including the tolerant concentrations of interfering sources (ions or DNAs) for certain degrees of accuracy, the limits of detection for certain analytes, recommended buffer solutions, and the required time for each procedure can be provided for users. For example, when our system is applied to DNA sequence discrimination, for higher accuracy, the concentration of the target DNAs should better be controlled roughly within the range from 1 pM to 10 nM, as found in the experiment.

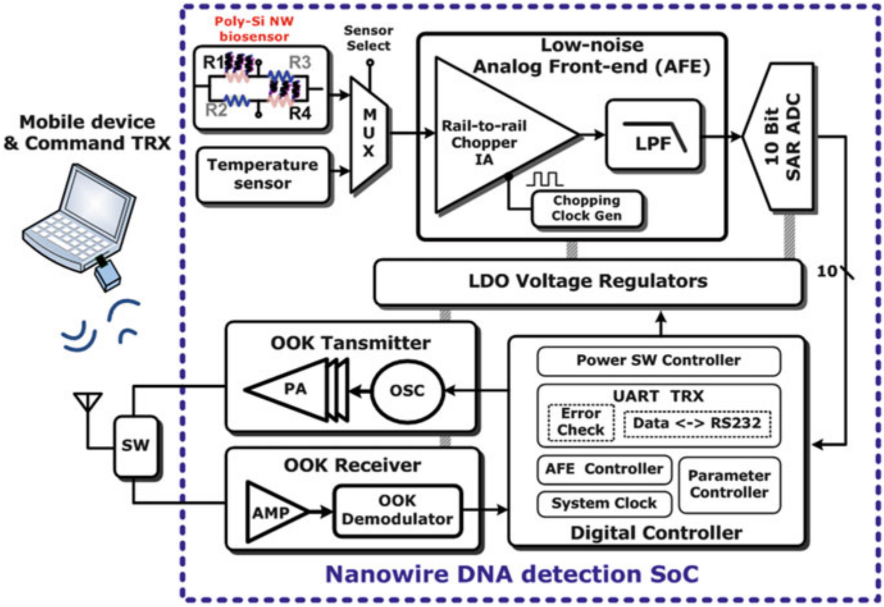


Fig. 38 The system architecture of the poly-Si NW based DNA detection SoC [2]

3.2 *Poly-Silicon-Nanowire-Based Hepatitis B Virus Detection DNA SoC*

There is still room for improving the characteristics of the previously presented DNA SoC. For clinical applications, systems with better sensitivity, higher selectivity and robotic characteristics are always desired. To address this need, DNA biosensors are realized with polysilicon nanowires (poly-Si NWs) by using commercialized CMOS technology. Due to the one-dimensional nano-scale morphology, NW FETs with large ratios of surface to volume can achieve high sensitivity as chemical and bio-molecules sensors [26, 27]. In addition to high sensitivity, the CMOS poly-Si NW-based bio-SoC can achieve label free and real time detection of HBV DNA. With the characteristics of low cost, high practicability and portability, the bio-SoC promisingly allows the access to point-of-care and outdoor applications.

The whole system architecture is shown in Fig. 38. Benefiting from the CMOS SoC technology, the poly-Si NW-based biosensor is integrated with an analog-front-end (AFE), a successive-approximation-register analog-to-digital converter (SAR ADC), and a digital controller to form a sensor SoC that exceeds traditional Si NW discrete measuring systems. In addition, an on–off keying (OOK) wireless transceiver is incorporated to provide the wireless capability for the DNA bio-SoC.

As mentioned in Sect. 1.1, a Wheatstone bridge is formed by poly-Si NW-based sensors in the full bridge arrangement. The change in the resistances of the sensors would result in a corresponding change in the output voltage of the

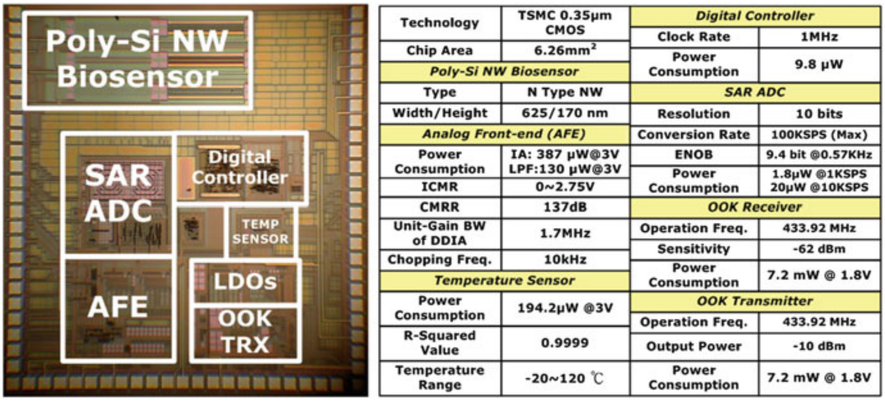
Wheatstone bridge. The differential output signal is properly amplified by the AFE with characteristics of low noise, high CMRR, and rail-to-rail input range. The AFE consists of a differential difference amplifier (DDA) and a low-pass filter (LPF) as shown in Fig. 38. It is worth mentioning that the in-band noise of the AFE is dominated by flicker noise and dc offset. Therefore, the chopper-stabilization technique is adopted to reduce low-frequency noise and dc offset. Moreover, the DDA should exhibit large input impedance to reduce its loading effect on the resistive biosensor.

The amplified signal is then digitized by the 10-bit SAR ADC at very low power consumption and then sent to the digital controller. The digital controller functions as a built-in micro-processer that provides the required control for each building block and translates the digitized signals into data with RS232 format so that the sensing results can be wirelessly delivered to external devices by the transceiver through OOK modulation. The data can be recorded and further analyzed in the external devices. In addition to the function of data transmission, the OOK transceiver also performs receiving and demodulation on commands from external devices.

To overcome the temperature drift of the resistive poly-Si NW biosensor, a temperature sensor that is proportional to absolute temperature (PTAT) with excellent linearity ($R^2 = 0.9999$ from −20 to 120 °C) is incorporated into the bio-SoC for temperature calibration. The measurement results from the temperature sensor and the bio-sensor would be used to build up a mapping table. Such a mapping table can be stored in the external devices and used to perform the calibration.

3.2.1 Experiment Results

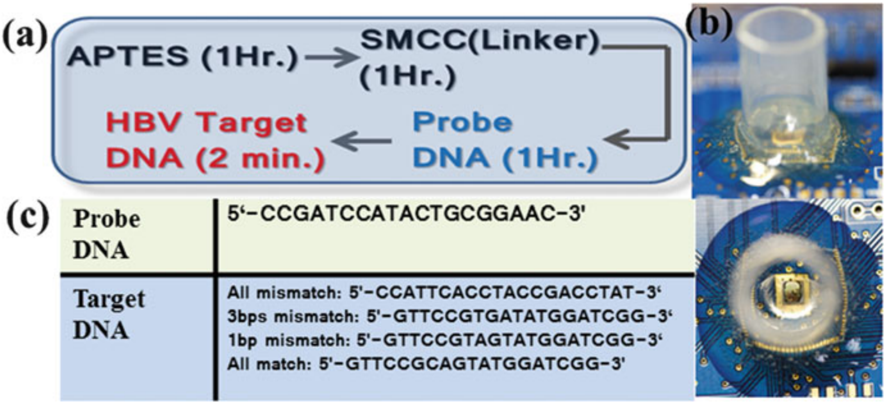
The poly-Si NW based bio-SoC is fabricated in TSMC 0.35-μm 2P4M CMOS process. Figure 39 shows the micrograph of the bio-SoC, along with a summary table of the system performance. Since the procedures to functionalize the sensor and the following experiment are conducted in aqueous environments, it is necessary to protect the on chip CMOS circuits. Moreover, it is important to prevent the bonding wires and pads from short-circuit condition due to ionic buffers. Therefore, after the bio-SoC is mounted on a printed-circuit-board (PCB) by wire-bonding, the epoxy Ab glue is used to cover the area of pads and bond wires to create the passivation. To keep liquid samples on the top of bio-SoC, a plastic tube is stood on the PCB board as a fluid channel that encloses the bio-SoC with bonding wires. During the experiment, all the bio-related protocols are carried out inside the fluid channel. To realize the DNA sensor, the probe DNA is first immobilized on the poly-Si NW. After the functionalizing procedure, the probe DNA, as a functional layer on the sensor, can capture the target DNAs with specific sequence. The steps of surface immobilization and hybridization are illustrated in Fig. 40a, while the photo of the bio-SoC and the fluid channel on the PCB, with the bounding pads and wires covered by AB glue, is shown in Fig. 40b. The HBV DNA sequences are shown in Fig. 40c.


Poly-Si NW Biosensor
SAR ADC
Digital Controller
TEMP SENSOR
AFE
LDOs
OOK TRX

Technology	TSMC 0.35μm CMOS
Chip Area	6.26mm^2
Poly-Si NW Biosensor	
Type	N Type NW
Width/Height	625/170 nm
Analog Front-end (AFE)	
Power Consumption	IA: 387 μW@3V LPF:130 μW@3V
ICMR	0~2.75V
CMRR	137dB
Unit-Gain BW of DDIA	1.7MHz
Chopping Freq.	10kHz
Temperature Sensor	
Power Consumption	194.2μW @3V
R-Squared Value	0.9999
Temperature Range	-20~120 ℃

Digital Controller	
Clock Rate	1MHz
Power Consumption	9.8 μW
SAR ADC	
Resolution	10 bits
Conversion Rate	100KSPS (Max)
ENOB	9.4 bit @0.57KHz
Power Consumption	1.8μW @1KSPS 20μW @10KSPS
OOK Receiver	
Operation Freq.	433.92 MHz
Sensitivity	-62 dBm
Power Consumption	7.2 mW @ 1.8V
OOK Transmitter	
Operation Freq.	433.92 MHz
Output Power	-10 dBm
Power Consumption	7.2 mW @ 1.8V

Fig. 39 The chip photo and performance summary table of the bio-SoC [3]



Probe DNA	5‘-CCGATCCATACTGCGGAAC-3'
Target DNA	All mismatch: 5'-CCATTCACCTACCGACCTAT-3‘ 3bps mismatch: 5'-GTTCCGTGATATGGATCGG-3‘ 1bp mismatch: 5'-GTTCCGTAGTATGGATCGG-3‘ All match: 5'-GTTCCGCAGTATGGATCGG-3'

Fig. 40 (**a**) DNA immobilization and hybridization flow chart, (**b**) bio-SoC on the PCB with bonding pads and wires passivated by epoxy Ab glue and the fluid channel stood on the board, (**c**) HBV DNA sequences [3]

Figure 41a shows the average voltage at the output of the AFE (amplification factor: 40×) for different concentrations of all-match target DNAs and the inset shows the relative voltage change $\Delta V/V_0$. The trend in the output voltage variation with respect to the DNA concentration can be expected from the sensing mechanism. Because the net charge of the DNA molecule is negative, the more target DNAs are hybridized with probe DNAs, the lower the AFE output voltage becomes. To prove the result is solely contributed by the DNA hybridization, hot de-ionized (DI) water at 90 °C is applied to de-hybridize the sample and remove target DNAs. The AFE output voltage then returns to nearly the same level as it is in the initial state when only the PBS buffer is applied. According to the experimental results in Fig. 41a, the detection limit of 10 fM is examined, so the bio-SoC can meet most clinical requirements. Figure 41b shows the average voltage at the output of the AFE for

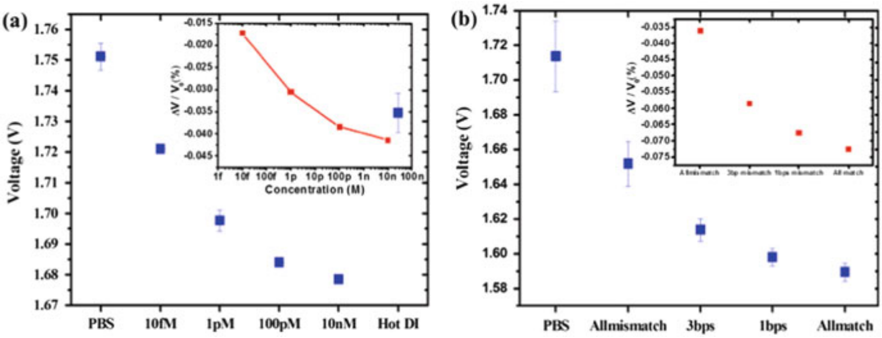


Fig. 41 (**a**) Sensitivity and (**b**) selectivity of the developed poly-Si NW based DNA detection SoC [3]

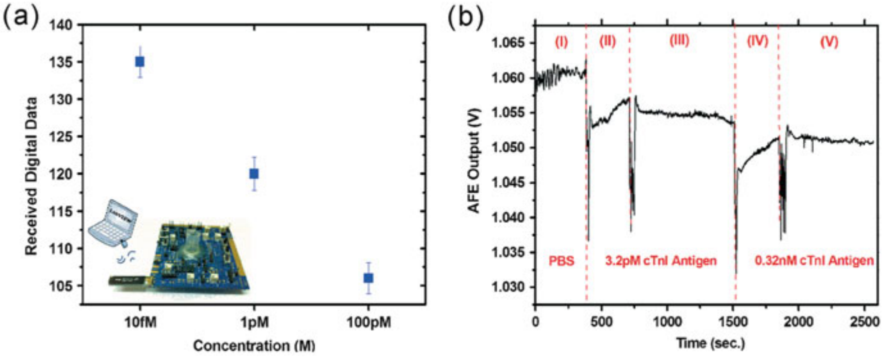


Fig. 42 (**a**) A experimental demonstration of a functional wireless bio-SSoC for HBV DNA detection. The small image shows the wireless setup of this experiment. (**b**) An experimental time history of cTnI detection. Region (I) represents the response of pure PBS buffer. At Region (II), 3.2 pM cTnI sample was injected into the testing reservoir. At Region (III), pure PBS buffer was used to wash away un-bound cTnI antigen. At Region (IV) and Region (V), the same experimental protocol was repeated for measuring the 0.32 nM cTnI sample [2]

different sequences of target DNAs. Obviously, the AFE presents the lowest output voltage for the all-match target DNA due to its highest binding affinity among target DNAs under test.

As aforementioned, the digital controller and the OOK transceiver are incorporated to achieve the wireless function of the bio-SoC. To demonstrate the wireless data link between a personal computer and the bio-SoC, a commercial antenna operating at 433 MHz is soldered on the PCB and connected to the bio-SCO, as shown in Fig. 42a. Another 433 MHz transceiver module with an antenna is used to receive OOK-modulated RF signals from the bio-SoC and to recover the signals back to digital data with RS233 format that can be fed to the personal computer. Through wireless transmission, the digital data received at the personal computer for different concentrations of match target DNAs is shown in Fig. 42a. In addition, the experiment on cardiac troponin I protein (cTnI) is performed to demonstrate the biomolecular diagnosis capability of the bio-SoC, as shown in Fig. 42b. The testing protocol of cTnI is similar to that of HBV ssDNA except it takes longer time to prepare the cTnI sample. During the measurement, a stable detection range from 3.2 to 320 pM can be achieved in the PBS environments. The measurement results show that the bio-SoC has a great potential to be employed in various applications.

3.3 Glucose Sensor SoC

As one of the most serious chronic diseases, it is predicted that the diabetes could affect the daily lives of 3 hundred millions patients in 2025. For patients with diabetes, the glucose concentration in blood needs to be strictly regulated, which requires routine blood glucose measurements. Nowadays, blood glucose tests are usually performed by electrochemical methods, such as electro-enzymatic methods or electro-catalytic methods [28]. The procedures of these regular blood tests are invasive and the test has to be performed four to six times each day, which is really troublesome and also quite painful to these patients. If the blood glucose concentration is over 200 mg/dl, which is quite normal for these patients, the wounds caused by these blood tests can be difficult to heal and may even get infected.

A reusable hydrogel-based glucose sensor SoC can provide a better alternative. Through a subcutaneous implant surgery, once and for all, the sensor SoC can be put under the skin of the patient and wirelessly delivers the testing results of the blood glucose. In this way, blood glucose monitoring becomes more convenient and friendly to patients. In fact, the function of continuous data tracking is also helpful for other patients who strongly rely on personalized medicines. The block diagram of a glucose sensor SoC is shown in Fig. 43. As mentioned in Sect. 1.1, a hydrogel-

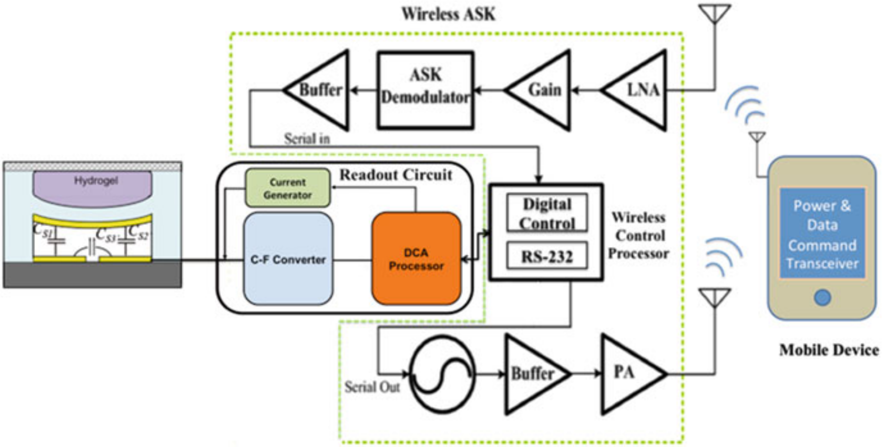


Fig. 43 Block diagram of the hydrogel-based glucose monitoring SoC [4]

based glucose sensor can convert a change in the glucose concentration into a capacitance change of the sensor. The glucose-induced capacitance changes can be translated into a digital output signal by the capacitive readout circuit introduced in Sect. 1.2.

Basically, the glucose sensor SoC operates the same as the other SoCs that are introduced in previous sections. A digital control unit is employed to translate the digital output signal into data with RS232 format and an ASK transceiver is incorporated to provide the wireless function. The ASK transceiver consists radio frequency circuits including a low noise amplifier (LNA), a multi-stage amplifier (Gain stage), a demodulator, a digital buffer, a ring-oscillator, a buffer stage and a power amplifier (PA) [29, 30]. Note that the power consumption for medical implants is an important issue and the ASK transceiver consumes the most power in this glucose sensor SoC due to its operating frequency. To reduce the average power consumption, the ASK transceiver will be waked up only when an external activation command is received.

3.3.1 Experiment Results

The hydrogel-based glucose sensor SoC is fabricated using TSMC 0.35-µm CMOS technology. The die area is 3 × 3 mm^2. The read out circuit dissipates the power of 285 nW from the 3-V supply and it provides the detectable capacitance range from 1.1 to 2.68 pF at such a low power consumption. As previously mentioned, the wireless circuitry consumes most of the energy and its power dissipation of is 11.9 mW. To integrate the hydrogel-based glucose sensor into the system, the sensor is fabricated on the top of the chip by using CMOS compatible micromachining techniques. The glucose sensor SoC is illustrated in Fig. 44. Notably, a spiral inductor with an outer dimension of 3 × 3 mm^2 is designed to capture the data and RF power through coil coupling [31] and placed under the active area of chip to reduce the fabrication cost. A rechargeable battery would be required for longer lifetime in the future [32]. As a medical implant, the SoC should be bio-compatible, so the package is coated with parylene material except for the AAO membrane.

The chip micrograph and the photo of the SoC with integrated glucose sensor after packaging are shown in Fig. 45a, b, respectively. Figure 46 shows in vitro measurement results. Before the glucose solution is applied, the glucose sensor would exhibit a small capacitance, which leads to a small binary value of the digital output result. Once the glucose solution of 200 mM is applied to the capacitive sensor, an obvious increase can be observed from the binary value of digital output result in a very short time. The increase in the binary value would become stable within 2 min. Experimental results show that the system achieves a limit of detection of 40 mM in terms of the glucose concentration. A summary of performance is given in Table 2.

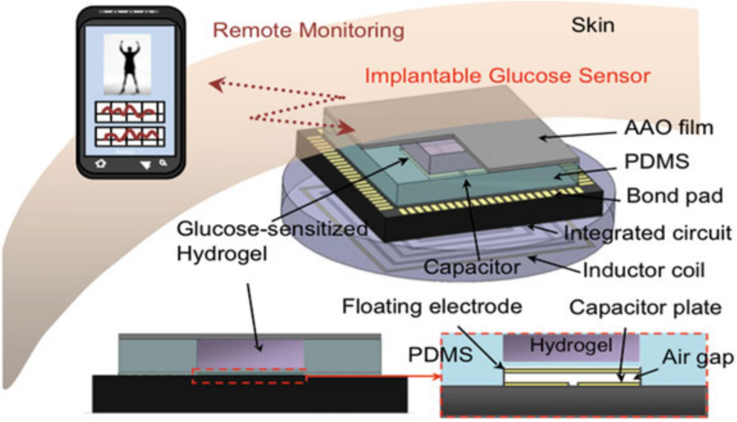


Fig. 44 Illustration of the hydrogel-based glucose monitoring SoC and its package [4]

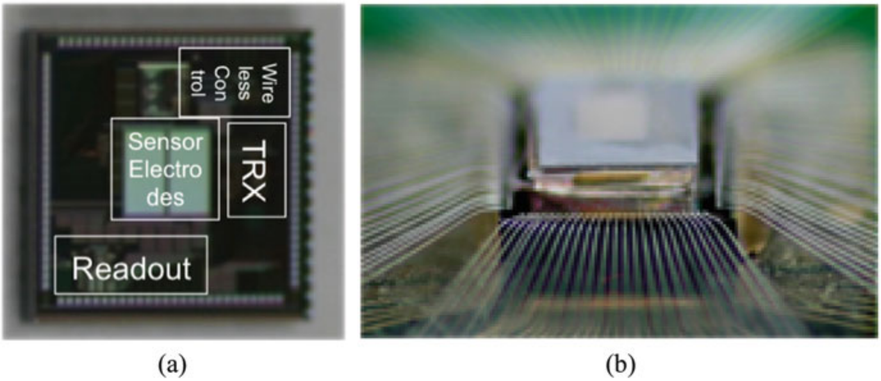


Fig. 45 Photo of (**a**) Chip (**b**) SoC package [4]

3.4 Reconfigurable Multi-Sensor SoC

In practical applications, several biomedical signals should be monitored simultaneously, which requires a multi-sensor SoC that can handle various types of biomedical signals by different sensing mechanisms. The biggest challenge in realizing a multi-sensor SoC may be the design of the post-IC processes. In the stage of deploying different sensors on a single chip, the chip needs to go through several post-IC processes. All the post processes must be compatible not only with the standard CMOS process but also with each other.

The architecture of a reconfigurable CMOS multi-sensor SoC is illustrated in Fig. 47. To achieve real-time monitoring on multiple physiological parameters, four types of sensors including the nanowire-based protein sensor, the hydrogel-based glucose sensor, the ISFET pH sensor, and the band-gap temperature sensor are

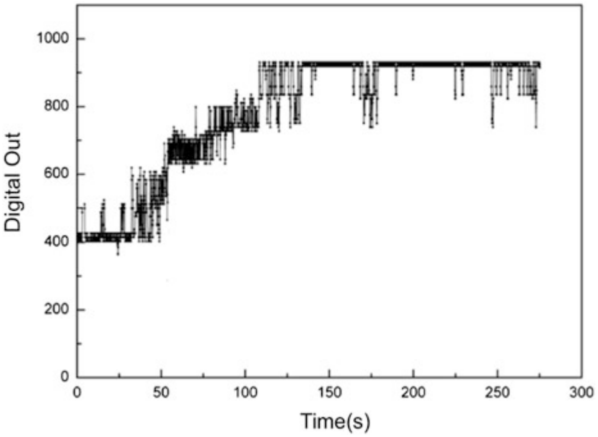


Fig. 46 Preliminary in vitro measurement result of Digital Out vs. time with 200 mM glucose concentration [4]

Table 2 Performance summary

Technology	TSMC 0.35 μm CMOS
Supply voltage	3 V
Chip size	9.01 mm^2
Readout sensing power consumption (CLK @ 10 kHz)	285 nW
TRX power consumption	11.9 mW
Operation frequency	402–405 MHz (MICS band)
Sensing capacitance range	1.1–2.68 pF

integrated into the SoC as these sensors are widely used in biomedical sensing applications. Moreover, two energy harvesting mechanisms are employed so that the SoC can pick up the solar energy through a 2×2 mm^2 GaAs solar cell with a condenser lens while grabbing the RF energy through electromagnetic coupling, which solves the issue of battery replacement in medical devices for long-term usage or implantable applications.

Figure 48 shows the block diagram of the reconfigurable multi-sensor SoC. Four types of sensing results (capacitive, resistive, current, and voltage types of analog signals) are processed by reconfigurable circuitries to reduce the chip area. The reconfigurable multi-sensor interface provides functions of multiplex and conversion for these analog signals. According to the received command, the interface would be appropriately configured by the digital processor to select one of the four types of input signals and convert the selected signal to a voltage-type of signal. The reconfigurable sensor interface is followed by a programmable gain amplifier (PGA) which provides amplification for the voltage signals. Both the interface and the PGA are based on switched-capacitor circuits, as mention in Sect. 1.2. The output signal of the PGA is then converted into digital output signals by a 10-bit

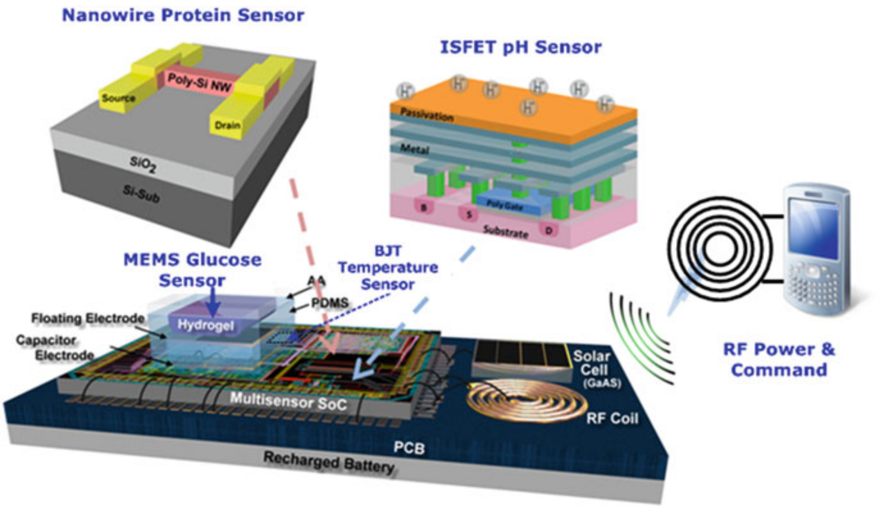


Fig. 47 Overall system architecture and cross-section view of sensors [5]

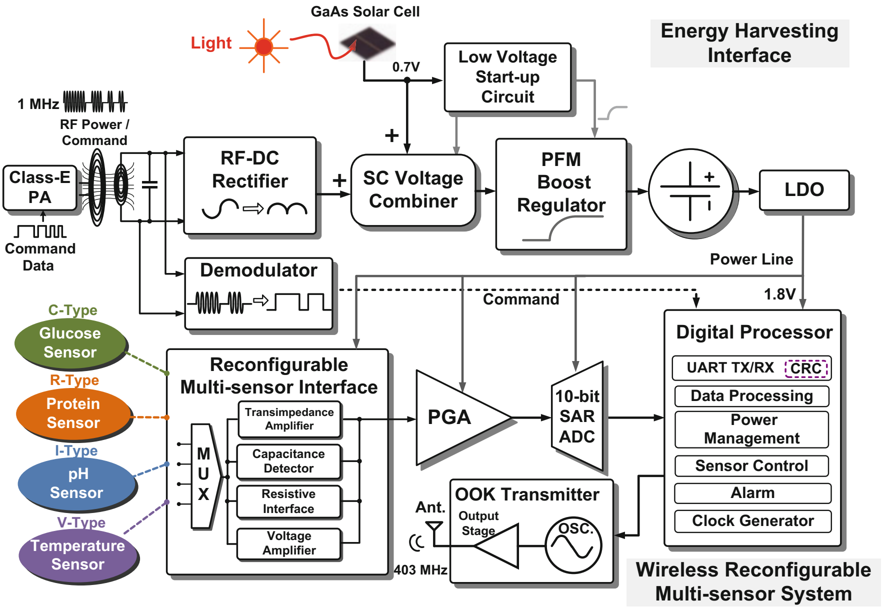


Fig. 48 Block diagram of the reconfigurable multi-sensor SoC [5]

successive approximation register analog-to-digital converter (SAR ADC). Through digital signal processing (DSP), the digital processor performs digital filtering on the digital output signals to reduce most surrounding interferences like power-line noise. Finally, the digital output signal will be translated into data in RS232 format and then wirelessly transmitted to an external monitor by the 403-MHz on–off keying (OOK) transmitter.

As previously mentioned, the system relies on two kinds of energy resources, the solar energy from a GaAs solar cell and the RF power through coupling at 1 MHz. The two energy harvesting methods have been successfully demonstrated in previous researches [33, 34] and proved to be applicable even inside the human body. An energy harvesting interface is design to collect the available energy as follows. The external RF power is coupled via an inductor-capacitor (*LC*) tuned network and the RF input voltage obtained from the coil is sent to a rectifier for RF-to-DC conversion. Then, a dual-input switch capacitor voltage combiner is employed to combine the energy from both sources.

To meet the supply voltage requirement of the system (1.8 V) or rechargeable batteries, the output voltage of the combiner is further raised by a pulse frequency modulation (PFM) boost converter. However, the voltage combiner and the PFM boost converter cannot be operated at the low voltage (=0.7 V) offered by the solar cell, especially when these circuits are implemented using 0.35-μm CMOS technology, where low-threshold-voltage devices are unavailable. For energy transducers with low output voltages, start-up circuits that operate at low supply voltages (<0.5 V) are usually necessary. Powered by the solar cell only, a low-voltage start-up circuit generates a boosted voltage (>1.8 V) to kick-start the energy harvesting interface.

To achieve the reconfigurable multi-sensor system, system parameters such as the type of sensors and associated signals, the amplification gain of the PGA, the calibration bit, the transmission mode, the operation mode, and the alarm threshold can be adjusted by users with the self-defined data/command. Data in RS232 format is carried by the 1-MHz signal through OOK modulation and wirelessly delivered to the system by coil coupling. A simple demodulator then recovers the data from the received signal. The data is further processed in the digital processor which controls and operates the system according to the data contents. The UART in the digital processor is in charge of the data conversion between serial format and parallel format. A cyclic redundancy check (CRC) is used to detect data errors in the digital processor. After being digitized in the ADC, the sensing results in digital form would be further processed by the digital impulse response infinite (IIR) low-pass filter presenting the corner frequency of 10/40 Hz to eliminate the surroundings interference such as the power-line noise at 50/60 Hz, which is critical for detection functions of medical alert systems. The system can work in different operation modes according to the received commands, as illustrated in Fig. 49. In the single detection mode, one of the four sensors would be selected according to the received command and the signals from the selected sensor would be continuously monitored at higher sampling-rate. In the cyclic detection mode, the reconfigurable sensor readout circuit would process the four types of sensor signals sequentially

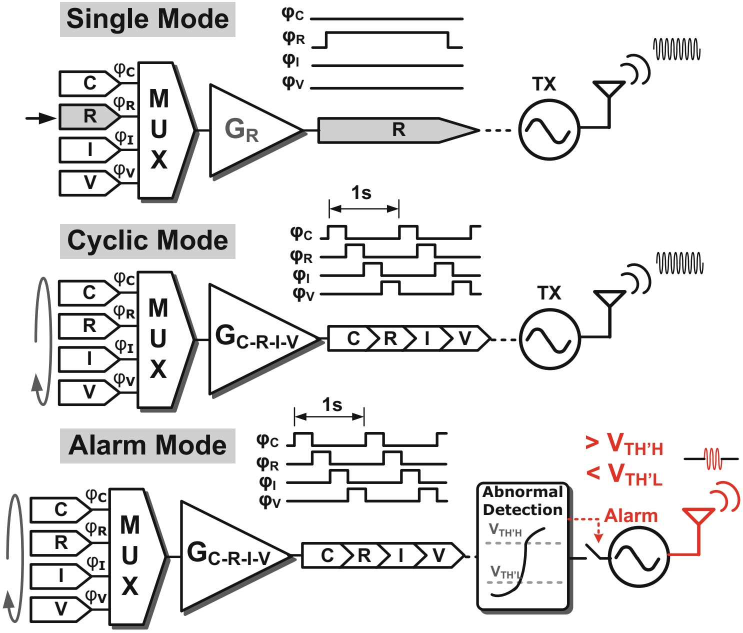


Fig. 49 Illustrations of the three operation modes [5]

and periodically to deliver the corresponding output voltages by time-division multiplexing (TDM). Moreover, the system can operate in the alarm mode to save the power consumption, which is suitable for low power applications. In the alarm mode, wireless transmission would not be turned on until an abnormal condition of the human body is detected, which dramatically reduces the average power consumption since the wireless circuitry is usually the most power-hungry building block in a medical implant. (The wireless transmitter in this system consumes up to 81 % of the total power.) Abnormal conditions can be defined by those in which the readout circuit delivers output voltages that either exceed or fall below specified thresholds ($>V_{TH'H}$ or $<V_{TH'L}$). For example, the blood glucose levels of diabetes patients might get abnormal under many conditions in which the alarms would be activated to turn on the wireless circuitry for a while till the blood glucose levels return to the normal range. The thresholds can be remotely adjusted by users or physicians to achieve customized detection.

3.4.1 Energy Harvesting Interface

The schematic of the dual-input energy harvesting interface circuit is depicted in Fig. 50. Instead of far-field antenna radiation at UHF bands [35], near-filed inductive coupling at 1 MHz is adopted for RF powering due to its characteristics such as the

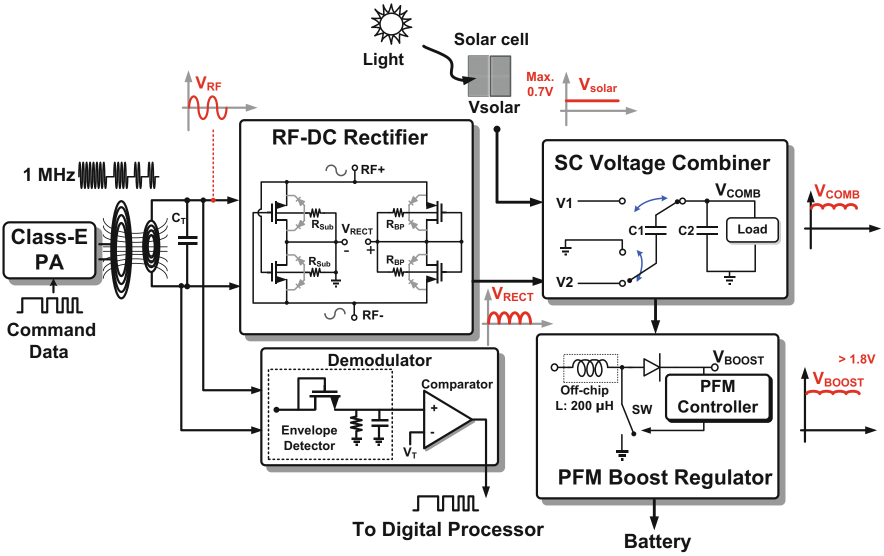


Fig. 50 Schematic of the dual-input energy harvesting interface [5]

potentially larger input voltage for the rectifier, weaker interference from nearby, and the lower propagation loss inside the human body. The rectifier is formed with four MOSFET switches to minimize the turn-on voltage [36]. Notably, since a switch conducts current in both directions, an unwanted reverse current would be generated during the rectifying process. Therefore, the rectifier is realized by using a two-way configuration [37], where the two NMOSs function as switches driven by differential input voltages while the two diode-connected PMOSs alternatively perform rectifying for half of each cycle. The substrates of PMOSs and NMOSs are weakly connected to the rectifier output voltage V_{RECT} and the ground via the resistors R_{BP} and the substrate resistors, respectively, to mitigate the reverse recovery currents through the PN junctions in the substrates. To recover the data from the OOK modulated RF signals, an envelope detector comprising a diode-connected NMOS, a parallel *RC* network, and a voltage comparator is employed to perform the demodulation. A shunt capacitor is incorporated with the coil to form a *LC* tuned network to optimize the input impedance so that the input RF amplitude can be maximized. Due to the impedance matching and the small threshold voltage of NMOS, the demodulator can handle the minimum input signal level of −30 dBm at the data rate of 2.4 kbps. To combine the voltages from the two transducers, a simple and high-efficiency (Max. 92 %) switched-capacitor voltage combiner is adopted. Let V1 and V2 represent the voltages from the solar cell and the rectifier, respectively. The operation mechanism of the voltage combiner is based on charge redistribution. Each clock cycle of the switched capacitor circuit is separated into two operation modes in which the switches would be configured as follows. In one mode, V1 is sampled and stored in C1. In the other mode, C1 is connected between the output terminals the rectifier and the combiner. After a number (=N) of cycles, the output voltage of the combiner (V_{COMB}) can be presented as:

$$\begin{aligned} V_{COMB}[N] &= (V1+V2)\cdot\frac{C1}{C1+C2}+V_{COMB}[N-1]\cdot\frac{C2}{C1+C2} \\ &= \frac{(V1+V2)\cdot C1}{C1+C2}\cdot\left[1+\left(\frac{C2}{C1+C2}\right)+\cdots+\left(\frac{C2}{C1+C2}\right)^{N-1}\right] \end{aligned} \tag{2}$$

By adopting identical capacitors C1 and C2, the combiner would eventually deliver an output voltage of V1 + V2 after a sufficiently large number of cycles to achieve voltage combining. It is worth mentioning that a current may flow back from the voltage combiner to the rectifier if the weak input power of the rectifier results in that $V_{COMB} > V2$. To prevent this potential reverse current, the intrinsic diode of the PMOS in the RF-DC rectifier is connected to the output terminal of the rectifier.

In case that the voltage generated from the combiner may not be high enough to charge the battery effectively, the voltage V_{COMB} is further stepped up by a boost switching regulator. Pulse frequency modulation (PFM) can dynamically adjust the switching rate and hence achieve the lower switching loss under light load conditions, as compared with pulse width modulation (PWM). Apparently, the PFM boost regulator is more suitable for low-power applications and therefore adopted in this sensor SoC. The schematic of the PFM boost regulator is depicted in Fig. 51. The PFM switching control circuitry consists of a resistive voltage divider (R_1 and R_2), a hysteresis comparator, and a ring oscillator with the fixed frequency of 1 MHz. By dynamically turning on the oscillator, the number of pulses can be increased (or decreased) to raise (or reduce) the output voltage (V_{BOOST}). Notably, the frequency of the oscillator is effectively set to zero when the oscillator is turned off. Through the hysteresis comparator, the output of the voltage divider would be compared with a reference voltage ($V_{REF} = 0.9$ V) to determine whether the oscillator is turned on

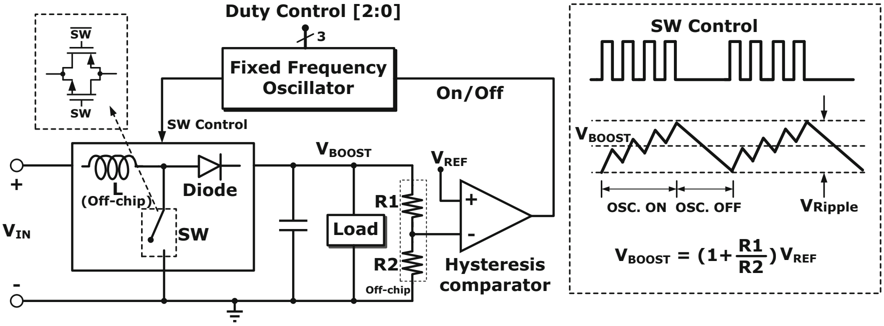


Fig. 51 Schematic of the PFM boost regulator [5]

to raise the output voltage or not. Based on the voltage regulation mechanism, the output voltage would eventually settle down around the level defined by

$$V_{BOOST} = \left(1 + {}^{R_1}\!/_{R_2}\right) \cdot V_{REF} \tag{3}$$

The resistive voltage divider employs two identical resistors (R1 = R2) of 10 MΩ to obtain the output voltage of 1.8 V and the intrinsic power consumption of the energy harvesting interface is 45 μW at 1.8 V supply voltage. The duty cycle of the switching clock can also be adjusted to optimize the efficiency of the boost regulator in different input voltage ranges.

As previously mentioned, the voltage from the solar cell may not be high enough to activate the energy harvesting interface circuit. Similar problems often occur in self-powered systems with integrated transducers for energy harvesting. Previous studies attempted to adopt batteries, to apply higher initial voltages or RF powering [38] to kick-start the voltage boosters. However, these systems still rely on manual intervention with additional external powering sources and hence fail to be truly self-powered. In order to achieve a self-powered system, a low-voltage (0.5 V) start-up circuit that can be activated solely by the solar cell (<0.7 V) is adopted to provide a stepped-up voltage to kick-start the operation of the energy harvesting interface. The schematic of this start-up circuit is shown in Fig. 52. Due to the weak inversion operation, the low-voltage clock generator and the clock booster can operate at a supply voltage of 0.5 V. The clock booster doubles the amplitude of the clock signal (CLK) from the low-voltage clock generator. Then the clock signal (CLKB) with sufficiently large amplitude is able to drive the switch in a step-up boost converter which would generate a boosted voltage stored in C1 to provide the start-up voltage (V_{START}). As V_{START} exceeds 1.8 V, the power-on detector applies the start-up voltage V_{START} to the primary energy harvesting interface and disables the start-up circuit by turning off the clock generator at the same time. The start-up circuit dissipates an intrinsic power of 2.3 μW from the 0.7-V supply. In particular, the energy interface circuit adopts large-size switches (W/L = 2,000/0.5 μm) to reduce the turn-on resistance (~4 Ω) and power loss. To drive these large-size switches whose intrinsic gate capacitance may be up to 5 pF, strong clock buffers with the setup time of 5.5 ns are adopted.

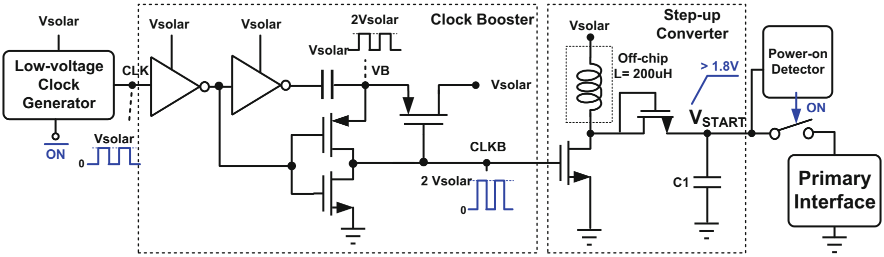


Fig. 52 Schematic of the low-voltage start-up circuit [5]

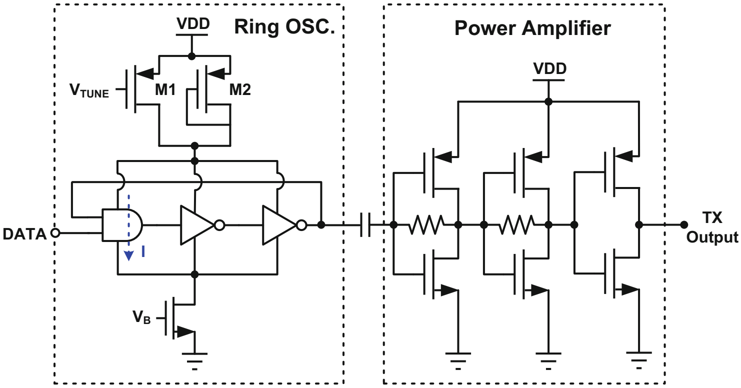


Fig. 53 Schematic of the low-power OOK transmitter [5]

3.4.2 Low-Power OOK Transmitter

The wireless circuitry tends to be more power-hungry than other building blocks due to its high operating frequency. To reduce the average power, OOK modulation scheme is adopted so that the transmitter consumes the least power during the transmission of a "zero". The OOK transmitter in the previous work [1] consumes a large power (11.9 mW) due to the inefficient power amplifier architecture. To achieve the better efficiency, the power amplifier is modified into the configuration of inverter stages in cascade. The OOK transmitter comprises a voltage-controlled ring oscillator and the inverter-type power amplifier (PA) [39], as depicted in Fig. 53. The ring oscillator generates the required carrier signal and is switched on/off by the DATA input to realize the OOK modulation. The carrier frequency (f_{OSC}) can be expressed by

$$fosc = \frac{1}{2 \cdot N \cdot \tau} = \frac{I}{2 \cdot N \cdot V_{OSC} \cdot C_g}, \tag{4}$$

where N is the number of stages, τ is the delay for each stage, V_{OSC} is the oscillation amplitude, C_g is the gate capacitance, and I is the current through each stage. To reduce the power consumption, the oscillator adopts small size transistors to minimize the gate capacitance C_g so that the oscillation at the desired frequency can be achieved with a lower current. Notably, one of the delay cells in the oscillator is realized by a NAND gate so that the oscillator is turned on/off by the data input of 1/0 to perform the OOK modulation. Through adjusting the equivalent resistance of PMOSs M_1 and M_2, the current through the oscillator as well as the oscillation frequency can be controlled by the voltage V_{TUNE}. The overall power consumption of this OOK transmitter is 762 μW.

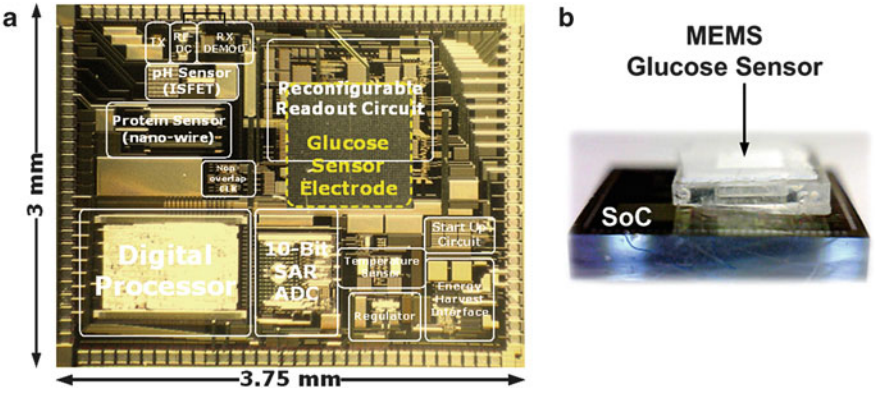


Fig. 54 (**a**) Chip photo and (**b**) side view of the chip after post-IC process [5]

3.4.3 Experiment Results

This self-powered reconfigurable multi-sensor SoC was fabricated in the standard TSMC 0.35-μm CMOS process followed by necessary post-IC processes. The chip micrograph is shown in Fig. 54, along with its side view after post-IC processes. The chip consumes an area of 3 mm × 3.75 mm. After the post-IC processes, the MEMS glucose sensor is stacked on this CMOS chip. For the sake of the post-process compatibility, the nano-wire protein sensor needs to be fabricated before the MEMS glucose sensor is implemented. The bonding wires and pads were carefully insulated by AB glue to avoid electrical short circuit during the experiments when the chip would be immersed into liquid solutions. On the contrary, the sensing areas of the on-chip sensors cannot be covered by the AB glue or anything else since these areas need to be exposed to the solutions. To ease the insulation procedure, all the sensors are placed at least 300 μm away from the bonding pads.

The integration of the analog building blocks including the reconfigurable multi-sensor interface, the PGA and the 10-bit SAR ADC is verified by using instruments or built-in devices to provide the four types of input signals. The sampling frequency of the sensor interface was set to 200 Hz. For each type of signal, the input range during the measurement is wider than the estimated output signal range for each sensor. The output voltages obtained from the readout circuit (left axis) and the digital output results delivered by the ADC (right axis) were simultaneously recorded, as shown in Fig. 55a–d. Within the input range of each signal, the sensor interface exhibits a coefficient of determination (R^2) close to unity, indicating that a linear conversion from each type of signal to a voltage is achieved. In other words, the circuitry can handle the sensor signals faithfully. The conversion for resistance input signals shows the worst linearity than the conversions for other types of input signals, because resistance signals go through the additional conversion of the Wheatstone bridge.

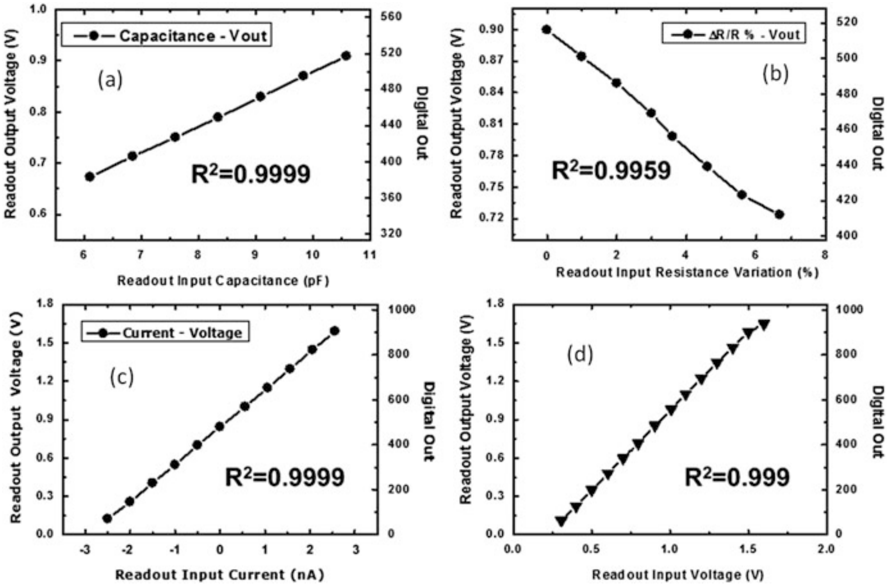


Fig. 55 Measured results: (**a**) capacitance, (**b**) resistance, (**c**) current, and (**d**) voltage detection [5]

Figure 56 show the measurement results obtained from the reconfigurable sensor interface and the ADC during the in vitro experiments with on-chip sensors. During these experiments, the chip was covered by AB glue, except for the exposed sensing area which was surrounded by a fluid channel. Then the phosphate buffered saline (PBS) buffer with pH = 7 was injected into the channel to initialize the test suite. Whenever a new detection process began, both the chip and the channel would be cleaned by the PBS buffer before the sample under test was injected. The data was recorded in all the procedures including the cleaning phase. Glucose sensor tests were performed on five samples with different concentrations, where the data was recorded for 15 min in each case. For the glucose concentration range from 4 to 8 mM, the output voltage rises quasi-linearly with the glucose concentration at the sensitivity of ~69 mV/mM, as shown in Fig. 56a. In the protein sensor test, cardiac protein Troponin-I was chosen as the target proteins to evaluate the feasibility of the system in heart disease detection. Four protein sensor tests were performed under different concentrations for the duration of 20 min.

Technically, the system can successfully identify the troponin-I protein according to Fig. 56b, where the readout delivers the voltage change of 6.8 % in response to the protein concentration of 300 fM. Finally, to test the on-chip band-gap temperature sensor, the chip was placed in a temperature chamber with the temperature varied from −20 to 120 °C in 14 steps. Particularly, the data was recorded after the temperature of the chamber became stable enough for each temperature point which

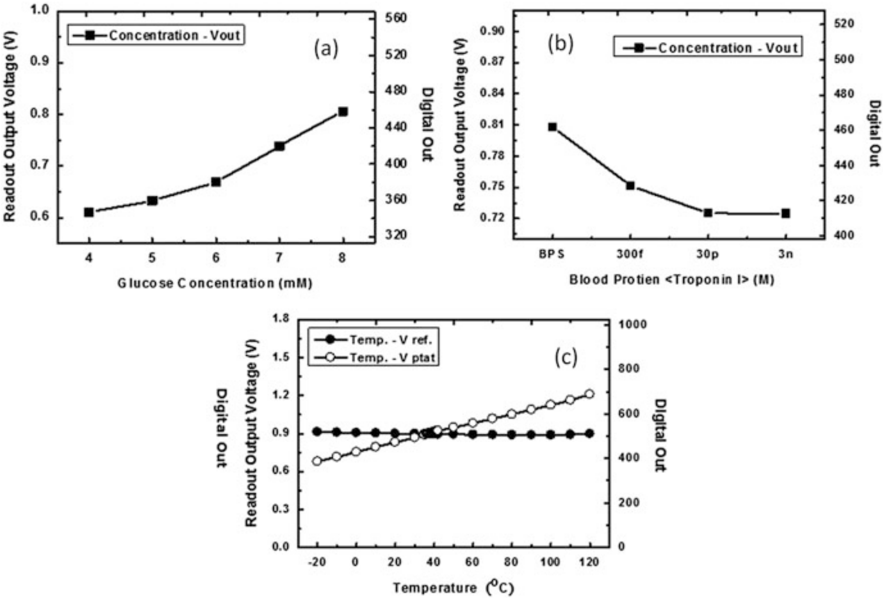


Fig. 56 In vitro experimental results obtained from the sensor interface and the ADC for: (**a**) glucose, (**b**) protein, and (**c**) temperature detection [5]

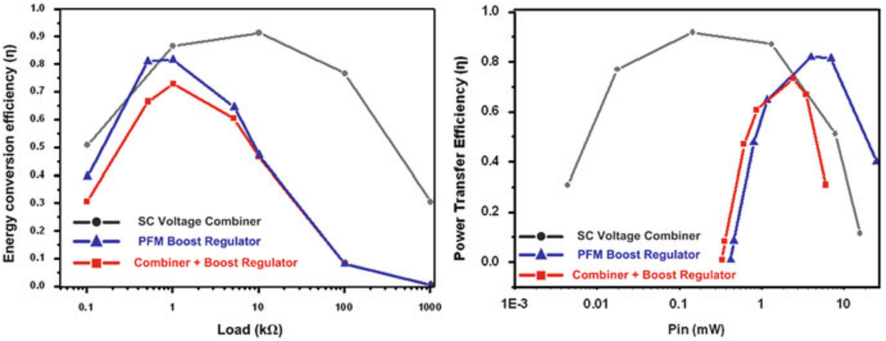


Fig. 57 Measured conversion efficiency of the energy harvesting interface [5]

took 30 min to achieve the accuracy of 1 °C. As shown in Fig. 56c, the output voltage is a linear function of the temperature with the sensitivity of 3.78 mV/°C.

The energy conversion efficiencies (η) of the energy harvesting interface circuit and its individual building blocks were evaluated for various load conditions. It took about 5 min to perform the measurement and data recording for each load condition. As shown in Fig. 57, the SC voltage combiner achieves the maximum efficiency of 92 % at the 10-kΩ load ($P_{in} = 150$ μW), while the PFM boost regulator achieves 82 % at the 1-kΩ load ($P_{in} = 4$ mW). The maximum efficiency of the whole energy

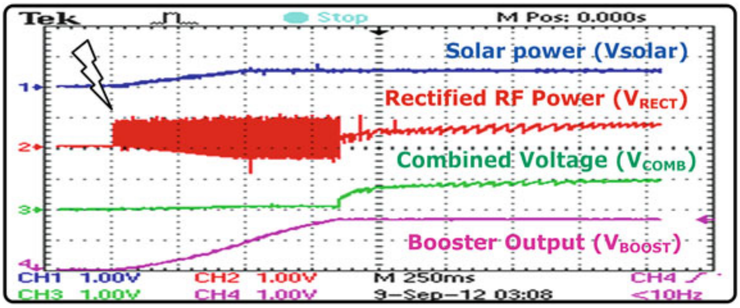


Fig. 58 Measured waveforms of energy harvesting process [5]

harvesting interface reaches 73 % at the 1-kΩ load ($P_{in} = 2.4$ mW) which falls in the vicinity of the equivalent resistance (~3.4 kΩ) of the entire wireless multi-sensor system.

The function of the whole energy harvesting interface is verified by simultaneously observing the output waveforms of the solar cell, the RF-DC rectifier, the voltage combiner, and the boost regulator with the oscilloscope, as shown in Fig. 58. After the RF power (~2.1 mW) and the sunlight (~1.36 mW/mm^2) were applied to this system, output voltages of the solar cell and the RF-DC rectifier gradually came to the dc voltages around 0.5 V. In one second, the voltage combiner successfully combined the two voltages into the dc voltage of 1 V which was upgraded to 1.8 V in the boost regulator in the meanwhile. Technically, the system operation can entirely rely on the solar cell due to the low-voltage start-up circuit and the boost regulator, namely, the system would autonomously power itself once the light shines on it.

The performance summary of the chip is listed in Table 3. The chart at the left-hand side of Fig. 59 illustrates the power breakdown of the system excluding the energy harvesting interface circuitry. The wireless transmitter consumes 762 μW, equal to 81 % of the total power consumption (942.9 μW). Practically, the system would better operate in the alarm mode to reduce the average power consumption, as previously mentioned. The power breakdown of the system in the alarm mode is illustrated at the right-hand side of Fig. 59, showing that the resistive sensor consumes most of the power (80 %).

Table 3 Performance summary

Technology	TSMC 2P4M 0.35 μm CMOS process					
Supply voltage	1.8 V					
On-chip sensors					*10-bit SAR ADC*	
Sensitivity	Glucose	Protein	pH value	Temp.	ENOB	9.4-bit
	69 mV/mM	6.8 % @ 300 fM	600 mV/pH	3.78 mV/°C	Sample rate	100 ksps (Max.)
Power	NA	144 μW	0–2.34 μW	3.78 μW	Power	10.26 μW
Reconfigurable multi-sensor readout					*OOK transmitter*	
PGA Gain	0–40 dB (7-bit resolution)				Carrier frequency	403 MHz
F_S	200 Hz (sample frequency)				Output power	−18 dBm
Power	2.57 μW (interface + PGA)				Power	762 μW
Noise	769 nV/$\sqrt{}$Hz (overall input referred)				*Digital processor*	
Input type	Capacitance	Resistance	Current	Voltage	LPF BW	10/40 Hz
Interface Conv. gain	53.4 mV/pF	91.8 mV/kΩ	293 mV/nA	1.23 V/V	Power	2.34 μW (normal)
						2.6 μW (alarm)
Linearity (R^2)	0.9999	0.9959	0.9999	0.999	*Bias circuit & buffer*	
Energy harvesting interface					Power	15.6 μW
Max. Efficiency (η)	Voltage combiner	PFM booster	Entire		Total power consumption: 942.9 μW	
	92 % @ 10 kΩ	82 % @ 1 kΩ	73 % @ 1 kΩ			

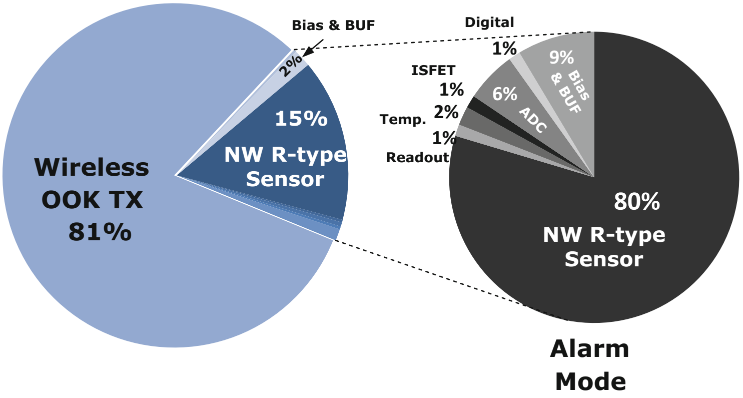


Fig. 59 Power distribution of the entire system [5]

References

1. Huang YJ, Huang CW, Lin TH, Lin CH, Chen LG, Hsiao PY, Wu BR, Hsueh HT, Kuo BJ, Tsai HH, Liao HH, Juang YZ, Wang CK, Lu SS. A CMOS cantilever-based label-free DNA SoC with improved sensitivity for hepatitis B virus detection. IEEE Trans Biomed Circuits Syst. 2013;7(6):820–30.
2. Huang C-W, Huang Y-J, Yen P-W, Tsai H-H, Liao H-H, Juang Y-Z, Lu S-S, Lin C-T. A CMOS wireless biomolecular sensing system-on-chip based on polysilicon nanowire technology. Lab Chip. 2013;13(22):4451–9.
3. Huang C-W, Huang Y-J, Yen P-W, Hsueh H-T, Lin C-Y, Chen M-C, Ho C-H, Yang F-L, Tsai H-H, Liao H-H, Juang Y-Z, Wang C-K, Lin C-T, Lu S-S. A fully integrated hepatitis B virus DNA detection SoC based on monolithic polysilicon nanowire CMOS process. In: 2012 IEEE symp. VLSI circuits dig.; June 2012. p. 124–125.
4. Kuo P-H, Kuo J-C, Yang Y-J, Wang T, Lu S-S. A hydrogel-based implantable wireless CMOS glucose sensor SoC. In: 2012 IEEE international symposium on circuits and systems (ISCAS 2012). Seoul, Korea; 2012.
5. Huang YJ, Tzeng TH, Lin TW, Huang CW, Yen PW, Kuo PH, Lin CT, Lu SS. A self-powered CMOS reconfigurable multi-sensor SoC for biomedical applications. IEEE J Solid State Circuits. 2014;49(4):851–66.
6. Meijer GCM, Wang G, Fruett F. Temperature sensors and voltage references implemented in CMOS technology. IEEE Sens J. 2001;1(3):225–34.
7. Zhang JC, Zhou JW, Mason A. Highly adaptive transducer interface circuit for multiparameter microsystems. IEEE Trans Circuits Syst I. 2007;54(1):167–78.
8. Fan Q, Sebastiano F, Huijsing J, Makinwa K. A 1.8 W 60 nV/Hz capacitively-coupled chopper instrumentation amplifier in 65 nm CMOS for wireless sensor nodes. IEEE J Solid-State Circuits. 2011;46(7):1544–52.
9. Yoo J, Yan L, El-Damak D, Altaf MAB, Shoeb AH, Chandrakasan AP. An 8-channel scalable EEG acquisition SoC with patient-specific seizure classification and recording/processor. IEEE J Solid-State Circuits. 2013;48(1):214–28.
10. Belloni M. A micropower chopper—CDS operational amplifier. IEEE J Solid-State Circuits. 2010;45(12):2521–9.
11. Lee B, Lee K-H, Lee J-O, Sohn M-J, Choi S-H, Wang S-W, Yoon J-B, Cho G-H. An electronic DNA sensor chip using integrated capacitive read-out circuit. In: 32nd annual international conference of the IEEE EMBS. 2010. p. 6547–6550.
12. Stagni C, et al. CMOS DNA sensor array with integrated A/D conversion based on label-free capacitance measurement. IEEE J Solid-State Circuits. 2006;41(12).
13. Lajnef N, Chakrabartty S, Elvin N, Elvin A. Piezo-powered floating gate injector for self-powered fatigue monitoring in biomechanical implants. In: Proc. IEEE int. symp. circuits and systems. 2007. p. 89–92.
14. Grassi M, Malcovati P, Baschirotto A. A 160 dB equivalent dynamic range auto-scaling interface for resistive gas-sensors arrays. IEEE J Solid-State Circuits. 2007;42(3):518–28.
15. Grassi M, Malcovati P, Baschirotto A. A 0.1 % accuracy 100 ohm–20 mohm dynamic range integrated gas-sensor interface circuit with digital output. In: Proc. Eur. solid-state electron circuits conf. 2005. p. 351–354.
16. Begueret J-B, Benbrahim MR, Li Z, Rodes F, Dom J-P. Converters dedicated to long-term monitoring of strain gauge transducers. IEEE J Solid-State Circuits. 1997;32(3):349–56.
17. Grassi M, Malcovati P, Baschirotto A. A 141-dB dynamic range CMOS gas-sensor interface circuit without calibration with 16-bit digital output word. IEEE J Solid-State Circuits. 2007;42(7):1543–54.
18. Behzad A, et al. A fully integrated MIMO multiband direct conversion CMOS transceiver for WLAN applications (802.11n). IEEE J Solid-State Circuits. 2007;42:2795–808.
19. Ruiz-Taylor LA, Martin TL, Zaugg FG, Witte K, Indermuhle P, Nock S, Wagner P. Monolayers of derivatized poly(L-lysine)-grafted poly(ethylene glycol) on metal oxides as a class of biomolecular interfaces. Proc Natl Acad Sci U S A. 2001;98(3):852–7.

20. Wang J, Kawde A-N. Pencil-based renewable biosensor for label-free electrochemical detection of DNA hybridization. Anal Chim Acta. 2001;431:219–24.
21. Wang J, Xu D, Erdem A, Polsky R, Salazar MA. Genomagnetic electrochemical assays of DNA hybridization. Talanta. 2003;56:931–8.
22. Rodriguez-Pardo L, Fariña J, Gabrielli C, Perrot H, Brendel R. Resolution in quartz crystal oscillator circuits for high sensitivity microbalance sensors in damping media. Sens Actuators B. 2004;103:318–24.
23. Senturia SD. Microsystem design. Boston: Kluwer; 2001.
24. Hansen K-M, et al. Cantilever-based optical deflection assay for discrimination of DNA single-nucleotide mismatches. Anal Chem. 2001;73:1567–71.
25. Zheng S, Choi JH, Lee SM, Hwang KS, Kim SK, Kim TS. Analysis of DNA hybridization regarding the conformation of molecular layer with piezoelectric microcantilevers. Lab Chip. 2011;11:63–9.
26. Zheng G, et al. Multiplexed electrical detection of cancer markers with nanowire sensor arrays. Nat Biotechnol. 2005;23:1294–301.
27. Cui Y, Lieber CM. Functional nanoscale electronic devices assembled using silicon nanowire building blocks. Science. 2001;291:851–3.
28. Peura RA. Blood glucose biosensors—a review. Medical Instrument Design. 1991. p. 51–64.
29. Chen CH, Hwang RZ, Huang LS, Lin S, Chen HC, Yang YC, Lin YT, Yu SA, Wang YH, Chou NK, Lu SS. A wireless bio-MEMS sensor for c-reactive protein detection based on nanomechanics. In: ISSCC Dig. Tech. Papers. February 2006. p. 562–563.
30. Chen C-H, Hwang R-Z, Huang L-S, Lin S-M, Chen H-C, Yang Y-C, Lin Y-T, Yu S-A, Lin Y-S, Wang Y-H, Chou N-K, Lu S-S. A wireless bio- MEMS sensor for c-reactive protein detection based on nanomechanics. In: IEEE Trans. biomedical engineering. February 2009, vol. 56, no. 2.
31. Simons RN, Miranda FA, Wilson JD, Simons RE. Wearable wireless telemetry system for implantable bio-MEMS sensors. In: Proceedings of the 28th IEEE EMBS annual international conference. New York City, USA, August 30–September 3; 2006.
32. Chan CK, et al. High-performance lithium battery anodes using silicon nanowires. Nat Nanotechnol. 2008;3(1):31–5.
33. Cong P, Chaimanonart N, Ko WH, Young DJ. A wireless and batteryless 10-bit implantable blood pressure sensing microsystem with adaptive RF powering for real-time laboratory mice monitoring. IEEE J Solid-State Circuits. 2009;44(12):3631–44.
34. Ayazian S, Akhavan VA, Soenen E, Hassibi A. A photovoltaic-driven and energy-autonomous CMOS implantable sensor. IEEE Trans Biomed Circuits Syst. 2012;6:336–43.
35. Masuch J. Co-integration of an RF energy harvester into a 2.4 GHz transceiver. IEEE J Solid-State Circuits. 2013;48(27):1565–74.
36. Mandal S. Low-power CMOS rectifier design for RFID applications. IEEE Trans Circuits Syst. 2007;54(6):1177–88.
37. Chiu H-W, Lin M-L, Lin C-W, Ho I-H, Lin W-T, Fang P-H, Lee Y-C, Wen Y-R, Lu S-S. Pain control on demand based on pulsed radio-frequency stimulation of the dorsal root Ganglion using a batteryless implantable CMOS SoC. IEEE Trans Biomed Circuits Syst. 2010;4(6): 350–9.
38. Zhang Y, Zhang F, Shakhsheer Y, Silver JD, Klinefelter A, Nagaraju M. A batteryless 19 W MICS/ISM-band energy harvesting body sensor node SoC for ExG applications. IEEE J Solid-State Circuits. 2013;48(1):199–213.
39. Liu Y-H, Lin T-H. A wideband PLL-based G/FSK transmitter in 0.18-μm CMOS. IEEE J Solid-State Circuits. 2009;44(9):2452–62.

Design of Ultra-Low-Power Electrocardiography Sensors

Xiaoyang Zhang, Yongfu Li, Lei Wang, Wei Zou, Yinan Sun, Yongpan Liu, Huazhong Yang, Yong Lian, and Bo Zhao

Abstract In this chapter, we present two key designs for ultra-low-power electrocardiography sensors, i.e., an event-driven analog-to-digital converter (ADC) and an on-off keying (OOK) transceiver. For the ADC, two QRS detection algorithms, pulse-triggered (PUT) and time-assisted PUT (t-PUT), are proposed based on the level-crossing events generated from the ADC. For the transceiver SoC, we propose a novel supply isolation scheme to avoid the instability induced by such a high receiver gain, use bond wires as inductors to reduce the transmitter power, and utilize near-threshold design (NTD) method for low power digital baseband. Fabricated in 0.13 μm CMOS technology, the ADC with QRS detector consumes only 220 nW measured under 300 mV power supply, making it the first nanoWatt compact analog-to-information (A2I) converter with embedded QRS detector. The transceiver SoC is fully integrated with a 10 Mb/s transceiver, digital processing units, an 8051 micro-controlled unit (MCU), a successive approximation (SAR) ADC, and etc. The receiver consumes 0.214 nJ/bit at −65 dBm sensitivity, and the Tx energy efficiency is 0.285 nJ/bit at an output power of −5.4 dBm. In addition, the digital baseband consumes 34.8 pJ/bit with its supply voltage lowered to 0.55 V, indicating its energy per bit is reduced to nearly 1/4 of the super-threshold operation.

1 Introduction

Cardiovascular disease has been identified among the leading health concerns in many countries. American Heart Association's recent research [1] shows that the annual direct cost of heart related disorders in the US amounts to $179 billion. The chances of total and fast recovery of the patient from heart related disorders

X. Zhang • Y. Li • L. Wang • Y. Lian
ECE Department, National University of Singapore, Singapore, Singapore
e-mail: elezhx@nus.edu.sg; liyongfu@nus.edu.sg; elewngl@nus.edu.sg; eleliany@nus.edu.sg

W. Zou • Y. Sun • Y. Liu • H. Yang • B. Zhao (✉)
Tsinghua University, Beijing, China
e-mail: zouw10@mails.tsinghua.edu.cn; syn08@mails.tsinghua.edu.cn; ypliu@tsinghua.edu.cn; yanghz@tsinghua.edu.cn; zhao_bo@tsinghua.edu.cn

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_14

are diminished due to the late detection of the symptoms. Therefore, patients diagnosed with heart defects, or those who are at high risk, need continuous Electrocardiography (ECG) monitoring solutions.

Wearable wireless ECG sensor is one of the best candidates for continuous heart condition monitoring [2–5]. However, the design remains challenging due to stringent constraints on weight, size, and power consumption. Achieving long battery life or self-powered sensor is the ultimate goal in such applications as it facilitates continuous recording of ECG signal without causing too much inconvenience to the patient. The significant improvement of energy efficiency in a wireless ECG sensor is only possible if the power can be reduced either by data compression or pre-filtering at the front-end and analog-to-digital converter (ADC) side or by duty cycling the wireless transmitter. This chapter focuses on the discussion of the two systems, i.e., analog-to-information (A2I) system with Real-time QRS detection and a fully integrated OOK transceiver system-on-chip (SoC).

The first system aims to reduce the effective data rate, which is directly proportional to the wireless power consumption. When the heart rate information is required instead of the ECG signal, the sensor performs QRS detection locally, reducing the data to several samples per second and the power consumption significantly. This is realised through the integration of the signal processing tasks into an analog-to-digital converter without incurring much hardware overhead. Two algorithms, i.e. PUlse-Triggered (PUT) QRS detection and time-assisted PUT (t-PUT), are introduced. Both algorithms are verified in simulation using all 48 modified lead II (MLII) ECG client records from MIT-BIH Arrhythmia Database [6], with over 99 % sensitivity and positive prediction for most records. Compared to the Nyquist ADC based system, the event-driven nature of A2I system not only reduces the number of sample points, but also improves power efficiency [7]. Fabricated in 0.13 μm CMOS technology, the ADC with QRS detector consumes only 220 nW measured under 0.3 V power supply, making it the first nanoWatt compact A2I converter with embedded QRS detector. The measurement results demonstrate the potential of the A2I system in terms of power efficiency and simplicity in hardware implementation.

The second system is a fully integrated wireless transceiver SoC system with a 10 Mb/s on-off keying (OOK) RF transceiver, digital processing units, a 8051 micro-controlled unit (MCU) and a successive approximation register (SAR) ADC. The receiver adopts envelop detector (ED) based structure to improve the energy efficiency. Conventional ED based structure has a poor sensitivity when reaching a bit rate of Mb/s level. Envelope detector (ED) based OOK receiver has shown a better Receiver energy efficiency, e.g. 0.5 nJ/bit in [8] and 0.295 nJ/bit in [9], respectively. However, ED based approach usually leads to poor receiver sensitivities, e.g. −37 dBm in [8] and −45 dBm in [9], which result in a limited communication distance. A 5 Mb/s super-regenerative OOK transceiver was reported in [10], which boosted energy efficiency to 0.363 nJ/bit. The high efficiency, however, was at the cost of several off-chip components and increased device size. To resolve the problem, we design a receiving (Rx) front-end with 77 dB gain at 10 Mb/s data rate, and introduce a better supply isolation scheme to avoid the instability induced by such a high gain. The transmitter is based on a 2 GHz digitally controlled oscillator (DCO), which uses bond wires as inductors to further reduce the power at transmitting (Tx) mode. The digital baseband is designed by a near-threshold design (NTD) method for low power consumption. The chip is implemented with 0.13 μm CMOS technology, measured results show that the receiver consumes 0.214 nJ/bit at −65 dBm sensitivity, and the Tx energy efficiency is 0.285 nJ/bit with an output power of −5.4 dBm. In addition, the digital baseband consumes 34.8 pJ/bit with its supply voltage lowered to 0.55 V, indicating its energy per bit is reduced to nearly 1/4 of the super-threshold operation. To achieve even higher energy efficiency, Impulse radio ultra-wideband (IR-UWB) could be adopted. IR-UWB consumes significantly less power than traditional carrier based transceiver because of its carrier-less nature. UWB transceivers can be designed by simple circuits without up/down converters. During pulses, they could also be turned off between pulses to be duty cycling. However, the communication range is typically smaller than the narrow band counterpart.

2 Event-Driven ADC with Real-Time QRS Detection

The architecture of the event-driven QRS processor is shown in Fig. 1. The major block, an event-driven ADC [11], converts analog input into level-crossing (LC) events. The subsequent real-time QRS detection block extracts the QRS information from the ADC's output stream.

The event-driven ADC includes two asynchronous comparators, a digital-to-analog converter (DAC) and an asynchronous digital control unit consisting of a latch, an event generator, a matched delay block and a shift register for event processing. The two comparators, C_{UPPER} and C_{LOWER}, continuously compare analog input with two voltage levels, V_{UPPER} and V_{LOWER}, generated by the DAC.

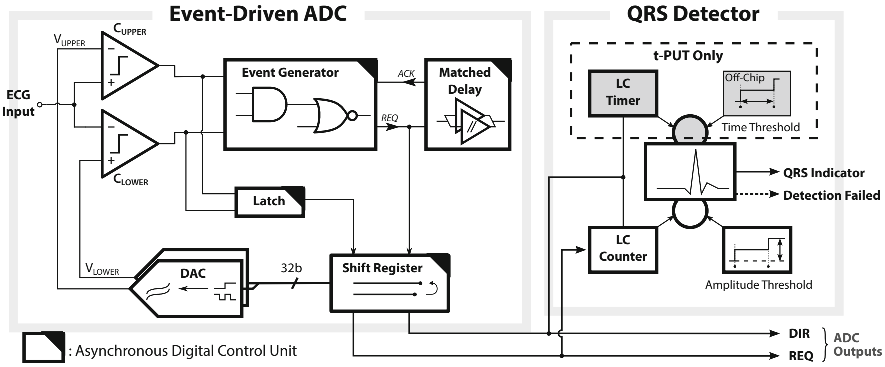


Fig. 1 Diagram of event-driven QRS processor

The difference between V_{UPPER} and V_{LOWER} is one least-significant-bit (LSB) voltage ΔV. When the input voltage level rises above V_{UPPER} such that $V_{INPUT} > V_{UPPER} > V_{LOWER}$, the comparator C_{UPPER}'s output turns high while C_{LOWER}'s remains low. The latch's output is high, so the event generator increases the shift register's output by 1 bit, raising the DAC's outputs V_{UPPER} and V_{LOWER} each by ΔV. This process is defined as a RISE level-crossing event, or {RISE} for simplicity. If C_{UPPER}'s output remains high after V_{UPPER} and V_{LOWER} update, another {RISE} event follows and raises the DAC's output levels further. The matched delay block controls the minimal interval of two events, which is slightly longer than ADC loop delay. It ensures that the condition $V_{UPPER} > V_{INPUT} > V_{LOWER}$ is satisfied eventually after the comparators respond to the new values. Similarly, when $V_{INPUT} < V_{LOWER}$, the shift register's output is decreased by 1 bit, and the DAC lowers V_{UPPER} and V_{LOWER} each by a ΔV. This is referred as a FALL level-crossing event and noted as {FALL}.

The ADC's outputs are encoded as delta-modulated 2-bit stream [12], DIR and REQ, in our system as indicated in the lower right corner of Fig. 1. DIR represents the signal direction, i.e. the rise or fall of input voltage level. REQ indicates the occurrence of the {RISE} or {FALL} events. An example of the ADC outputs is illustrated in Fig. 2 for an ECG signal. Whenever there is a {RISE} (or {FALL}), event indicator REQ outputs a short pulse, and the level direction indicator DIR turns high to indicate a rise in voltage (or remains low for a voltage fall). Generated from {RISE} or {FALL} events by the digital control units, the REQ and DIR outputs represent the input activity, and are the input signals for the subsequent QRS detector.

The QRS detector shown in the right part of Fig. 1 consists of a LC counter, an amplitude-threshold-setting block, and related logic controls. The t-PUT QRS requires additional timer and time-threshold-setting blocks. The blocks for t-PUT only are shaded in grey. They are switched off when only PUT is activated. Both the LC counter and the LC timer receive information from the ADC. The LC counter

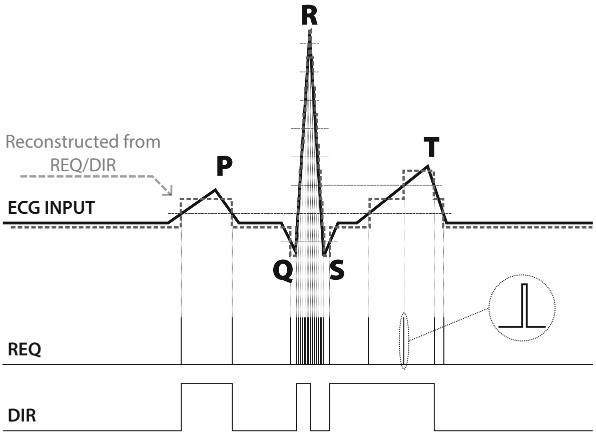


Fig. 2 Delta-modulated event-driven ADC outputs for an ECG signal

counts the number of monotonic {RISE} or {FALL} events. The LC timer measures the duration of every monotonic event sequence. Details about the QRS detection and circuit implementations are discussed in the following sections.

2.1 QRS Detection Algorithms and Performance Evaluations

The operational principle of PUT and t-PUT algorithms is illustrated in Fig. 3. As t-PUT is improved from PUT, the PUT algorithm is introduced first. The PUT involves three main steps.

1. The algorithm starts with identifying the Q wave. Every trough point in the input is a possible candidate of Q wave. It is identified by a {FALL} followed by a {RISE}, noted as a {FALL, RISE} sequence, in the ADC output. A valid Q wave should consist of a {FALL, RISE} followed by a number of uninterrupted {RISE} events. The LC counter in the QRS block tracks the

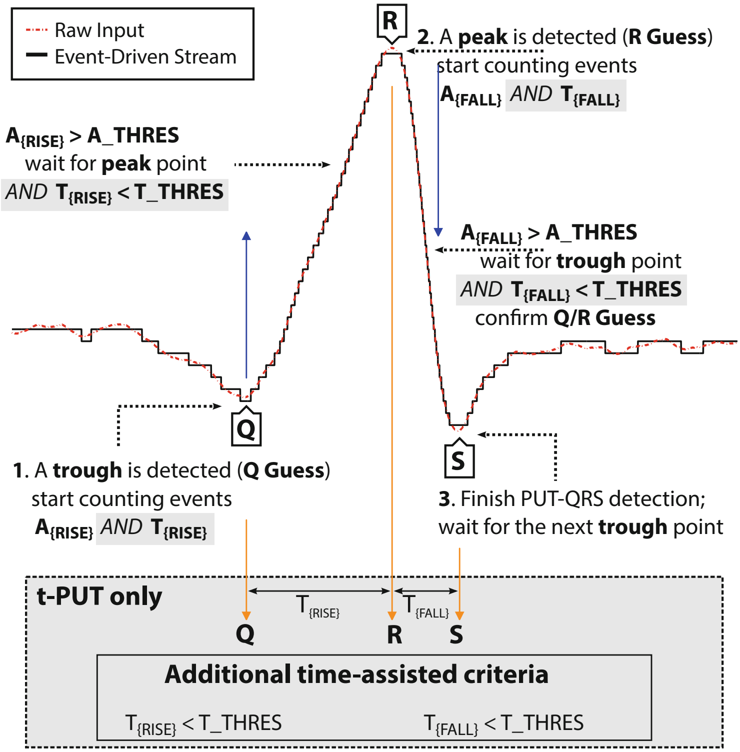


Fig. 3 Pulse-triggered (PUT) and time-assisted PUT (t-PUT) QRS detection algorithms, with t-PUT related part in *grey boxes*

number of uninterrupted {RISE} events. The counter starts counting when a {FALL, RISE} occurs. Each subsequent {RISE} event increases the counter by 1. The counter resets its value when one of the following two conditions is met: (1) the pre-defined threshold value A_THRES is reached; (2) a {FALL} event occurs before the counter reaching A_THRES. A_THRES is defined as the minimal number of {RISE} events for a rising edge to be qualified as a Q-R interval. If the LC counter resets due to the first condition, the Q wave identification process completes and the algorithm moves to Step 2 for R peak detection. Otherwise, the algorithm goes back to beginning and waits for a new {FALL, RISE} sequence.

2. The R wave identification and confirmation start with peak detection. The identified peak is first marked as an unconfirmed R guess when a {FALL} is detected in the Q-R edge. This is referred as a {RISE, FALL} sequence. To confirm the detected peak is indeed an R peak, the LC counter starts to count the number of uninterrupted {FALL} events. The counter resets itself based on the two conditions similar to those in Step 1. The only difference is that the A_THRES represents the minimal number of {FALL} events for a falling edge to be qualified as an R-S interval. Note that it uses the same value of A_THRES for identifying Q-R and R-S intervals because of the similarity between these two intervals. If the counter resets due to the first condition, the R wave is successfully detected, i.e. the previous R peak guess is confirmed. The output “QRS Indicator” is asserted. The algorithm moves to next step to complete the detection process. Otherwise, the output “Detection Failed” is asserted. The algorithm goes to the beginning and waits for a new {FALL, RISE} sequence.
3. The S wave detection is straightforward, i.e. detecting the arrival of a {RISE} event on the R-S edge. Once the {RISE} is detected, the QRS detector resets the outputs and is ready for next QRS.

By choosing a proper A_THRES value, PUT-QRS is capable of distinguishing true QRS waves from P/T waves or small fluctuations, which are usually smaller in amplitude compared with QRS peaks. In our design, A_THRES is set to 7 for both Q-R and R-S edges, representing a 3-bit counter for the LC. A high-performance front end also improves the PUT-QRS performance by suppressing the noise and the power-line interference.

However, when a high T wave occurs or the amplitude of ECG signal changes abruptly, PUT-QRS may generate false QRS results under its fixed threshold setting. To solve this problem, a LC timer is added to improve the PUT-QRS, which is shown earlier in a grey box in Fig. 1. With the LC timer, the new time-assisted PUT makes use of both the amplitude (number of monotonic LC events) and the time (duration of monotonic LC events) information from the event-driven ADC’s outputs.

The timing characteristics of QRS complex add another dimension to QRS detection [13] and form the basis for t-PUT QRS identification. A normal QRS duration is less than 0.1 s. Therefore a large pulse lasting longer than 0.1 s is unlikely to be a healthy QRS complex. This additional criterion helps to differentiate the true QRS complex from other large pulses.

The additional components in t-PUT are highlighted with grey shaded areas in Figs. 3. It can be seen that a LC timer and a time threshold T_THRES are introduced in t-PUT. The T_THRES is defined as the longest possible duration of a Q-R (or an R-S) wave. Although the LC counter and the LC timer count the amplitude and the duration of a rising (or falling) edge, the operations of these counters are different. The LC counter resets itself once its value reaches A_THRES or a {FALL} (or {RISE}) arrives before it accumulates to A_THRES. The LC timer, on the other hand, does not stop counting until a {FALL} (or {RISE}) occurs in a Q-R (or an R-S) edge. The Q-R or R-S edge is confirmed in t-PUT algorithm only if the timer value T is less than T_THRES and counter value A > A_THRES. Such additional criterion improves QRS detection accuracy over PUT, especially for ECG signals with large P or T waves.

The time-assisted PUT detector addresses the detection errors due to high T wave. The additional time information is used to filter out T waves which are high in amplitude but last much longer than that of a normal QRS complex. Figure 4 shows the instant heart rates calculated from both detectors and annotated heart rates in the database. The input is a 30-min ECG from Record 222. The upper graph is the result from PUT-QRS detector. There are 11 large positive spikes shown in the PUT heart rate, each representing a false positive detection error. Also there are 5 missed beats in PUT-QRS. The plot in the middle shows the result from the t-PUT algorithm, and lower one is the annotated heart rates from the MIT-BIH database. All the false negative detects and some false positive detects are corrected in t-PUT QRS.

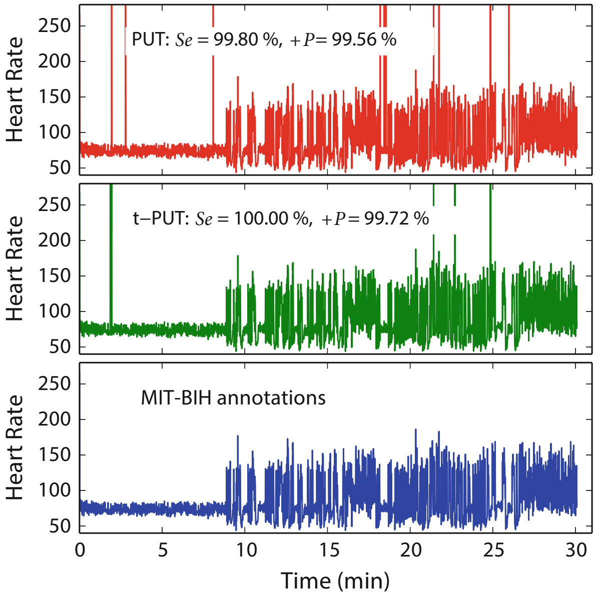


Fig. 4 Heart rate calculated based on R-R interval

Table 1 Performance comparison with published QRS detection methods

Method	Se(%)	+P(%)	Ref
Wavelet transform	99.90	99.94	[14]
Input-feature correlated	79.33[a]	98.55[a]	[15]
Filter bank	99.59	99.94	[16]
Genetic algorithm	99.60	99.94	[17]
Mathematical morphology	99.38	99.94	[18]
PUT	97.63	97.33	–
t-PUT	97.76	98.59	–

[a] Results using all records in MIT-BIH database based on our simulations without 5–35 Hz bandpass filtering

One advantage of PUT-QRS algorithm is its very low hardware complexity and potential for low power implementation. More importantly, the simplicity of PUT-QRS does not drastically reduce the detection accuracy as shown in Table 1. Further improvements can be made to PUT algorithm by making A_THRES programmable, which can be tuned by either physicians or a calibration algorithm.

The performance of PUT and t-PUT QRS detectors is compared with some well-established methods based on traditional digital signal processing (DSP) techniques and an A2I based QRS detector as shown in Table 1. Sensitivity (*Se*) and positive prediction ($+P$) are used to represent the detection accuracy.

$$Se(\%) = \frac{TP}{TP+FN} \tag{1}$$

$$+P(\%) = \frac{TP}{TP+FP} \tag{2}$$

where TP is the number of total QRS peaks from database annotations, FN is the number of false negative errors, and FP is the number of false negative errors.

Our extensive simulations also suggest that the PUT and t-PUT algorithms are robust when handling abnormal ECG signals. The reasons are two-fold. First, the PUT treats each QRS independently, which promises fast start-up and consistent result for every QRS peak. Second, the PUT detector is less affected by low-frequency noise and baseline fluctuations. It can quickly recover from detection errors even when there are abnormal peaks or troughs. Figure 5 illustrates a Z from Record 105, which has two exceptional pulses, one high R peak and one low trough. The unexpected high R peak affects the detection accuracy if previous peaks are involved in the decision of current peak as IFC does. Similarly, a sudden low trough drastically changes the average value for Q troughs, and leads to detection errors in the subsequent Q waves. It is clear from Fig. 5 that the PUT handles such abnormal ECG well.

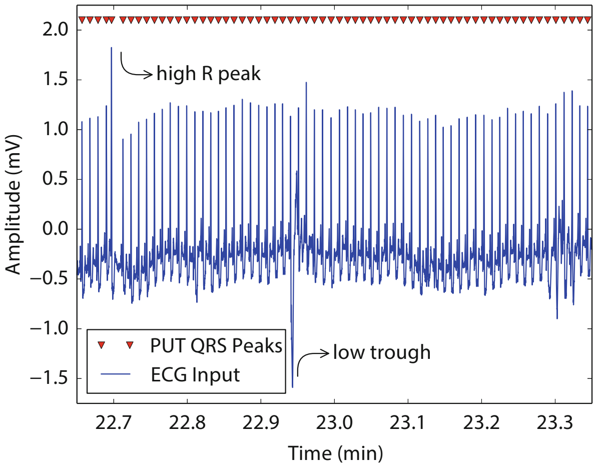


Fig. 5 Simulated QRS detection results for abnormal ECG signals

2.2 Circuit Design Considerations

ECG signals are within the frequency band of 0.05 $\sim$ 250 Hz. Such a low frequency allows the event-driven system to work at a very low speed. This feature provides us an opportunity to aggressively lower the supply voltage to 0.3 V. As discussed in [19–21], the performance of event-driven ADCs is mainly determined by ADC's feedback loop delay, the comparator's resolution, and the DAC's resolution. Under 0.3 V supply voltage, all transistors operate in the subthreshold or cutoff region. The static current for analog blocks such as comparators is substantially reduced, which restricts the circuit speed and input voltage range. The designer needs to find a proper balance between power consumption and system performance. This section highlights the design considerations under such a low supply voltage.

2.2.1 0.3 V Process-Insensitive Comparator

The three-stage comparator used in our design is shown in Fig. 6. Stage 1 is a differential amplifier with rail-to-rail input range [22]. Stage 2 provides gain for the comparator and Stage 3 is an inverter buffer generating full-scale output. Several factors should be considered when designing under 0.3 V.

First, the rail-to-rail input range is crucial for analog circuits when supply voltage is reduced. This requires both PMOS and NMOS differential amplifiers for the first stage. In Stage 1, M_{P3} and M_{P4} are the active loads for $M_{N1,2}$ NMOS input pair. With current tail M_{NB}, they form an NMOS differential amplifier. Similarly, M_{PB}, $M_{P1,2}$, and $M_{N3,4}$ build up a PMOS amplifier. Both amplifiers' outputs are connected

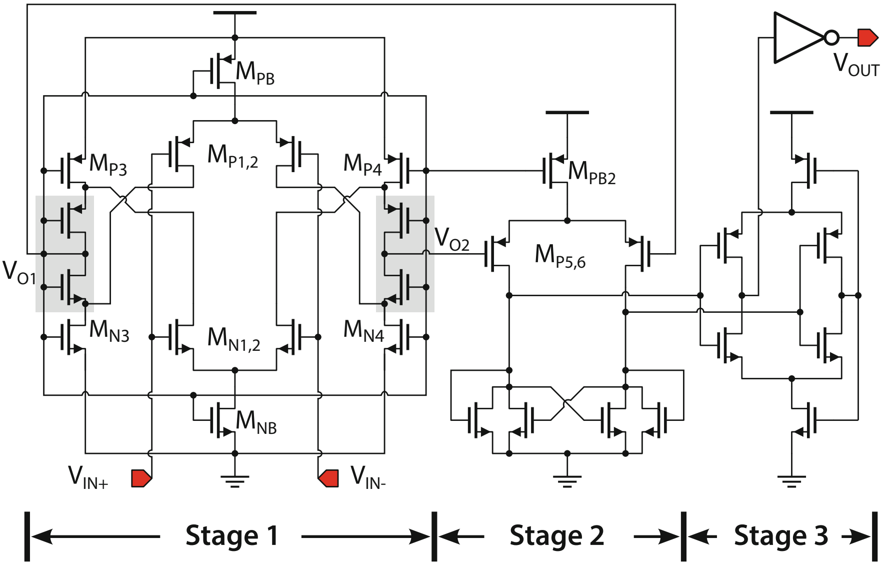


Fig. 6 3-stage comparator used in the event-driven ADC

through four transistors highlighted in grey in Fig. 6. These four transistors generate outputs V_{O1} and V_{O2} of Stage 1 by averaging the outputs of NMOS and PMOS amplifiers.

Second, high CMRR provides reliable open-loop gain under any input situations. Since V_{O1} is connected to the two diode-connected loads M_{P3} and M_{N3}, its voltage depends little on the input. By proper sizing, V_{O1} is set at 0.13 V, which biased the input PMOS transistors of Stage 2 $M_{P5,6}$ in moderate inversion region with enough headroom. Through this common-mode shifting [23], Stage 2 is able to generate high differential gain regardless of the input voltage. Without Stage 1, the gain of Stage 2 and the output delay severely depend on the input common-mode as the V_{DS} of the tail PMOS M_{PB2} is small. Considering V_{O1}'s insensitivity to input voltages, all the current tails are biased using V_{O1}. This voltage biasing technique has a disadvantage that the total current of Stage 1 is exponential to the supply voltage. But it saves the extra biasing circuits and current, and the performance is guaranteed in wide supply voltage [22].

Last but not least, the comparator uses large transistors to suppress process variation, flicker noise, and input offset. In Stage 1, the length for $M_{P1,2}$ and $M_{N3,4}$ is 3 μm. The cross-coupled loads in Stage 2 increases the DC gain. To avoid negative loading, the cross-coupled NMOS transistors are half the size of the diode-connected ones in this stage. Post-layout simulations under all corners and temperature from −10 to 80 °C show the comparator has delay less than 3.6 μs and ENOB over 8.1 bits under 300 mV supply. The comparator also fully operates under power supply from 0.2 through 1.2 V, with all-situation worst-case delay of 34.6 μs and ENOB of 7.2 bits.

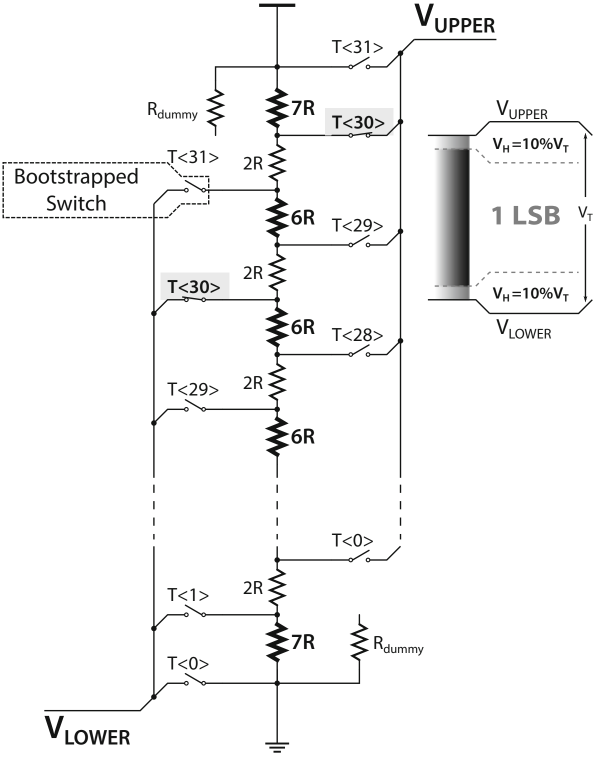


Fig. 7 DAC design with hysteresis

2.2.2 Low-Voltage DAC and System Hysteresis

Figure 7 shows the 5-bit DAC architecture. It generates two voltage references V_{UPPER} and V_{LOWER} from one string resistor ladder network. The DAC connects V_{UPPER} and V_{LOWER} to specific voltage levels through bootstrapped switches [24], which are controlled by the 32-bit shift register block. Simulations show that 5-bit DAC resolution is sufficient for reliable PUT QRS detection. The increase of resolution improves the QRS detection accuracy, especially for ECGs with large variation in amplitude. However, higher resolutions require better resistor matching and less feedback delay, which lead to higher power consumption and larger chip area.

The DAC includes 10 % hysteresis V_H for both V_{UPPER} and V_{LOWER} within each LSB ΔV to mitigate noise and fluctuations. Hysteresis removes erroneous level-crossing events due to noise, interference, or other fluctuations. The exact hysteresis value is set by the resistor ratio. As shown in Fig. 7, a resistor of 2R is inserted between 6R (or 7R) resistors. The difference between V_{UPPER} and V_{LOWER} is slightly larger than 1 LSB ΔV. In most cases when one of the switches T<1:30> is on, we have

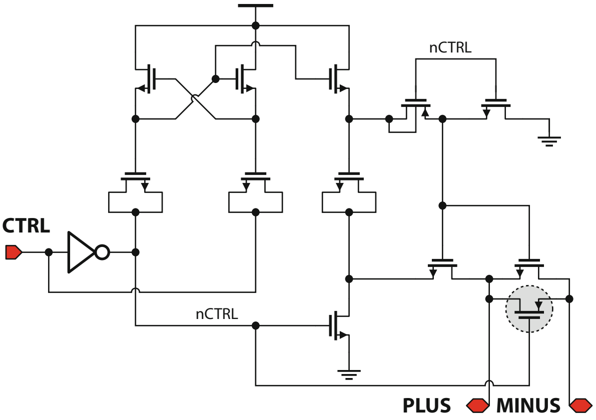


Fig. 8 Bootstrapped switches used in low-voltage DAC

$$\begin{aligned} V_T &= V_{UPPER} - V_{LOWER} \\ &= \tfrac{2R+6R+2R}{2^5 \cdot (2R+6R)} V_{DD} \\ &= 125\,\% \cdot \Delta V \end{aligned} \tag{3}$$

In this implementation, the value of R is 13 kΩ. By using a different resistor ratio, the hysteresis value changes accordingly. Increasing hysteresis improves PUT and t-PUT detection accuracy by filtering out noise. However, higher hysteresis removes ECG details, resulting in poorer ADC performance. A 10 % hysteresis provides to be a good trade-off between QRS detection accuracy and ADC performance according to our simulations.

The switch design is given in Fig. 8. All switches used in DAC are bootstrapped. As linearity is not an issue for small signals, an extra PMOS, circled in grey in Fig. 8, is added to boost the conductivity. It further improves the circuit speed and DAC accuracy under 0.3 V.

2.2.3 Asynchronous LC Timer and Delay Cell

The level-crossing timer used for time measurement in t-PUT is shown in Fig. 9. It contains 16 asynchronous unit delay cells and a counter, all connected in a loop. The LC timer measures the duration from a rising edge to the following falling edge (or from a falling edge to the following rising edge) of the DIR signal, which corresponds to the duration of uninterrupted {RISE} (or {FALL}) events. The rising and falling edges of DIR signal are first converted to a pair of START and STOP pulses as shown at the left of Fig. 9. The START pulse then propagates through the chain of 16 delay cells to the counter. The delayed pulse increases the counter value by 1 and triggers the counter to generate a new pulse, which is looped back to the input of delay chain. This process continues, until the STOP pulse arrives and

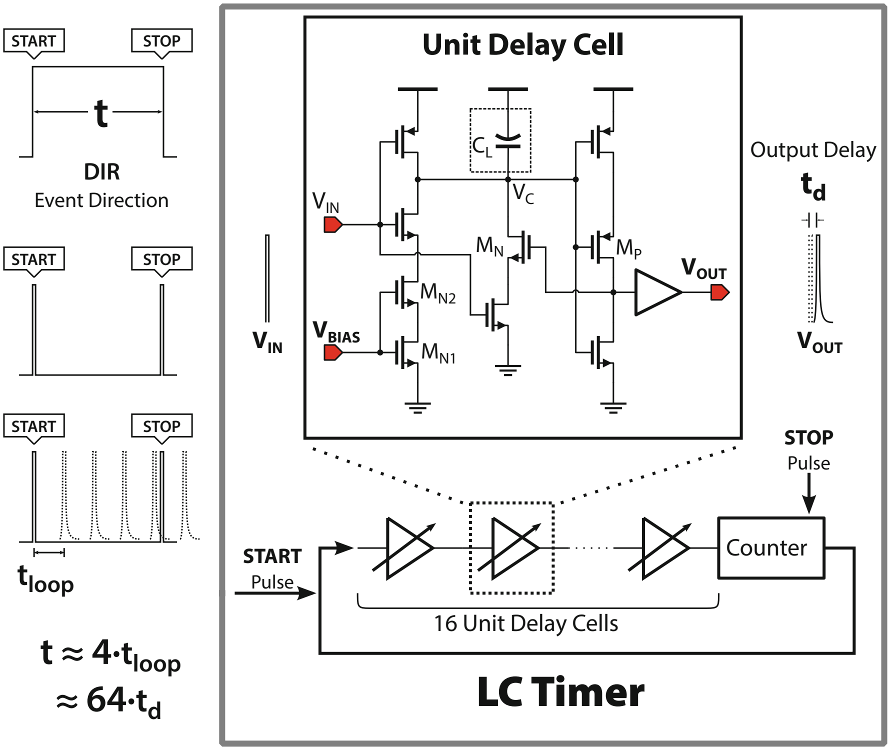


Fig. 9 LC timer and delay cell for timing control

the counter stops counting. The LC timer outputs the counter value and resets the counter, waiting for next START pulse. As each delay cell provides a delay of t_d, the counter value N can be converted into the duration of DIR signal, i.e. $t = N \cdot 16 \cdot t_d$.

The delay cell is based on a current-starved digital buffer [25, 26]. When V_{IN} is low, C_L is charged and V_{OUT} is turned low. When V_{IN} turns high, C_L gets discharged slowly through M_{N2} and M_{N1}, until the voltage at V_C falls to a threshold value V_{TH} and turns M_P moderately on. The positive feedback [27], formed by M_N and M_P, accelerates the remaining discharging process. The acceleration helps to reduce short current. Two small transistors M_{N2} and M_{N1} are used to limit the discharging current, which increase the delay time to several milliseconds. This delay time is adjustable by changing the V_{BIAS} voltage.

The matched delay block in the ADC uses the same delay cell to avoid race conditions. When a new event is generated, the matched delay block locks this event for t_{lock} time, until the DAC updates V_{UPPER} and V_{LOWER}, and comparators are ready to respond to the new references. Together with the event generator, the matched delay block establishes a simple hazard-free 4-phase asynchronous handshaking protocol. In our design, a delay chain consisting of 8 delay cells are used to generate tunable delay t_{lock}.

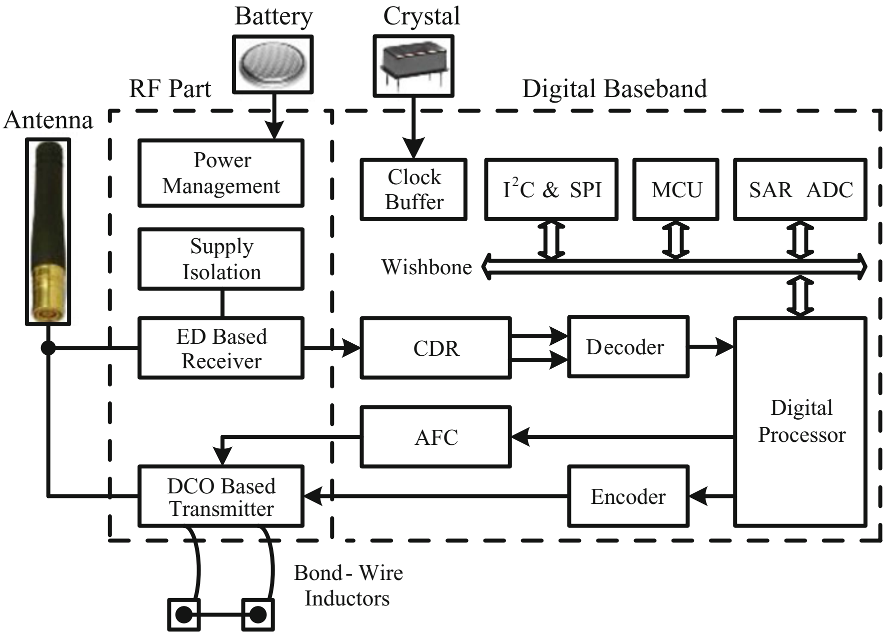


Fig. 10 System architecture

3 Transceiver SoC

The SoC architecture is shown in Fig. 10, assisted by an antenna, a 1.5 V button battery, a crystal, and bond-wire inductors. There are two main blocks in the SoC, i.e. RF part and digital baseband. The RF transceiver contains an ED based receiver and a DCO based transmitter, whereas a novel low-power supply isolation scheme is introduced for the receiver. The blocks of digital baseband include clock-data recovery (CDR), automatic frequency calibration (AFC), MCU, encoder, decoder, and etc. The AFC module is used to calibrate the frequency deviations induced by the bond-wire inductors of transmitter. The digital submodules exchange data by a wishbone bus, and I^{2}C and SPI bus are also supported. A 10-bit SAR ADC with 50-kS/s sampling rate is also adopted here as an interface for sensor.

3.1 Transceiver Design

The block diagram of OOK transceiver is presented in Fig. 11. The receiver consists of a low-noise amplifier (LNA), a cascaded amplifier (CA) for gain enhancement, and several other blocks for demodulation, i.e. ED, baseband amplifier (BA), and comparator. The use of CA is mainly for the sensitivity improvement. To achieve a

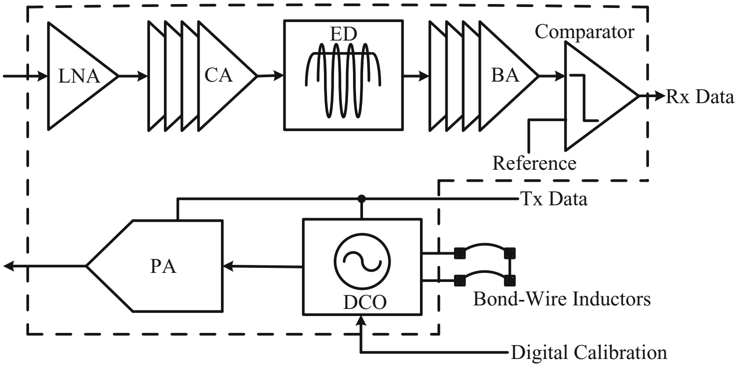


Fig. 11 Transceiver architecture

sensitivity of −60 to −70 dBm while taking into consideration of at least 100 mV input for the ED, the front-end must provide gains in the range of 60–70 dB. LNA itself is not sufficient for such a large gain, so an inverter-based structure is adopted for CA to achieve a high gain.

The transmitter contains a DCO and a power amplifier (PA), where baseband data are used to directly control the power switch of the two modules for lowering the average power at Tx mode. The DCO frequency deviations induced by bond-wire inductors are calibrated by digital baseband.

3.1.1 Power Isolation Scheme for Receiver

The overall gain for LNA and CA is set to 70 dB. To balance noise figure and power, the LNA is designed with current-reused common-gate typology [28] for 10 dB gain. The rest of gain is provided by the CA, which was set to 67 dB by taking into consideration of process variations. One of issues in design high gain inverter-based CA is stability, especially under low-power condition where all inverters are single-end without bias current. As a result, every two stages form a closed loop with a forward path for RF signal and a backward path for power line, as shown in Fig. 12. Instability may appear if the gain of each loop is higher than 0 dB. Figure 13 shows the close-loop gains versus frequencies at different forward gains (FGs). It can be seen that the loop gain is raised to 0 dB level when the forward gain becomes higher than 66 dB. In our design, the Rx front-end gain is set to 77 dB which will lead to unstable state at the output of Rx front-end.

To maintain the stability while achieving 77 dB of gain, we introduce a new supply isolation scheme. Different from traditional stable power management schemes that utilize complex regulators to minimize the supply interactions among sub-modules [29], we present a power management scheme based on simple supply isolators (SIs), as shown in Fig. 14, where every sub-module is protected by a SI, except the BA. The SI consists of R, C and M1 as shown in dotted box in Fig. 14. The corner frequency of RC is set to 2.5 MHz.

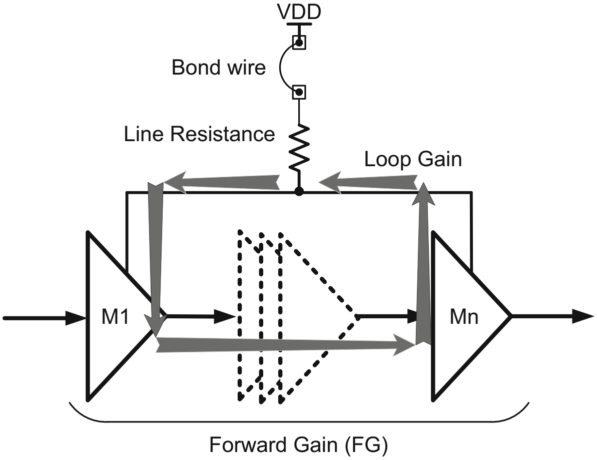


Fig. 12 Close loop between each two stages

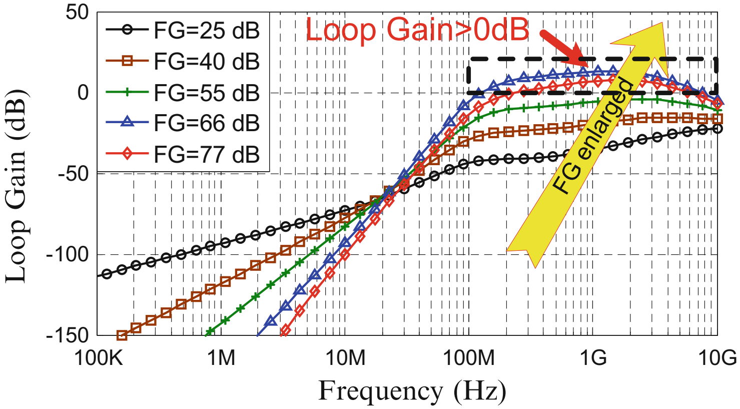


Fig. 13 Close-loop gains at different FGs

The supply current of LNA changes little since it processes small signals. For the inverter-based CA, two adjacent inverters are attached to one SI; so overall supply current of these two stages has small fluctuations because the currents of these two inverters are complementary to each other, as show in the dashed box of Fig. 14. The ED is a victim of supply noise due to its low current of 30 μA, so it is protected by the SI. BA is the most significant supply-noise aggressor for two reasons: First, its signal amplitude and power consumption are relatively large; second, the BA is used to amplify the low-frequency signals at analog baseband, generating low-frequency supply noise which is not easy to remove by the SI. Therefore, a dedicated shunt regulator [30] is introduced to block its supply noise.

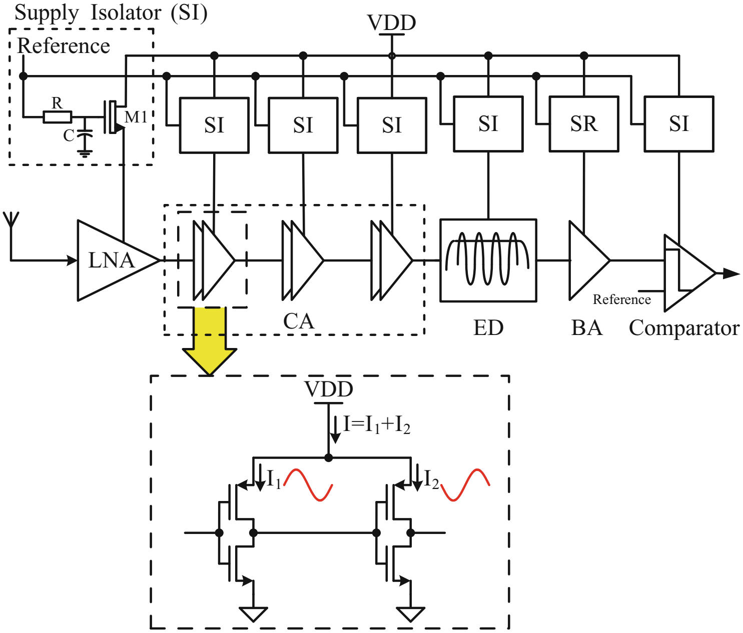


Fig. 14 Low-power supply isolation scheme

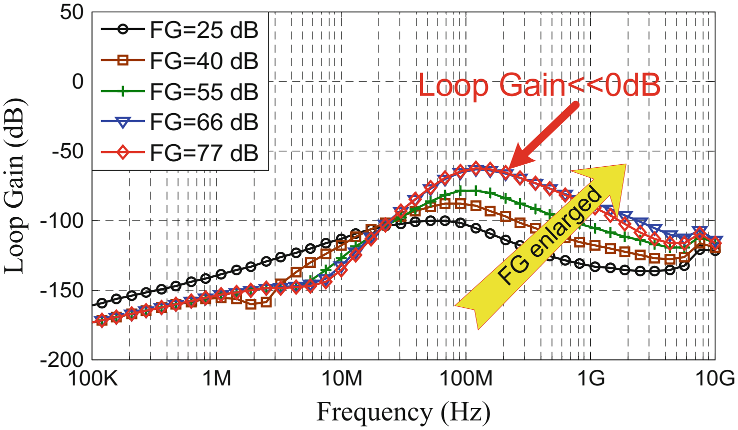


Fig. 15 Close-loop gains under the supply isolation scheme

With newly introduced power isolation scheme, the close-loop gain is kept far below 0 dB while the forward gain attains 77 dB as illustrated in Fig. 15, so the receiver maintains its stability under such a high gain.

The current consumption of LNA, CA, ED, and BA are 430 μA, 700 μA, 30 μA, and 140 μA, respectively. As a result, the overall power of receiver can be less than 1.43 mA.

3.1.2 Bond-Wire Transmitter

The OOK transmitter contains a bond-wire DCO and a PA, as shown in Fig. 16. The Tx data modulates the transmitter by switching the supply power of all blocks. The maximum RF power is −5.4 dBm.

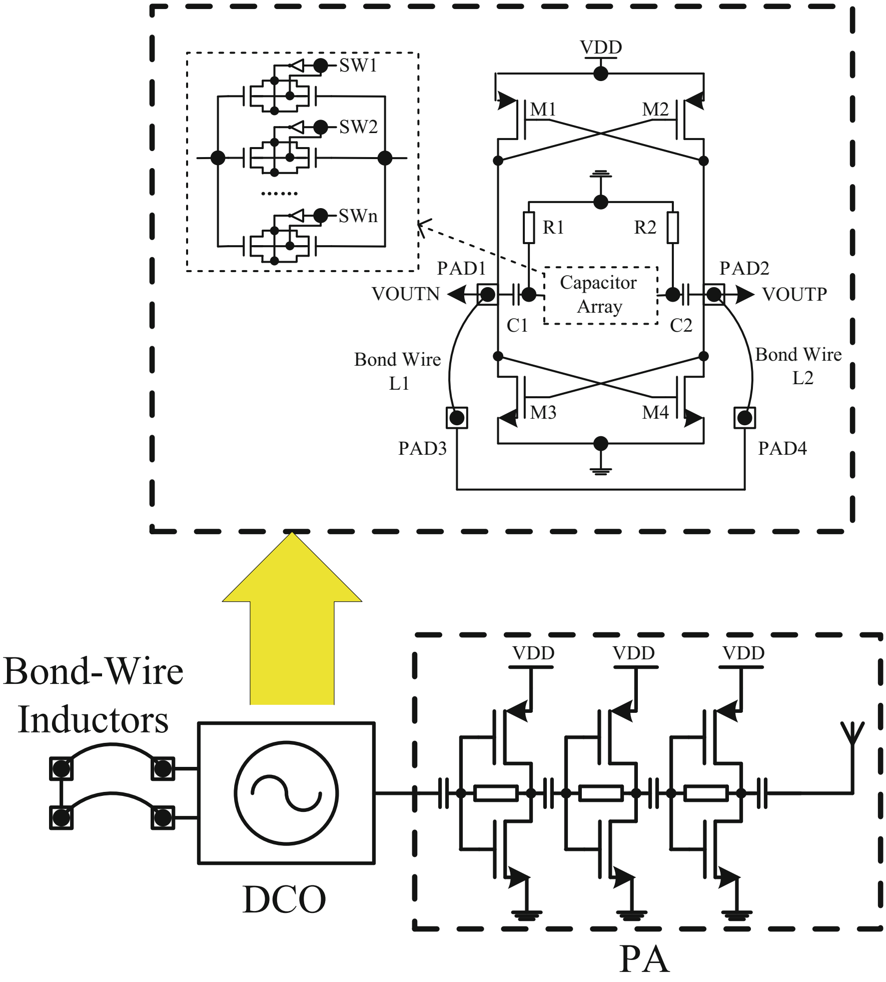


Fig. 16 Transmitter architecture

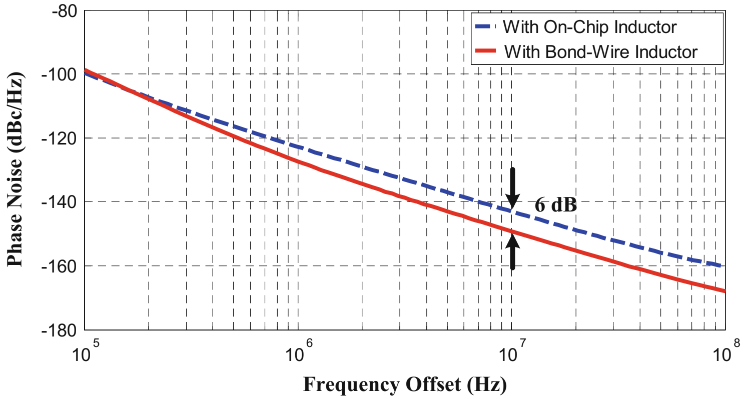


Fig. 17 Bond-wire induced noise improvement of DCO

The DCO uses bond wires as inductors. The DCO circuit is shown in Fig. 16, where two on-die pads are connected to the differential output port, whereas another two are the bond pads for packaging. A digitally controlled MOS capacitor array is used to calibrate the frequency deviations caused by the bond wires. The AFC is realized by the digital baseband.

Bond-wire DCO helps reduce the power due to high Q-factor of bond wires. As the Tx/Rx data rate is 10 Mb/s, the phase noise at 10 MHz offset decides the error vector magnitude (EVM) of transmitter. For the same DCO, and we compare its phase noise in case of on-chip inductors and bond-wire inductors in Fig. 17, indicating that our bond-wire technique improves phase noise by 6 dB at data rate of 10 MHz. Therefore, the power can be reduced by 50 % since the oscillator phase noise is inversely proportional to the square of current consumption [31].

3.2 Digital Baseband

The power reduction of digital baseband is realized by NTD. Compatible with the OOK RF specifications, the digital baseband contains direct sequence spectrum spread (DSSS) and packet processing units. To address process, voltage and temperature (PVT) issues in NTD, we filtered out some standard cells with minimal size or multi-stack transistors that are vulnerable to PVT variations. Therefore, the digital part can work robustly under an ultra-low supply voltage such as 0.4 V, supporting a low-speed mode with 30 μW power consumption.

4 Measurement Results

4.1 Event-Driven ADC and QRS Detector

This ADC-QRS chip is fabricated in a 0.13 μm standard CMOS process. It includes the event-driven ADC, the complete PUT-QRS detector, and the LC timer for t-PUT detection. The total core area is $420 \times 850\,\mu m^2$. The die photo is shown in Fig. 18.

The event-driven ADC is first tested using 0.3 V full-swing 50-Hz sinusoidal input. The delta-modulated outputs REQ and DIR are first captured to reconstruct the input, and the recontructed input is resampled at a higher frequency of 25 kHz before power spectrum analysis through (Fast Fourier Transform) FFT. Based on the FFT result, the signal-to-noise and distortion ratio (SNDR) is 28.3 dB. We also measured the maximum input frequency without slope overload [32] at 1.2 kHz. At higher frequencies the ADC loop delay is too large to track the level-crossing events.

The functions of the overall system are verified with the help of a Fluke ECG simulator. The Lead II output of the simulator is first amplified through an SR560 low-noise voltage preamplifier to around 0.3 V_{PP}, and then connected to the ADC's input. Figure 19 shows the measured PUT-QRS detection results, delta-modulated outputs and the reconstructed signal. As designed, the QRS output is only activated during the R-S interval. Under 0.3 V supply, the total power consumption of the system is 220 nW using the ECG input. The system also fully functions under higher supply voltage up to 0.6 V.

To test the t-PUT QRS detector, the delay chain used in the LC timer needs to be characterized. By changing the bias voltage V_{BIAS} of the delay cells from 0 to 0.3 V, the delay t_{loop} is tuned from 14.1 to 976 μs. In order to measure the QRS duration of about 0.1 s, an external 8-bit counter is required for the LC timer, assuming using the maximum delay by setting $V_{BIAS} = 0$.

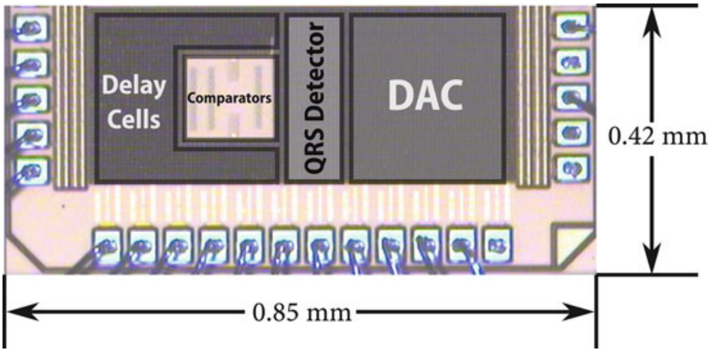


Fig. 18 Micro-photograph of the fabricated event-driven system chip

Fig. 19 Chip testing results using ECG simulator input

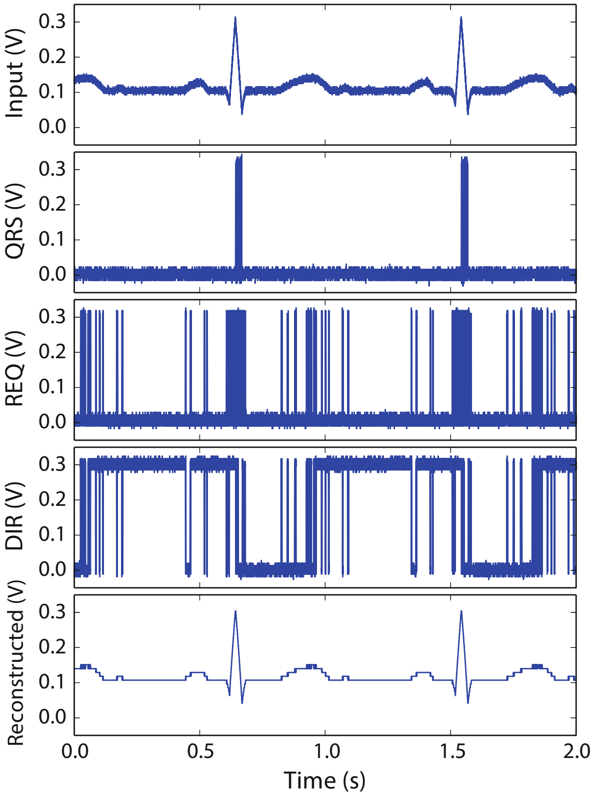


Figure 20 gives the power consumption and area breakdowns through post-layout simulations. The analog blocks, including the two comparators and the DAC, consume over 84 % of the total power, while the QRS detector consumes less than 9.5 %. Therefore, the total power consumption changes little with the input frequency, which is also verified in our measurements. To our best knowledge, it has the lowest power consumption among all reported QRS detectors. In terms of area, the resistor array in the DAC takes up over half of the total chip area, followed by the delay cells.

Tables 2 and 3 compare this work with recently published event-driven ADCs and low-power QRS detectors. Our ADC uses the lowest supply voltage and consumes the lowest power. The presented QRS detectors achieve similar performance for most MIT-BIH database records compared to [37]. There are only 4 records, i.e. 200, 203, 217, and 233, where [37] demonstrates better sensitivity and positive prediction. Note that the comparison excluded Record 207 because [37] does not count episodes of ventricular flutter in the record. The overall better performance of [37] was achieved at the cost of larger area and an order of magnitude higher power.

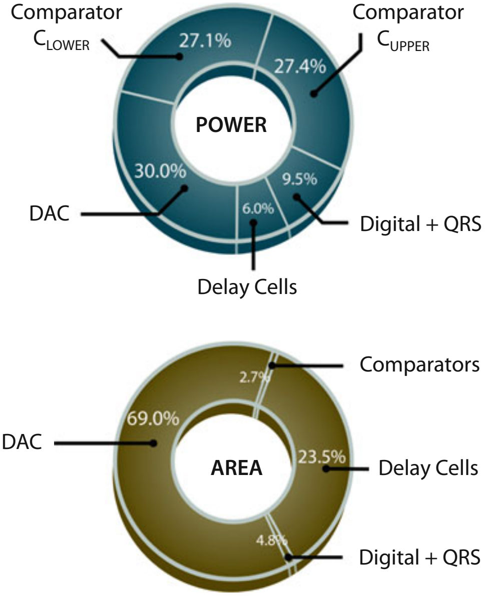


Fig. 20 Power and area breakdowns for the whole system

Table 2 Comparison of low-power event-driven ADCs

	[25]	[33]	[34]	[35]	[36]	This work
Process (μm)	0.09	0.18	0.13	0.5	0.18	0.13
Supply (V)	1	0.7	0.8	3.3	0.8	0.3
SNDR (dB)	47	43.2	47	31	49	28.3[b]
ENOB (bits)	7.5	6.9	7.5	4.8	7.9	4.4
Max freq. (kHz)	20	4	20	5	5	1
Power (μW)	50	25[a]	3	8.25	0.313	0.22[b]
Area (mm^2)	0.06	0.96[a]	0.36	0.06	0.045	0.36[b]

[a] Analog part only

[b] ADC and QRS; measured at 50 Hz input

Table 3 Comparison of low-power QRS detectors

	[37]	[38]	[39]	[40]	t-PUT
Process	0.35 μm	45 nm	0.35 μm	0.18 μm	0.13 μm
Supply (V)	1.8	0.34	3.3	–	0.3
Se (%)	99.31[a]	∼96[b]	99.81[c]	95.65	97.76[c]
$+P$ (%)	99.70[a]	∼95[b]	99.80[c]	99.36	98.59[c]
Power (μW)	0.83	0.33	2.7	2.21	0.034
Area (mm^2)	1.1	0.49	–	0.68	0.10

[a] Excludes counts for ventricular flutter in Record 207

[b] Estimated from Fig. 5 in [38]

[c] Simulation results only

4.2 OOK Transceiver

The transceiver SoC is fabricated with 0.13 μm CMOS technology. The chip core can operate from a 1.5 V button battery, and the die area is about 3.4 × 2.5 mm, including the testing buffer and IO pads. The die photo is shown in Fig. 21, which contains the LNA, CA, BA, ED, DCO, PA, digital baseband, and etc.

For 0.1 % BER, the receiver achieves a sensitivity of −65 dBm at a data rate of 10 Mb/s. The transient waveform is shown in Fig. 22, where the input RF signal is a −65 dBm carrier modulated by a 10 Mb/s random Tx signal. Note that the demodulated data is inverted with reference to Tx signal, so the Rx data are consistent with the Tx data in Fig. 22. The average power of receiver is 2.14 mW, which results in an energy efficiency of 0.214 nJ/bit.

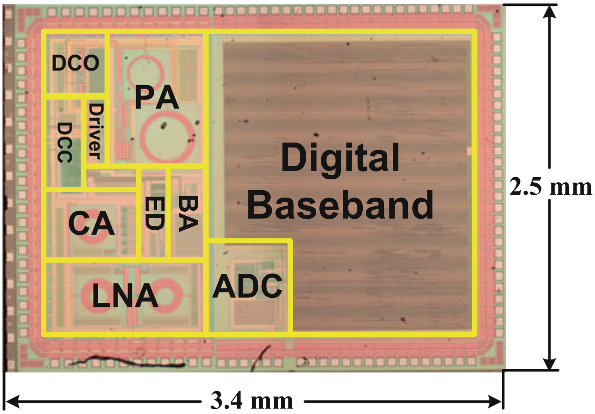


Fig. 21 Die photo

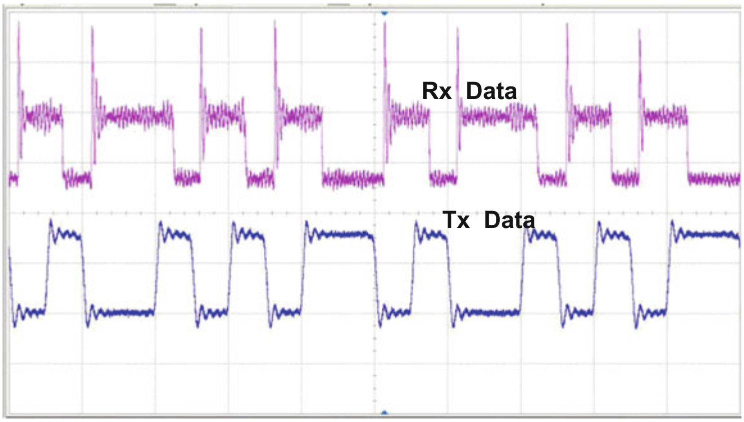


Fig. 22 Waveforms of Tx and Rx data

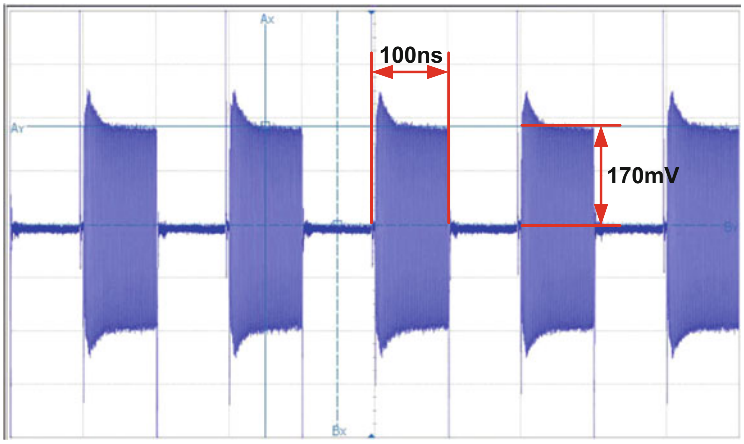


Fig. 23 Tx output signal

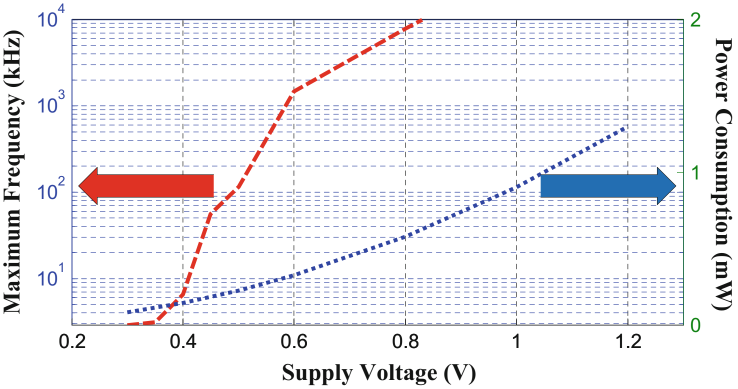


Fig. 24 Power consumption of digital baseband

The Tx output signal is shown in Fig. 23. The amplitude is about 170 mV and the settling time is less than 20 ns, indicating that the output power is −5.4 dBm. At 10 Mb/s data rate, the power consumption of transmitter is 2.85 mW, resulting the Tx energy efficiency of 0.285 nJ/bit at a data rate of 10 Mb/s.

The power performance of NTD digital part is also measured. The maximum working frequency and power consumption are plotted in Fig. 24. The energy efficiency is 130 pJ/bit under normal 1.2 V supply voltage, whereas the maximal energy efficiency reaches 34.8 pJ/bit at 0.55 V. In addition, the digital baseband supports a low-speed mode, when the power consumption can be reduced to 30 µW by lowering the operating voltage to 0.4 V.

Table 4 Performance comparison

		JSSC'07 [8]	ESSCIRC'11 [41]	BioCAS'11 [9]	ISSCC'11 [10]	This Work
Common	Type	Only RF	Only RF	Only RF	SoC	SoC
	Process	180 nm	180 nm	180 nm	90 nm	130 nm
	Carrier	916.5 MHz	425 MHz	406 MHz	2.4 GHz	2 GHz
	Data rate	1 Mb/s	2 Mb/s	2 Mb/s	5 Mb/s	10 Mb/s
Rx	Sensitivity	−37 dBm	−80.2 dBm	−45 dBm	−75 dBm	−65 dBm
	Power	0.5 mW	3.4 mW	0.59 mW	1.815 mW	2.14 mW
	nJ/bit	0.5	1.7	0.295	0.363	0.214
Tx	Output	−2.2 dBm	−2.17 dBm	−17 dBm	0 dBm	−5.4 dBm
	Power	3.8 mW	3.1 mW	0.84 mW	3.68 mW	2.85 mW
	nJ/bit	3.8	1.55	0.42	0.736	0.285

The key specifications of the SoC are summarized in Table 4, and compared with some typical OOK chips in recent citations. Our transceiver consumes the lowest energy per bit at both Rx and Tx modes. In addition, our chip is implemented without additional off-chip components.

5 Conclusions

This chapter presents two ultra-low-power systems for the wearable wireless ECG sensor, i.e., a A2I system with Real-time QRS detection and a fully integrated OOK transceiver SoC. Low speed requirements of the event-driven circuit allow the use of a low supply voltage in order to reduce power. Low power and good accuracy are achieved through the pulse-triggered and time-assisted pulse-triggered QRS algorithms that directly utilize the information embedded in level-crossing events to identify QRS complex. The algorithms are verified through simulations using signals from MIT-BIH ECG Arrhythmia Database. Implemented in 0.13 μm CMOS technology, the A2I-QRS chip consumes 220 nW under 0.3 V supply voltage for typical ECG input, which demonstrates the potential of using A2I system in ultra-low-power designs for wearable biomedical applications.

Next, the design of an energy efficient 10 Mb/s OOK transceiver SoC for WBAN applications is discussed. The problem of poor sensitivity of conventional ED based OOK receiver is resolved by a novel supply isolation scheme, whereas the low power advantage of ED based structure is maintained, i.e. 0.214 nJ/bit at Rx mode. By using bond wires as DCO inductors, the energy efficiency of transmitter is reduced to 0.285 nJ/bit at output power of −5.4 dBm. The energy per bit of digital baseband has been improved to nearly 1/4 when compared to the super-threshold operation, which also makes the design a perfect candidate for low-power biomedical sensor application.

References

1. Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Hailpern SM, Heit JA, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Virani SS, Wong ND, Woo D, Turner MB. Heart disease and stroke statistics - 2012 update. Circulation. 2012;125(1):e2–e220.
2. Zou X, Xu X, Yao L, Lian Y. A 1-V 450-nW fully integrated programmable biomedical sensor interface chip. IEEE J Solid-State Circuits. 2009;44(4):1067–77.
3. Jocke SC, Bolus JF, Wooters SN, Jurik AD, Weaver AC, Blalock TN, Calhoun BH. A 2.6-μ W sub-threshold mixed-signal ECG SoC. In: Proceedings of IEEE Symposium on VLSI Circuits; 2009. pp. 60–1.
4. Deepu CJ, Xu XY, Zou XD, Yao LB, Lian Y. An ECG-on-chip for wearable cardiac monitoring devices. In: Proceedings of Fifth IEEE International Symposium on Electronic Design, Test and Application DELTA '10; 2010. pp. 225–8.
5. Yazicioglu RF, Kim S, Torfs T, Kim H, Van Hoof C. A 30 μW analog signal processor ASIC for portable biopotential signal monitoring. IEEE J Solid-State Circuits. 2010;46(1):209–23.
6. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–20.
7. Kurchuk M, Tsividis Y. Signal-dependent variable-resolution clockless A/D conversion with application to continuous-time digital signal processing. IEEE Trans Circuits Syst I. 2010;57(5):982–91.
8. Daly D, Chandrakasan A. An energy-efficient OOK transceiver for wireless sensor networks. IEEE J Solid-State Circuits. 2007;42(5):1003–11.
9. Liu X, Demosthenous A, Jiang D, Vanhoestenberghe A, Donaldson N. A stimulator asic with capability of neural recording during inter-phase delay. In: Proceedings of the ESSCIRC (ESSCIRC), 2011 ; Sept 2011. pp. 215–18.
10. Vidojkovic M, Huang X, Harpe P, Rampu S, Zhou C, Huang L, Imamura K, Busze B, Bouwens F, Konijnenburg M, Santana J, Breeschoten A, Huisken J, Dolmans G, De Groot H. A 2.4 GHz ULP OOK single-chip transceiver for healthcare applications. In: IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC); 2011. pp. 458–60.
11. Tsividis Y. Event-driven data acquisition and continuous-time digital signal processing. In: Proceedings of IEEE Custom Integrated Circuits Conference; 2010. pp. 1–8.
12. Inose H, Aoki T, Watanabe K. Asynchronous delta-modulation system. Electron Lett. 1966;2(3):95–6.
13. Wang Y, Deepu CJ, Lian Y, A computationally efficient QRS detection algorithm for wearable ECG sensors. In: Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2011. pp. 5641–4.
14. Li C, Zheng C, Tai C. Detection of ECG characteristic points using wavelet transforms. IEEE Trans Biomed Eng. 1995;42(1):21–8.
15. Agarwal R, Sonkusale SR. Input-feature correlated asynchronous analog to information converter for ECG monitoring. IEEE Trans Biomed Circuits Syst. 2011;5(5):459–67.
16. Afonso V, Tompkins W, Nguyen T, Luo S. ECG beat detection using filter banks. IEEE Trans Biomed Eng. 1999;46(2):192–202.
17. Poli R, Cagnoni S, Valli G. Genetic design of optimum linear and nonlinear QRS detectors. IEEE Trans Biomed Eng. 1995;42(11):1137–41.
18. Trahanias P. An approach to QRS complex detection using mathematical morphology. IEEE Trans Biomed Eng. 1993;40(2):201–5.
19. Sayiner N, Sorensen H, Viswanathan T. A level-crossing sampling scheme for A/D conversion. IEEE Trans Circuits Syst II. 1996;43(4):335–9.

20. Schell B, Tsividis Y. Analysis of continuous-time digital signal processors. In: Proceedings of IEEE International Symposium on Circuits and Systems; May 2007. pp. 2232–5.
21. Balasubramanian V, Heragu A, Enz C. Analysis of ultralow-power asynchronous ADCs. In: Proceedings of IEEE International Symposium on Circuits and Systems; 2010. pp. 3593–6.
22. Bazes M. Two novel fully complementary self-biased CMOS differential amplifiers. IEEE J Solid-State Circuits. 1991;26(2):165–8.
23. Chatterjee S, Tsividis Y, Kinget P. 0.5-V analog circuit techniques and their application in OTA and filter design. IEEE J Solid-State Circuits. 2005;40(12):2373–87.
24. Abo AM. Design for reliability of low-voltage, switched-capacitor circuits. Ph.D. dissertation. Berkeley: University of California; 1999.
25. Schell B, Tsividis Y. A continuous-time ADC/DSP/DAC system with no clock and with activity-dependent power dissipation. IEEE J Solid-State Circuits. 2008;43(11):2472–81.
26. Kurchuk M, Tsividis Y. Energy-efficient asynchronous delay element with wide controllability. In: Proceedings of IEEE International Symposium on Circuits and Systems; 2010. pp. 3837–40.
27. Kim G, Kim M-K, Chang B-S, Kim W. A low-voltage, low-power CMOS delay element. IEEE J Solid-State Circuits. 1996;31(7):966–71.
28. Jeong C, Qu W, Sun Y, Yoon D, Han SK, Lee SG. A 1.5V, 140 μA CMOS ultra-low power common-gate LNA. In: IEEE Radio Frequency Integrated Circuits Symposium (RFIC); 2011. pp. 1–4.
29. Feng L, Mao Y, Cheng Y. An efficient and stable power management circuit with high output energy for wireless powering capsule endoscopy. In: IEEE Asian Solid State Circuits Conference (A-SSCC); 2011. pp. 229–32.
30. Maxim A, Poorfard R, Johnson R, Crawley P, Kao J, Dong Z, Chennam M, Nutt T, Trager D, Reid M.A fully integrated 0.13 μm CMOS Low-IF DBS satellite tuner using automatic signal-path gain and bandwidth calibration. IEEE J Solid-State Circuits. 2007;42(4):897–921.
31. Zhao B, Lian Y, Yang H. A low-power fast-settling bond-wire frequency synthesizer with a dynamic-bandwidth scheme. IEEE Trans Circuits Syst Regul Pap. 2013;60(5):1188–99.
32. Steele R. Delta modulation systems. London:Pentech Press; 1975.
33. Trakimas M, Sonkusale SR. An adaptive resolution asynchronous ADC architecture for data compression in energy constrained sensing applications. IEEE Trans Circuits Syst I. 2011;58(5):921–34.
34. Weltin-Wu C, Tsividis Y. An event-driven, alias-free ADC with signal-dependent resolution. In: Proceedings of IEEE Symposium on VLSI Circuits; 2012. pp. 28–9.
35. Tang W, Osman A, Kim D, Goldstein B, Huang C, Martini B, Pieribone VA, Culurciello E. Continuous time level crossing sampling adc for bio-potential recording systems. IEEE Trans Circuits Syst I. 2013;60(6):1407–18.
36. Li Y, Zhao D, Serdijn W. A sub-microwatt asynchronous level-crossing adc for biomedical applications. IEEE Trans Biomed Circuits Syst. 2013;7(2):149–57.
37. Ieong C-I, Mak P-I, Lam C-P, Dong C, Vai M-I, Mak P-U, Pun S-H, Wan F, Martins RP. A 0.83-μ W QRS detection processor using quadratic spline wavelet transform for wireless ECG acquisition in 0.35-μm CMOS. IEEE Trans Biomed Circuits Syst. 2012;6(6):586–95.
38. Abdallah R, Shanbhag N. A 14.5 fJ/cycle/k-gate, 0.33 V ECG processor in 45nm CMOS using statistical error compensation. In: Proceedings of IEEE Custom Integrated Circuits Conference; Sept 2012. pp. 1–4.
39. Zhang F, Lian Y. QRS detection based on multiscale mathematical morphology for wearable ECG devices in body area networks. IEEE Trans Biomed Circuits Syst. 2009;3(4):220–8.
40. Wang H-M, Lai Y-L, Hou M, Lin S-H, Yen B, Huang Y-C, Chou L-C, Hsu S-Y, Huang S-C, Jan M-Y. A ±6 ms-accuracy, 0.68 mm^2 and 2.21 μW QRS detection ASIC. In: Proceedings of IEEE International Symposium on Circuits and Systems; 2010. pp. 1372–75.
41. Liu J, Li C, Chen L, Xiao Y, Wang J, Liao H, Huang R. An ultra-low power 400 MHz OOK transceiver for medical implanted applications. In: European Solid-State Circuis Conference (ESSCIRC); 2011. pp. 175–8.

A Sensor-Fusion Solution for Mobile Health-Care Applications

Chen-Yi Lee, Kelvin Yi-Tse Lai, and Shu-Yu Hsu

Abstract As Internet of Things (IoT) and wearable devices become the next wave of emerging markets, it is very necessary to provide a set of sensors related to target applications so that decision-making becomes more reliable from collected data. However multi-sensor approach results in higher data rate and larger power consumption, which may not be accepted by application requirements, e.g. mobile health-care. In this paper, a sensor-fusion approach will be introduced to provide an energy-efficient and data-reliable solution. By exploiting event-driven architecture, energy efficiency can be enhanced; on the other hand, analysis accuracy can be further improved with the support of multi-data sets. As a result, both power consumption and data bandwidth can be minimized with better accuracy to meet those specifications in battery-operated devices. A test vehicle related to mobile health-care applications will also be introduced to demonstrate the feasibility of our proposal.

1 Introduction

Smart sensors have been highly demanded in traditional 3C (computer, communication, and consumer electronics) products and emerging medical, green-energy, and vehicle applications for better life with improved service quality. These sensors are deployed to sense the physical/chemical change volumes, which are then converted to digital codes for follow-up data processing. In the past, research in smart sensors has been concentrated on the improvement of sensors' quality, such as sensitivity, dynamic range, response time, . . . etc. Recently due to the progress of CMOS-MEMS technology, it is possible to integrate both sensor devices and readout circuits for a better form-factor to meet those system specifications in mobile devices. Moreover, power consumption can be further optimized to improve energy-per-bit efficiency by taking into account the behavior of target sensors.

C.-Y. Lee (✉) • K.Y.-T. Lai • S.-Y. Hsu
Department of Electronics Engineering and Institute of Electronics,
National Chiao Tung University, Hsinchu, Taiwan
e-mail: cylee@faculty.nctu.edu.tw

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_15

Very often an analog-to-digital converter (ADC) is exploited as part of the readout circuits to meet this conversion purpose, where its sampling rate depends on both sensor's characteristics and application requirements. Related works on low-power ADC for sensors can be found in [1, 2]. Recently time-to-digital converter (TDC) has been exploited as readout circuits by taking into account both sensing behavior and system requirements [3–5]. The advantages of this TDC approach are: (1) event-driven approach when sensor is activated; (2) portable solution which can be ported to different technology nodes; (3) adjustable output resolution which can be optimally tuned to meet system requirements. As a result, this TDC-based approach provides an area and power efficient solution for smart sensors.

This single-sensor, e.g. motion or accelerator sensor, together with various application software packages (so-called APP's) has created a lot of new applications and services in hand-held devices. For mobile health-care monitoring, Hsu et al. has proposed an ECG-SoC solution [6] with machine learning capability for heart-disease detection, where accuracy can be greatly improved by investigating different ECG features. However as application toward intelligent services, it's very necessary to provide more data sets for better analysis accuracy. As a result, multi-sensor (or so-called sensor-fusion) solutions are requested to meet the demands of these emerging applications. But this multi-sensor approach also demands higher power consumption and data bandwidth, which may not be accepted in bandwidth/power-constrained devices. To overcome the above-mentioned issue, an improved master-slave event-driven architecture is proposed in this paper, where slave sensors and corresponding data sets will be activated and collected only when master sensor requests. This sensor fusion concept has been verified on a mobile health-care platform for heart-disease detection. Simulation results show that overall power consumption can be further reduced even with three sensors (ECG, Motion, Respiration), while achieving better analysis accuracy than that from single ECG sensor platform.

The rest of this paper is organized as follows. In Sect. 2, we'll review the concept of event-driven architecture for single sensor platform. Both periodic and non-periodic sensed data sets will be investigated to see how the event-driven approach can be exploited to save both power consumption and data bandwidth. Then a master-slave (MS) event-driven architecture will be proposed to handle multiple data sets generated by different sensors in Sect. 3. In Sect. 4, we'll show a mobile health-care platform for heart-disease detection as a test vehicle to demonstrate the unique features of our sensor-fusion proposal.

2 Review of the Proposed Event-Driven Architecture

The goal of event-driven architecture is to derive an energy-efficient solution for realizing smart sensor designs. In other words, the required energy-per-bit should be minimized during data generation. Very often, optimization at system architecture level can reach better power/energy efficiency compared to those at lower circuit

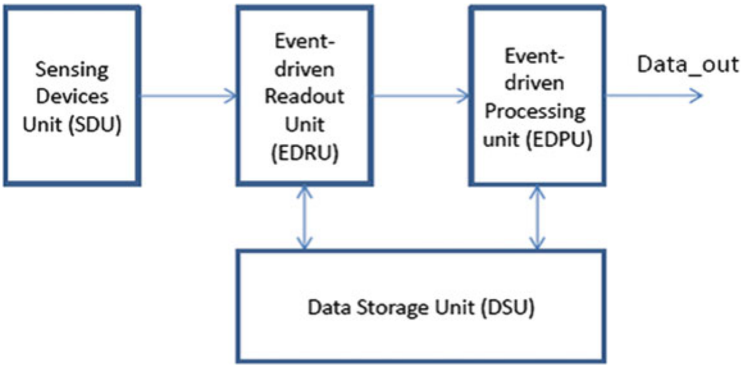

Sensing Devices Unit (SDU)
Event-driven Readout Unit (EDRU)
Event-driven Processing unit (EDPU)
Data_out
Data Storage Unit (DSU)

Fig. 1 An event-driven architecture for smart sensors to achieve better energy-efficiency

and process technology levels [7]. However the design cycle may get longer due to trade-off among large exploration spaces. Thus it is very necessary to derive an application-specific architecture to shorten the design phase at system-level design.

For sensor-based applications, an event-driven approach is proposed as shown in Fig. 1. This proposed architecture mainly consists of the four major blocks: the sensing devices units (SDU) designed to convert physical/chemical parameter changes into passive components, such as capacitance (C) and resistance (R); the event-driven readout unit (EDRU) is designed to detect the volume changes of these passive components and activate the follow-up operations whenever the detected volume is over a threshold; event-driven processing unit (EDPU) is designed to deal with the event-related sensed data for status/syndrome detection so that only significant sensed data or related features will be further processed; data storage unit (DSU) is designed to buffer sensed data and features which are demanded by both EDRU and EDPU. The required power consumption model is simplified as follows:

$$\mathrm{P_T} = (1 - P_{ed}) * \mathrm{P_s} + P_{ed} * \mathrm{P_a}, \ldots \tag{1}$$

where $\mathrm{P_T}$, $\mathrm{P_s}$, and $\mathrm{P_a}$ stand for total power, static power, and active power of these four modules; P_{ed} is the probability to generate effective events when sensor is activated.

In practical designs, $\mathrm{P_a}$ can be much larger than $\mathrm{P_s}$ because many low-power design techniques can be applied to reduce static power. Thus if P_{ed} is lower, the total power $\mathrm{P_T}$ becomes less as well. This also implies that only effective data will be transmitted to save network bandwidth.

Figure 2 shows a possible solution to realize the EDRU for a capacitive-type sensor device with differential structure. It can be found that the differential outputs from sensing device are connected to C2T to generate a time-pulse, which is mainly for non-periodic data generation sources. The duration of this time pulse is corresponding to the capacitance difference (C+ and C−) and hence the humidity level in this example.

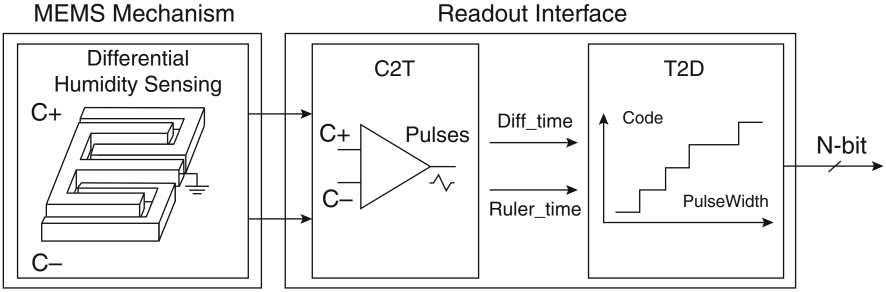


Fig. 2 The proposed event-driven readout circuit for capacitive-type MEMS sensors

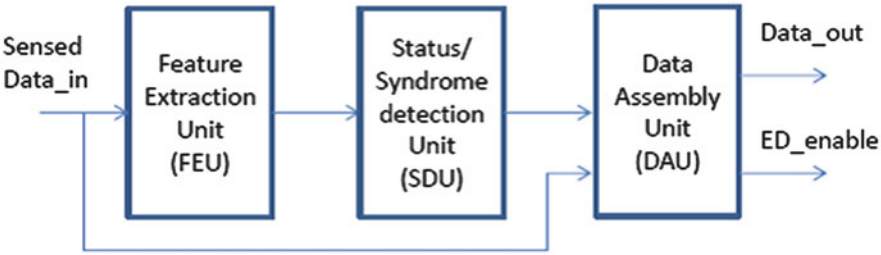


Fig. 3 Event-driven processing based on status/syndrome detection for periodic data sources

Note that there are two outputs from C2T serving as volume detection (Diff_time) and calibration (Ruler_time) respectively. This Diff_time signal is also served as an event indicator to enable the follow-up code conversion whenever over a pre-defined threshold. The last stage, T2D, is for code generation and its output resolution can be dynamically adjusted to meet various system requirements under different operation modes. This proposal has been successfully integrated into both accelerator and humidity sensors [5, 8] with significant power saving.

Figure 3 shows a block diagram for the EDPU, which is designed to detect certain status/syndromes from periodic data sources so that only those sensed data with significant values need to be further processed. Thus a feature extraction unit (FEU) is first included. Then a status/syndrome detection unit (SDU) is designed to compare those extracted features with off-line training ones. Finally data assembly unit (DAU) is included to packetize data (sensed data and features) as output for follow-up data processing. Note that an event-driven indicator "ED_enable" is included as well. Since event-driven concepts are applied to both EDRU and EDPU, both clock-gating and data-gating schemes can be exploited to reduce dynamic power. In the meantime, dynamic body biasing (DBB) and power-gated cells can be applied as well to further reduce static power. A case study with this concept for ECG-based mobile applications can be found in [6, 9].

3 Sensor-Fusion Architecture

In some applications, single-sensor approach may not provide the sufficient data sets for follow-up decision-making. For example, heart rate (HR) can be detected from the received ECG data by detecting R-peaks [9]. However HR is also highly related to body motion when ECG signals are measured. With the support of motion sensor, it will be easy to monitor users' health condition in a more effective and correct way. This becomes more important as health-care towards off-hospital services (or so-called mobile health-care service concept), where only bio signals will be received and detected by medical teams and service centers.

To meet this application trend, both form factor and power consumption have to be further optimized to meet mobile requirements. However due to the demand of more sensors to provide extra data sets, very often both form factor and power consumption will be increased as well. Although form factor can be improved by using the advanced technology, e.g. CMOS-MEMS, there remains a challenge in reducing power consumption to meet battery-powered requirements of mobile devices. As a result, a sensor-fusion architecture is proposed in Fig. 4, which is derived from an event-driven proposal for single-sensor as shown in Fig. 1. Note that both architectural proposals are very similar except some minor functions are modified to cover more sensors. Also a wireless interface unit (WIU) is included to provide a link for data transmission and command assignment between service center and wireless sensing node.

The proposed sensor-fusion architecture consists of a master-slave sensor-set. A master-sensor is selected and designed to generate the required signals to be detected, while the other slave sensors are included to provide support data sets

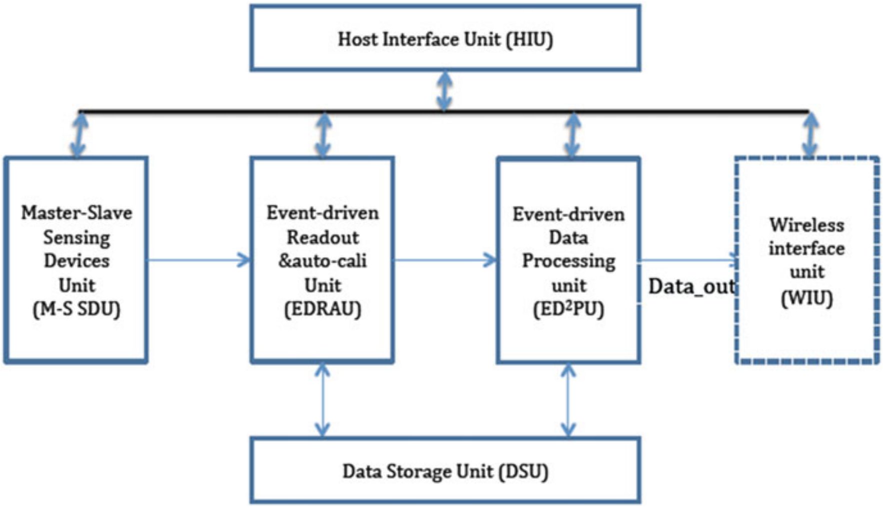

Host Interface Unit (HIU)
Master-Slave Sensing Devices Unit (M-S SDU)
Event-driven Readout &auto-cali Unit (EDRAU)
Event-driven Data Processing unit (ED²PU)
Data_out
Wireless interface unit (WIU)
Data Storage Unit (DSU)

Fig. 4 The proposed sensor fusion architecture, where wireless interface is included

whenever requested. With this architectural model, the total power can be estimated as follows:

$$\mathrm{P_{TS}} = (1 - P_{eds}) * \mathrm{P_s} + P_{eds} * \mathrm{P_a} + \mathrm{P_{od}} \ldots \tag{2}$$

Note that $\mathrm{P_{TS}}$, $\mathrm{P_s}$, $\mathrm{P_a}$ and $\mathrm{P_{od}}$ stand for total power, static power, active power and power overhead of the sensor-fusion solution respectively; P_{eds} is the probability to generate effective events when sensor is activated. Here P_{eds} is derived from the following:

$$P_{eds} = P_{ed} \ * \ P_{su} \tag{3}$$

where P_{su} is the conditional probability from slave sensors to support master sensor and meets the following condition: $0 < P_{su} <1$. Since P_{eds} is always less than P_{ed}, it is found the total power from this sensor-fusion approach can be further reduced if P_{od} is controlled well.

4 An Event-Driven Sensor-Fusion Solution for Mobile Health-Care Applications

To demonstrate the feasibility of our proposed sensor-fusion architecture, a demo-platform based on ECG signal for heart disease detection has been set up as shown in Fig. 5. An ECG-SoC for mobile health-care application has been selected as the master sensor. This chip is designed with both feature extraction and auto classification based on machine learning mechanism, in addition to analog front-end sensing circuits. Also low-power design techniques have been exploited to reduce power consumption and enhance energy efficiency when each event is detected. A block diagram for this ECG-SoC and the corresponding chip die-photo in 90 nm CMOS technology is shown in Figs. 6 and 7 respectively [9].

To enhance detection accuracy, two slave sensors are included to provide support data sets. A 3-D motion sensor is exploited to provide “motion” status, where event-driven readout circuit is exploited to save power consumption. In other words, only motion event is detected, the readout circuit is activated to generate digital output codes based on the capacitance change volume. The chip die photo is shown in Fig. 8, where sensing devices occupy most of the chip area [5]. Another slave sensor is a humidity-based breathing sensor, designed for respiration monitoring. The humidity level causes the capacitance change volume, which will be detected by event-driven readout circuit. Figure 9 shows the die photo of this humidity sensor, where sensing devices occupies most of the chip area [8]. Note that the power consumption of both slave sensors is highly related to those events causing capacitance volume change. The static power can be under sub-uW if calibration is done well.

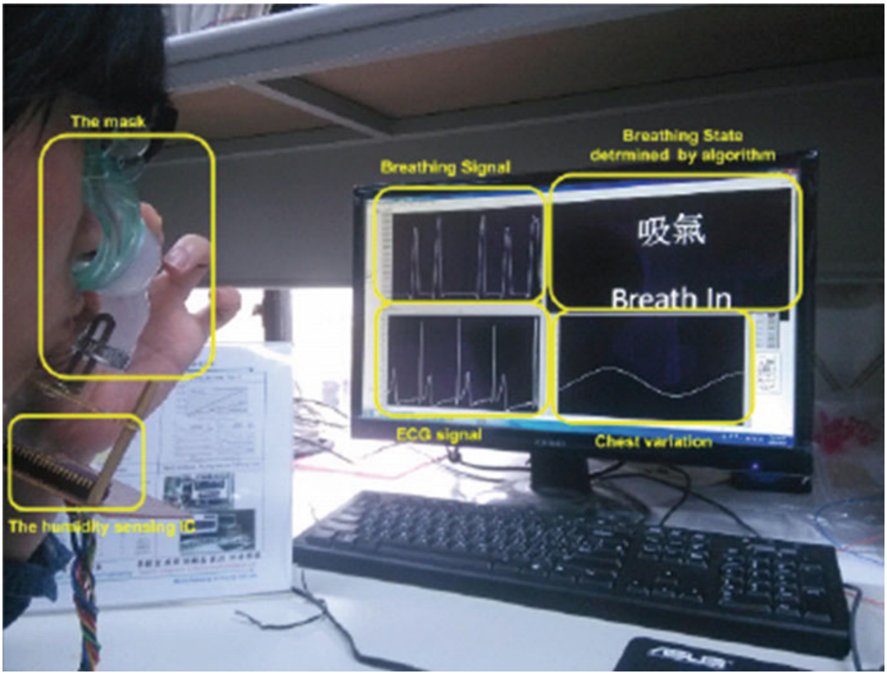


Fig. 5 A sensor-fusion solution for mobile healthcare applications, where ECG unit is the master sensor, motion and breathing units are slave sensors to generate support data sets

Those classified patterns related to health-care will be identified by medical teams and medical service centers. This provides a real-time health-care service when the sensor-fusion device is activated by end-users. Also note that there are more than 70 features extracted from the received ECG signals. As a result, this platform can be exploited for health-care monitoring in mobile conditions. Below we'll use two cases as shown in Fig. 10 to illustrate the sensor-fusion concept, especially those related to reduce power consumption.

4.1 Case 1: Basic Mode for Heart Rate (HR) Estimation

This mode is defined as the basic requirement when sensor node is activated. If the estimated HR is within a certain range defined as "health", the other two slave sensors are in idle state. On the other hand, the motion sensor will be activated only when HR falls into a status to be double confirmed. This implies that both $\boldsymbol{P_{su}}$ and $\mathbf{P_{od}}$ have to be calculated to find the total power for this mode. Assume the following:

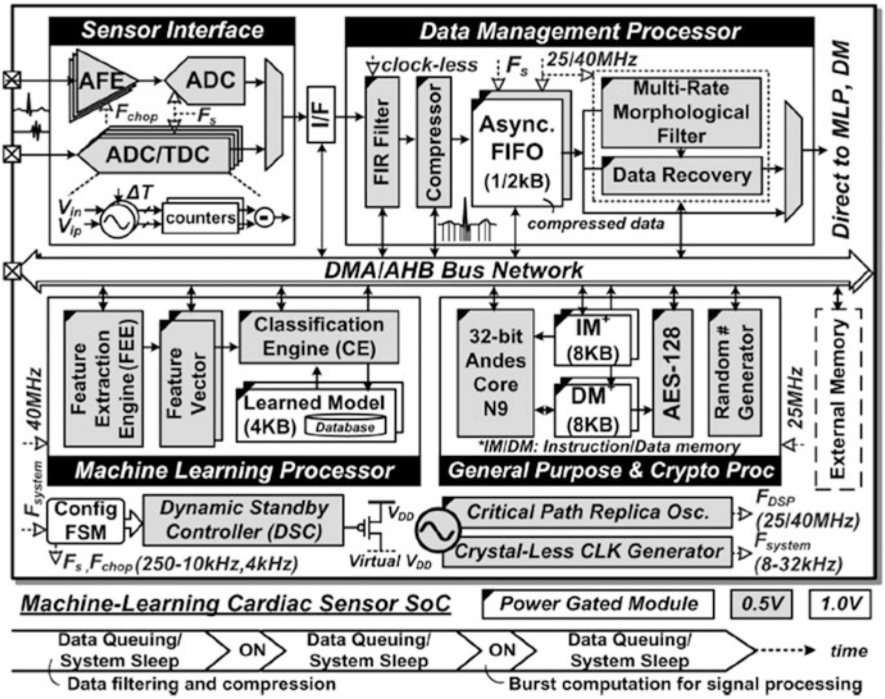

Sensor Interface
AFE
ADC
F_{chop}
F_s
ADC/TDC
ΔT
V_{in}
V_{ip}
counters
I/F
Data Management Processor
clock-less
F_s
25/40MHz
FIR Filter
Compressor
Async. FIFO (1/2kB)
Multi-Rate Morphological Filter
Data Recovery
compressed data
Direct to MLP, DM
DMA/AHB Bus Network
Feature Extraction Engine (FEE)
Feature Vector
Classification Engine (CE)
Learned Model (4KB)
Database
40MHz
F_{system}
Machine Learning Processor
32-bit Andes Core N9
IM* (8KB)
DM* (8KB)
AES-128
Random # Generator
25MHz
*IM/DM: Instruction/Data memory
General Purpose & Crypto Proc
External Memory
Config FSM
Dynamic Standby Controller (DSC)
V_{DD}
Virtual V_{DD}
Critical Path Replica Osc.
Crystal-Less CLK Generator
F_{DSP} (25/40MHz)
F_{system} (8-32kHz)
F_s, F_{chop} (250-10kHz,4kHz)
Machine-Learning Cardiac Sensor SoC
Power Gated Module
0.5V
1.0V
Data Queuing/ System Sleep
ON
Data Queuing/ System Sleep
ON
Data Queuing/ System Sleep
time
Data filtering and compression
Burst computation for signal processing

Fig. 6 Block diagram of the ECG-SoC for mobile health-care applications [9]

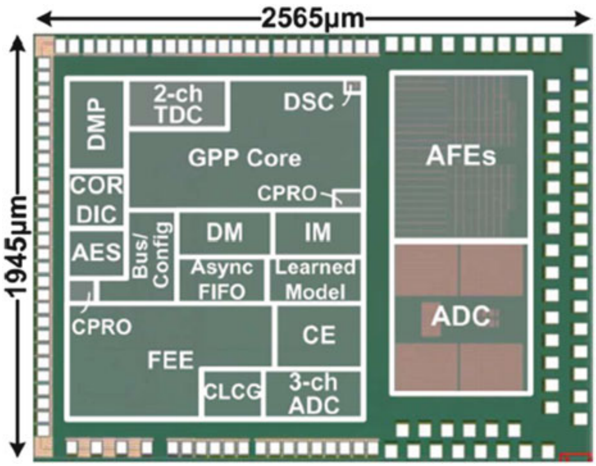


Fig. 7 Die photo of the ECG-SoC chip in 90 nm CMOS process

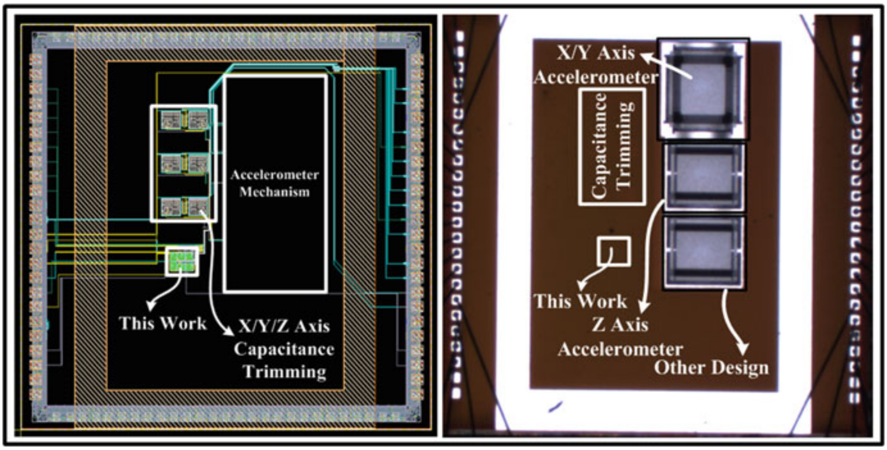
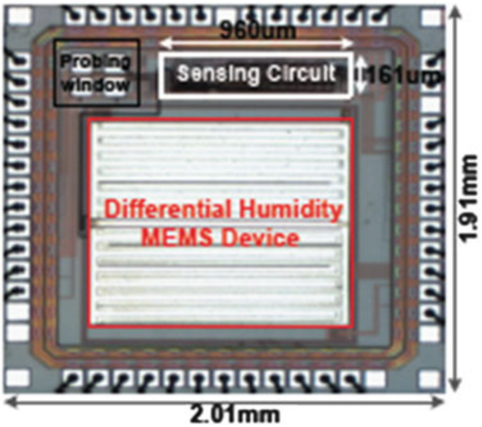


Fig. 8 Die photo of the 3-D motion sensor chip, where "This Work" implies the readout circuit whose area is much less than that of MEMS-based sensing devices

Fig. 9 Die photo of the humidity sensor chip fabricated in 0.35 um CMOS



$$\mathrm{P_{od}} = 0.1\mathrm{P_a};\ \mathrm{P_s} = 0.05\mathrm{P_a};\ P_{su} = 10\ \%;\ P_{ed} = 1;$$

the total power $\mathrm{P_{TS}}$ can be calculated as $\mathrm{P_{TS}} = (1-0.1)*0.05\mathrm{P_a} + 0.1\mathrm{P_a} + 0.1\mathrm{P_a} = 0.245\mathrm{P_a}$, which is less than that from single ECG sensor. Note that these data are from 90 nm CMOS technology and may have some differences if different technologies are considered.

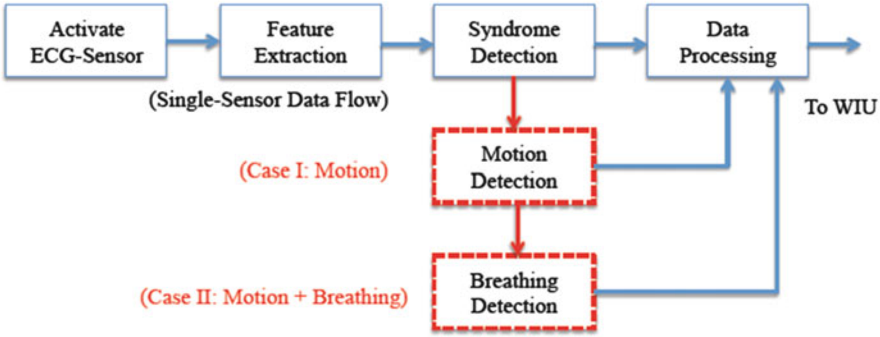


Fig. 10 Activating slave sensors when master sensor detects certain pre-defined syndromes

4.2 Case 2: Arrhythmia Detection

This mode is designed to detect specific heart disease based on cardiac syndrome derived from a set of selected features. In this case, the master ECG sensor is always activated. Both motion and breathing sensors are requested to provide support data sets. Again both $\boldsymbol{P_{su}}$ and $\mathbf{P_{od}}$ have to be calculated to find the total power for this mode. Assume the following:

$$\mathrm{P_{od}} = 0.3\mathrm{P_a};\ \mathrm{P_s} = 0.2\mathrm{P_a};\ P_{su} = 30\ \%;\ P_{ed} = 1;$$

the total power $\mathrm{P_{TS}}$ can be calculated as $\mathrm{P_{TS}} = (1-0.3)*0.2\mathrm{P_a} + 0.3\mathrm{P_a} + 0.3\mathrm{P_a} = 0.714\mathrm{P_a}$, which is again less than that from single ECG sensor.

Note that power overhead $\mathrm{P_{od}}$ results from slave sensors with data processing. However the probability P_{eds} to process and transmit these important data sets is reduced. Hence overall power consumption is reduced.

5 Conclusion

In this paper, we have first reviewed an event-driven architecture to enhance energy efficiency for single sensor designs and then introduced a sensor-fusion solution for mobile healthcare applications. By taking into account both data statistics and application-specific architecture, overall power/energy efficiency can be further improved. The proposal has been illustrated on a 3-sensor platform with ECG serving as the maser sensor. Simulation results from two cases demonstrated the feasibility of our proposal, especially in energy efficiency and analysis accuracy. The proposal can be applied to other wearable and Internet-of-Things (IoT) applications to achieve better energy efficiency in battery-powered devices.

Acknowledgements This work is supported by NSC of Taiwan under grant NSC 102-2218-E-009 -025. The authors would like to thank Chip Implementation Center of Taiwan for testing services and all members of NCTU/SI2 Lab for fruitful discussions.

References

1. Yazicioglu RF, et al. A 30μW analog signal processor ASIC for biomedical signal monitoring. In: ISSCC; 2010. p. 124–5.
2. Zou X, et al. A 1V 22μW 32-channel implantable EEG recording IC. In: ISSCC; 2010. p. 126–7.
3. Xia S, et al. A capacitance-to-digital converter for displacement sensing with 17b resolution and 20us conversion time. In: ISSCC; 2012. p. 198–200.
4. Freyman L, et al. A 346 μm^2 reference-free sensor interface for highly constrained microsystems in 28 nm CMOS. In: ASSCC-2013; Nov 11–13. Singapore; 2013. p. 105–108.
5. Lai KYT, et al. A 0.0354 mm^2 82 μW 125 KS/s 3-axis readout circuit for capacitive MEMS accelerometer. In: ASSCC-2013; Nov 11–13. Singapore; 2013. p. 109–112.
6. Hsu S-Y, et al. A 48.6-to-105.2 μW machine-learning assisted cardiac sensor SoC for mobile healthcare monitoring. In: Symposium on VLSI Circuits; 2013. p. 252–3.
7. De Man H. Ambient intelligence: gigascale dreams and nanoscale realities. In: ISSCC; 2005. p. 29–35.
8. Lai KYT, et al. A 3.3V 15.6b 6.1pJ/0.02%RH with 10 ms response humidity sensor for respiratory monitoring. In: ASSCC-2014; Nov 10–12. Kaoshiung, Taiwan; 2014. p. 293–296.
9. Hsu S-Y, et al. A 48.6-to-105.2 μW machine learning assisted cardiac sensor SoC for mobile healthcare applications. IEEE J Solid-State Circuits. 2014;49(4):801–11.

Part VI
Deployment and Service of Smart Sensors in the Society

An IoT Browsing System with Learning Capability

Wen-Tsuen Chen, Chih-Hang Wang, Yen-Ju Lai, and Po-Yu Chen

Abstract IoT browser provides a novel way for people to interact with objects through the Internet. Comparing with traditional web browsing environment, IoT browsing environment has some unique features such as the way of interacting with the objects, the importance of spatial-temporal information of objects, and the necessity of resource reuse. In this chapter, we propose a novel IoT browsing system with a learning capability middleware for IoT browser integrated with context-aware services. The system adopts a service-oriented architecture and provides device interoperability, resource reusability and spatial-temporal awareness. More specifically, the system can provide tailored services to meet the preferences of users by using flow-based programming to establish event flows on the IoT browser. Furthermore, learning is introduced to help the users build their event flows by suggesting them the next possible events based on the historical context information of sensors (or smart things). We develop a prototype of the proposed system and demonstrate the applicability of the system in a home browsing scenario. The prototype shows that with the help of the IoT browsing system, heterogeneous devices can cooperate with each other to provide IoT services in accordance with context inference results. In addition, each device can be reused by multiple services with minimum human supervision to reduce hardware and deployment costs.

1 Introduction

The Internet of Things (IoT) has attracted tremendous attentions recently. When all kinds of "smart things" (e.g., sensors, mobile devices, etc.) interconnected with each other via Internet, they will bring various novel applications or services into human's

W.-T. Chen (✉)
Department of Computer Science, National Tsing Hua University, 300 HsinChu, Taiwan

Institute of Information Science, Academia Sinica, Nankang, 115 Taipei, Taiwan
e-mail: wtchen@cs.nthu.edu.tw; chenwt@iis.sinica.edu.tw

C.-H. Wang • Y.-J. Lai • P.-Y. Chen
Department of Computer Science, National Tsing Hua University, 300 HsinChu, Taiwan
e-mail: s100062591@m100.nthu.edu.tw; clearsky0305@gmail.com; chenpy@mx.nthu.edu.tw

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_16

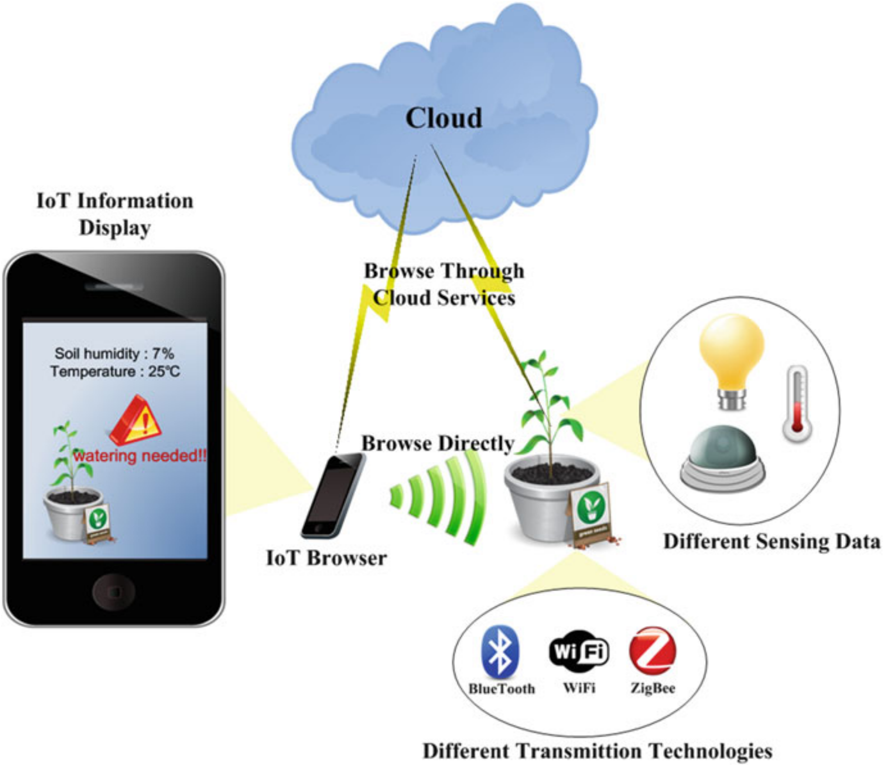


Fig. 1 An IoT browsing scenario. The information of the plant pot distributed in various sensors can be locally or remotely browsed through the IoT browser

daily life, environment monitoring, and healthcare in the near future. For using these applications and services, IoT browser is a convenient tool by which people are able to use smart phones or other client devices to browse not only digital entities but also physical objects [1]. It provides a means for people to easily explore the IoT and simplifies the operations for using the IoT services. A practical IoT browsing scenario illustrated in Fig. 1 [1] reveals some distinctive features in contrast with the traditional web browsing. In IoT, data sources are distributed in multitude of devices, and their spatial-temporal information is critical because people have to know where the devices are and when the sensing data is recorded. Besides, browsing of physical objects can be locally through a local area network or remotely through the cloud, and the information will be displayed in compliance with user’s rights and status.

To realize the IoT browsing environment with the above features, the following issues need to be considered. First, since various devices may be designed with different communication technologies and protocols, an interface(s) accounting for interoperability is important by which connected heterogeneous devices can exchange information and mutually control each other to work together for the IoT services [2, 3]. Furthermore, owing to the fact that multiple services often request the information from the same devices, resource reusability ought to be considered in IoT browsing. Without a mechanism to share the devices with other services, each service provider has to install more proprietary devices accompanying higher hardware and deployment cost. An efficient method to manage the relations between context data and spatial-temporal information is also required and should be done automatically for scalability. Finally, context awareness and access control should be addressed to make the browsing more intuitive and display the information in the most suitable way with respect to data types to the intended users. To sum up, in order to handle the above issues, a new system is needed to augment the traditional web browsing environment to support IoT browsing.

In this chapter, therefore, we propose an IoT browsing system with a learning capability middleware for IoT browser integrated with context-aware services. The system adopts a service-oriented architecture with a middleware to provide device interoperability, resource reusability and spatial-temporal awareness. More specifically, users can ask the system to provide services they prefer by using flow-based programming (FBP) to build their event flows on the IoT browser where each event represents a sensing result or an action of smart things and a chain of events comprises a service. Furthermore, learning is introduced to help the users build the event flows by suggesting them the next possible events based on the historical context information of sensors (smart things). The functions of the middleware are mainly located in the cloud, while the local gateways only collect different sensing information from heterogeneous devices (sensors or smart things); hence, the cost of the local gateways is reduced. In particular, the modules on the local gateways and the cloud jointly provide interoperable interfaces for one or a set of device implementations through the web service technologies and make concept of the device implementation details as an abstract service [3, 4]. Although web services help for integration of sensor systems (or smart things) with the Internet, this integration is inflexible and costly since it needs human supervisions for concept mapping between sensor systems and multiple applications. For this reason, Ontology [5], Resource Description Framework (RDF), and semantic annotation techniques are adopted to manage the context data and spatial-temporal information due to their advantages, e.g., better supporting for logical inference [6], and achieve automatic (or semi-automatic) sensor data collection and device registrations [7]. Furthermore, ontology can also provide domain concept management and data format transformation that exempt human from the tedious works of integration details in reusing devices among multiple services since they have the common set of concept about context. Benefitting from the existing logic reasoning mechanism and the concept sharing of ontology, context-awareness can be provided to make the browsing more intuitive [8].

A prototype of the proposed system for IoT browser is developed and tested in a home browsing scenario. The test scenario exhibits the importance of maintaining relations between context and spatial-temporal information in IoT browsing. With the help of the IoT browsing system, heterogeneous devices can cooperate with each other to provide services in accordance with context inference results, and be reused by multiple services to reduce hardware and deployment cost.

The remainder of this chapter is organized as follows. Section 2 reviews the related technologies and the works about web services, semantic technologies, context-aware systems, FBP, and learning. In Sects. 3 and 4, we describe the design of the proposed IoT browsing system and the implementation of a prototype with a test scenario respectively. Finally, we conclude this chapter and discuss the future works in Sect. 5.

2 Background and Related Works

Since IoT browsing is similar to traditional web browsing, some web technologies can be applied to our system architecture and middleware. In this section, we review the related web technologies, semantic technologies and previous works, and explain how they could improve IoT browsing experience.

2.1 IoT Browser

IoT browser provides a novel convenient approach to interacting with IoT devices. The work in [1] provided a sentient visor system prototype for browsing IoT and discussed some conceptual considerations for developing IoT browser such as contextual reasoning, service discovery, semantic communication, mixed-reality user interface and multimodal information. However, they did not describe the implementation details that take into account of these considerations.

2.2 Web Services

Since IoT browser is also a kind of web browser, many web technologies, such as web services and semantic web, can be utilized or extended for an IoT browser. Web services originally aim to provide the interoperability among applications and software built on different platforms [9]. In particular, web services provide network functions through web standards and protocols (e.g., HTTP, XML and SOAP). Regardless of differences in developing languages and platforms, any application can use the existing web services (i.e., the public network functions) with the web standards. Recently, web services are used in embedded systems to support interoperable machine-to-machine (M2M) interaction by exchanging messages through the web protocols and XML-based languages [4]. XML is a markup language that is both human and machine readable. The benefits of XML include simplicity, openness, extensibility, self-description and so on [10]. Today, XML is widely used for representing data structures in web services.

2.3 Sensor Web

Integrating IoT devices to the web for interoperability is called "Sensor Web" in the literatures [11, 12]. There are two main ways to implement web services in Sensor Web platform: Service-Oriented Architecture (SOA) and Resource-Oriented Architecture (ROA). Both architectures can be represented in different web service styles such as Simple Object Access Protocol (SOAP) or Representational State Transfer (REST) [13] and each has its own advantages. In the following, we discuss the characteristics of each architecture.

2.3.1 Service-Oriented Architecture

With SOA, a public interface is created for one or a set of device implementations, which makes the concept of device implementation details as an abstract service [3]. This abstract service can be accessed by applications and identified by a Uniform Resource Identifier (URI). Therefore, SOA has the advantages of service encapsulation, componentization, dynamicity, flexibility, etc. [14]. Web services are the most prominent technologies today to realize SOA [15] and Fig. 2 represents the mapping between web services and the major components of SOA [16]. Service Providers publish their services with the Service Broker, and Service Requesters find desired services by searching Service Broker's registry of published services, and then bind to them to consume the services.

2.3.2 Resource-Oriented Architecture

Some previous works apply ROA to tackle interoperability problem in the past few years [17, 18]. Applying ROA to integrate resources into the web is also called "Web of Things" (WoT) [19]. In ROA, all resources such as processes, data and devices can be seen as entities that provide some services. For example, in work [17], they assume that all the IoT devices support web service architecture by embedding tiny

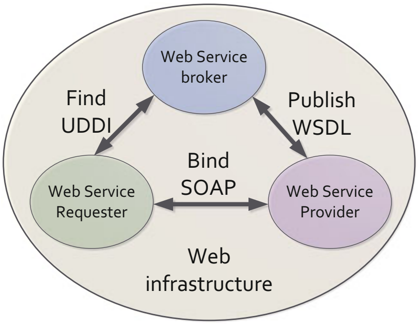


Fig. 2 The mapping between SOA and web services technologies. *UDDI* universal description, discovery and integration, *WSDL* web services description language, *SOAP* simple object access protocol

web servers. ROA applies the REST architectural style [13] that rather than using complex mechanisms such as SOAP, applications access and control IoT devices via HTTP methods (e.g., GET, POST, PUT and, DELETE) explicitly.

In summary, ROA is considered to be more scalable and more robust than the SOA architecture since it is a more distributed and simpler solution. However, the unnecessary cost of making entire smart things support web architecture is non-negligible [11]. In addition, it is not easy to realize complex use cases such as disaster management systems with ROA approaches since they need more detailed sensor information models and richer functionality on access, discovery, tasking and event handling [11]. In contrast, the SOA encapsulates device implementations into services and makes the implementations couple with each other better. Thus it can easily realize more complex use cases. We adopt the SOA approach to designing our IoT browsing system.

2.4 Semantic Technologies

With semantic interoperability technologies, beyond the ability of communicating and exchanging data, devices are capable of automatically interpreting and using the data. In addition, semantic technologies can be utilized to assist people in arranging their knowledge [20]. The main purposes of semantic technologies are easing the integration of resources (any objects, data or concepts), increasing the utility of information and helping information access/analysis [21]. Semantic technologies allow data to be defined and linked in such a way that it is machine-interpretable to achieve effective collaboration and reuse of data across applications. Such data are referred to as Linked Data [22, 23].

The main semantic technologies include Resource Description Framework (RDF), Web Ontology Language (OWL), Semantic Rule Language (SWRL) and SPARQL query language [24]. Ontology is an explicit specification of conceptualization [25] and capable of describing the concepts in a domain (e.g., mathematics, medical science, natural phenomena, etc.). It comprises classes (concepts), slots (roles or properties) and facets (role restrictions). Ontology together with individuals (instances) constructs a knowledge base [5], which is a particular kind of database for knowledge management, to represent data in the RDF graph data model rather than relational data model. Figure 3 is an example of RDF graph data model. It constructs tight relationship among resources to enable inference. The URLs shown in Fig. 3 are defined by OWL. SPARQL is an RDF query language to retrieve and manipulate data stored in RDF Triplestore (i.e., RDF graph database) as SQL for relational database. Finally, SWRL is used in logic reasoning or ontology matching.

In the IoT browsing scenario, semantic technologies combine virtual world with physical world and facilitate users to query/manage devices and data. They can relieve human efforts in managing data format, domain concepts and integration details. For example, SP-ACT [26] rewrites information from IoT devices in RDF

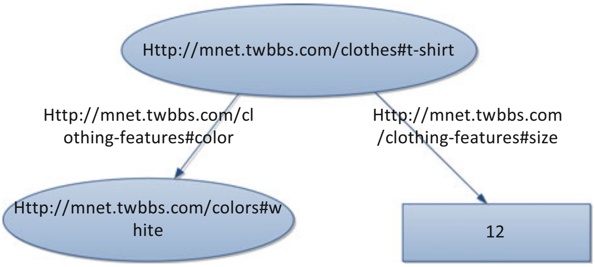

Http://mnet.twbbs.com/clothes#t-shirt
Http://mnet.twbbs.com/clothing-features#color
Http://mnet.twbbs.com/clothing-features#size
Http://mnet.twbbs.com/colors#white
12

Fig. 3 RDF graph data model

format, and adopts OWL reasoning paradigm and the execution of SPARQL rules to recognize human activities. With machine-interpretable data, automatic sensor data collecting, processing and reasoning can be achieved [27]. Moreover, Linked Data facilitate the IoT devices to find other devices and interlink with external IoT sources. The SPITFIRE project [28] is an example that uses RDF and SPARQL to search real world entities by their high-level meaning (e.g., searching for an "empty" meeting room rather than multiple sensors' sensing data).

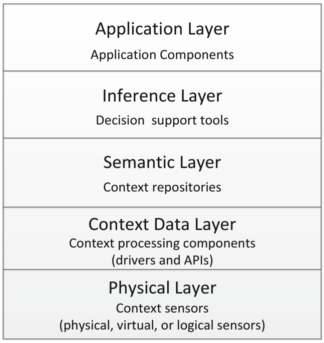
2.5 Context-Aware Services

Context is any information that can be used to characterize the situation of an entity [29], and context awareness means that the applications or services adapt their behaviors to the context variations [30]. The advantages of context-aware services are twofold: the first is reducing the unnecessary user operation loading. The applications can automatically collect the context (both user context and environment context) so that the users do not need to set or define the relevant information about them. The second advantage is that the applications can predict user expectation based on the inferred status or situations to change their behaviors or provide suggested services automatically.

For context-aware applications, context modeling is needed. Context model is an infrastructure that defines and stores context data in a machine interpretable form that facilitate the manipulation of context information. Many context modeling approaches have been proposed (e.g., Key-Value models), and ontology-based modeling, which models context with ontologies, is considered as one that can fulfill most of the requirements of context-aware computing [31]. Most context processing is designed in layered architecture that is generally separated into five layers as shown in Fig. 4 [32].

Physical layer contains various sensors, and the context data are extracted from the raw sensing data by the abstraction functions in the Context data layer. Semantic layer reasons the context data to higher level information that are stored for later use. Finally, Inference layer infers the real-time context information by rules (usually along with learning algorithms) to guess what user (people or machines)

Fig. 4 Layered structure of context-aware systems



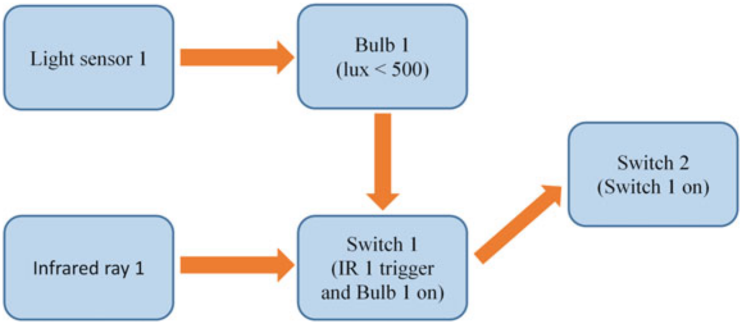


Fig. 5 An example of flow-based programming

status is, and forwards it to Application layer to provide suitable services. This architecture facilitates each layer to combine any inputs from lower layers to make the best decisions. For example, in [30], they design an ontology-based module-layer framework that each module layer can use the inference information from the lower layer to enable more accurate context inference.

2.6 Flow-Based Interface

In IoT scenario, one may combine different sensors or smart things to provide a specific service. For example, one may use CO_2 sensors and temperature sensors to determine whether there is a fire and may use the spatial and temporal information of sensors to determine the accident location. In the example shown in Fig. 5, the relation between each device is critical to determine the correctness of service. Flow-based programming (FBP) [33] defines services as networks of "blocks" that exchange information across predefined connections by message passing. FBP can change the relation between blocks by reconnecting the blocks. In Fig. 5, each "block" represents an "event" where each event is a sensing result or an action of smart things and the connection between two blocks represents the relation between two events. The starting point of an arrow is the precondition that needs to be satisfied when the next event is triggered and the end of the arrow is the corresponding action. If all preconditions of an event are true, the corresponding action of the event will be taken. In this example, when the lux detected from light sensor is lower than 500, the bulb1 will be opened. Then, switch 1 turns on automatically if both bulb1 is turned on and infrared ray (IR) 1 is triggered, and switch 2 turns on after switch 1 turns on. There are many previous works that exploit the flexibility of FBP to design their systems. For example, the work in [34] applies FBP for anti-steal system and [35] employs it to develop smart home applications.

2.7 Learning

Multiple works have designed event recognition and discovering mechanism based on the longest common subsequence algorithm [36, 37] to determine the occurrence of each event or service in the future time. Rashidi et al. [36] proposes a method to discover and recognize frequent activities that naturally occur in an individual routine. The method is improved in [37] to better handle real life data by dealing with different occurrence frequencies. In most environment, however, one may prefer only the next occurred event than a long sequence of events and wishes the system can serve the user automatically (e.g. after going home, the user may want to watch TV and wish the system can turn on TV automatically). Therefore, we design a learning algorithm to discover the patterns of events and to suggest users the next possible occurred event based on historical context information of sensors (or smart things).

3 System Design

In this chapter, we propose an SOA system with a middleware solution to address interoperability, context management and reasoning issues in IoT browsing environment. In this section, we overview the system architecture and the basic functions of the middleware, and then explain how the proposed system manages the spatial-temporal information and provides learning schemes to enhance the browsing operations.

3.1 System Architecture

The IoT browsing system architecture consists of IoT devices, middleware (distributed in local gateways and the cloud) and application interfaces, as shown

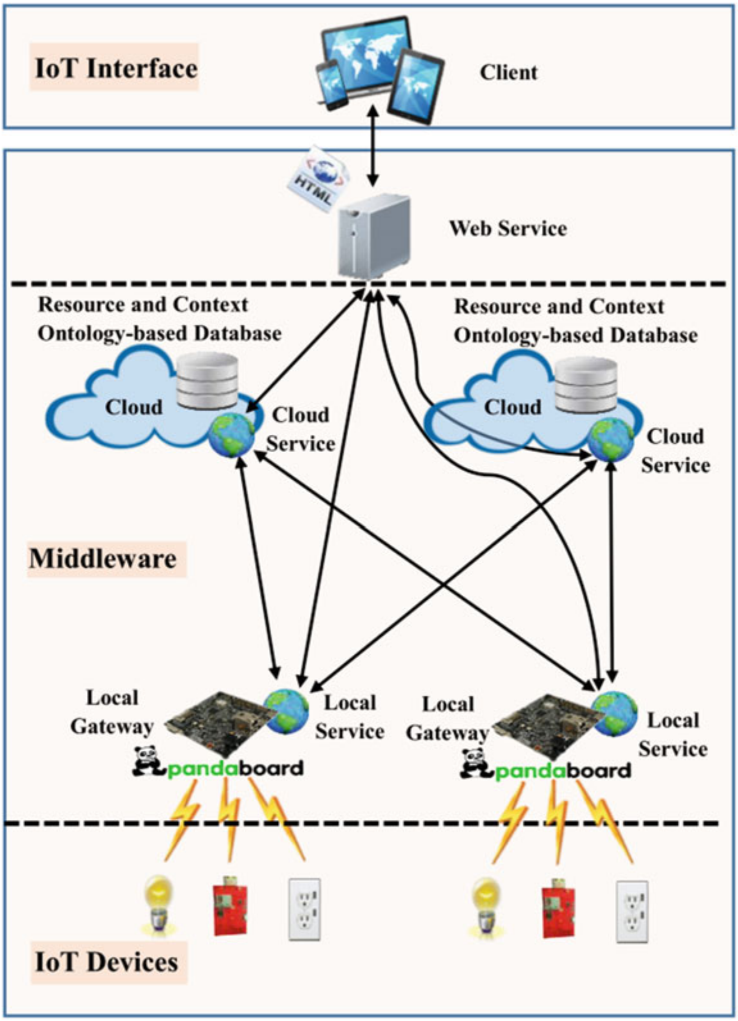


Fig. 6 IoT browsing system architecture

in Fig. 6. In particular, the functions of the middleware can reside in local gateways and/or the cloud depending on the capability of the gateways. The powerful gateways supporting more functions can provide users the capability to browse the objects around them efficiently. However, it requires more resources (i.e. memory, and storage space) and computational ability on the gateways. The resource-limited gateways contain only the device interface modules, leaving most functions in the cloud, and the cost is less. We adopt the limited one in our current implementation of the IoT browsing system.

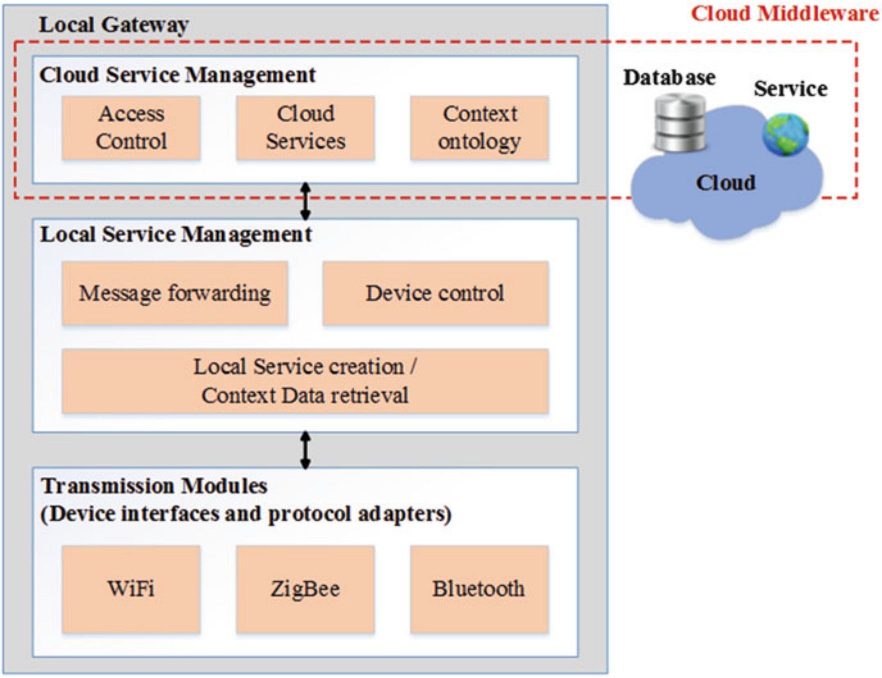


Fig. 7 Function blocks of middleware

Various hardware devices with different communication technologies and protocols (e.g., WiFi, ZigBee and Bluetooth) are available in IoT, and they are responsible for interacting with the physical world such as sensing data or doing specific tasks. The observed raw sensing data can be text, graph, audio, etc. These devices may work alone or together with other devices to provide context data or services.

The middleware depicted in Fig. 7 maintains many transmission modules on the local gateways, and they communicate with server process via inter-process communication (IPC). A transmission module is responsible for communicating with its associated IoT devices and forwarding the messages from the devices to the cloud via server process. After the module creates connections with the devices, its associated devices receive the request for the device description files (e.g., SensorML profile) to register in the cloud services.

Cloud services are hosted in the cloud and utilize a large number of devices under multiple local gateways. After the devices are registered to the cloud services through the local gateways, the cloud services notice the existence of these devices and retrieve their information, such as their locations, parent gateways and the services provided by them. In the meantime, the cloud services update their resources and context ontologies that record the device profiles along with the spatial-temporal information of the IoT devices.

The raw sensing data observed by the registered sensors are collected by the local gateways and will be forwarded to the cloud. Then, the cloud service transforms them into context data according to the metadata and updates the context ontology-based databases for later rule-based inference to complex status. Benefiting from the linked sensor data in RDF and the concept sharing in ontology, the sensing data can be shared among multiple cloud services. Last but not the least, enforcing access control to the devices is essential to prevent their being accessed by unauthorized malicious users.

The reasons for introducing local gateways in the system architecture include the hardware limitation in terms of wireline/wireless communication technologies and overhead of transmitting metadata (i.e., device description files and context metadata profiles). Many devices still need a gateway to connect them to the Internet. Though some devices are used for local services and need not to be queried from the Internet, some client devices (e.g., smartphones) may not support specific communication technologies such as ZigBee to communicate with the devices. From another point of view, the overhead of transmitting semantically-annotated XML-based metadata to the cloud may increase the power consumption, which is critical to power-constrained devices. With the help of local gateways, the power consumption for the communication of these devices could be reduced.

The application interfaces of our designed IoT browser, developed by web application development techniques (e.g., HTML, JavaScript, CSS, Flash and server-side scripting languages like ASP and PHP) or native application development techniques (e.g., iOS and Objective-C for iPhone native apps), can be accessed by smart phones or other client devices. We adopt jsPlumb [38] which provides visually element connection on web pages as our user interface. In our browser, each element and link represents an IoT device (e.g., sensor or smart thing) and the relation between IoT devices respectively. Users can modify the relations among IoT devices to ask the system provide services they want. Moreover, the system can suggest the users how to build the services that are more suitable for them by learning.

3.2 Resource and Context Ontology

The primary difference of IoT browsing from traditional web browsing is to handle the spatial-temporal information, which is absent in traditional web browsing and is critical to IoT browsing, since the users demand an intuitive way to know what devices they can use within a specific area and time. The spatial-temporal information can be either absolute location (e.g., longitude and latitude) or relative location (e.g., inside a building and at specific room) with associated time/date information. Moreover, the network topology information (e.g., the devices under which local gateway) and logical relationship (e.g., which devices work together) are important as well.

In Fig. 7, the middleware provides functions to construct resource and context ontologies that can manage the above information and store relationships between places, users, devices and data in ontology-based databases (i.e., RDF triple stores), which are optimized for storage/retrieval of RDF triples [39]. For example, when a user wants to turn off all the lights at a specific bedroom at home through an IoT service, the service can search the context ontology via SPARQL to know which devices are located in this bedroom and retrieve the instructions from the resource ontology to turn off the lights through the interoperable interfaces provided by the local gateway. Resource and context ontologies are designed to be application-specific to best suit the application purpose, and they can be mapped to global domain ontologies (e.g., DBpedia and W3C SSN Ontology) for being integrated into other resource ontologies (i.e., domain reference). This domain globalizing approach avoids too application-specific domain models or too generic annotations [40].

3.3 Context Aware and Source Display

In the IoT environment, new ways for user interacting with the resources (e.g., sensing data or devices in physical world) are emerging. IoT Browser allows users, for example, to use smart phones to trace current events easily by a simple sensing action to activate a camera at the site of events. However, the IoT devices may contain many kinds of information that are suitable to be displayed in different ways. For example, when browsing home, user may want to know the environmental conditions (e.g., current temperature and CO_2 concentration), control the lights and air conditioners, or interact with the home members. However, a smart phone or other client devices may not be able to display all the information at the same time due to the restriction of the display size. For the purpose of intuitive browsing, the context reasoning is a proper solution to reduce the complicated operation of utilizing IoT services, and IoT browsers show the IoT information in the most suitable way according to the user preference and the environment context. For example, after a family member just finishes exercise, when he/she browses the house, the browser will display the temperature of the house and control the air conditioner automatically. Further, if he/she browses home while working, the browser may display the state of home appliances so that these appliances could be in proper conditions before arriving home.

3.4 Flow Based Interface and Learning

In our system, we allow users to specify a sequence of events, such as users' daily routines. We use FBP to develop our web interface. Users can configure the relation among IoT devices through the interface to determine the sequence of event flow according to their habits or preferences. However, sometimes the “real” sequence of event flow may be different from the pre-determined sequence specified by the users. Thus, we provide a learning mechanism, named Frequent Event Suggested (FES), which can help users determine their flow sequence and automatically provide the next possible event after an event is triggered. FES consists of two phases, namely *pattern discovering* and *list suggestion* that are detailed in the following.

- *Pattern Discovering*
  In the pattern discovering phase, FES searches the historical records in the ontology database to find the event pattern (sequence), which is similar to the event flow built so far by the user through the IoT browser. All the events are stored with information (i.e. time, and location) that can be used to determine the relation among relevant events. FES first employs the longest common subsequence (LCS) algorithm to find all sequences that match the user pattern. Then, FES discards some sequences based on time threshold (the threshold can be pre-defined by users) or location information of the events. For example, if the user pattern is A-B-C and FES finds an A-B-C event sequence from the ontology database with event C occurring 8 h (time threshold) after event B, the sequence can be discarded since the period between occurrences of event B and event C was too long, and thus we can infer that these two events have no relation. The *Pattern Discovering* algorithm is shown in Algorithm 1.

Algorithm 1. *Pattern Discovering*

```
Input: Pattern $\mathbf{P}$
Output: Sequence set $\mathbf{S}^* = \{S_1, S_2, \ldots, S_N\}$ where each sequence $S_n$ matches input pattern $\mathbf{P}$
Run LCS algorithm to find all sequences $\mathbf{S}^*$ that matches input pattern $\mathbf{P}$
for each sequence $S_n$ in $\mathbf{S}^*$ do
    Check if each event $E_{i_n} \in S_n$ has relation with each other (i.e. time, and location)
    if not do
        discard sequence $S_n$
        Refresh $\mathbf{S}^* = \mathbf{S}^* \setminus S_n$
    end if
end for
return $\mathbf{S}^*$
```

- *List Suggestion*
  In the list suggestion phase, FES first finds all the next possible events based on the sequences found in Pattern Discovering phase, that is, the events that may occur next to the sequences in the ontology database. Then, FES sorts the events based on their occurrence frequency in different domains (e.g., days or time periods) and suggests the sorting list to the user. As an example, we denote each domain as a type of day and consider two types of day, that is, workday and holiday. Suppose that there’s a user who does different things after coming home

on different types of day. If it's a holiday, the user usually watches TV. On the other hand, the user usually takes a shower if it's a workday. Therefore, when the event "main door open" is triggered (i.e., the user is coming home), two possible events may occur. If it's a holiday, the next event is "turn on the TV". Contrarily, if it's a workday, the next event is "turn on the water heater". Therefore, FES can provide the user different suggestions based on the user habits in different types of day. The *List Suggestion* algorithm is shown in Algorithm 2.

Algorithm 2. *List Suggestion*

```
Input: Sequence set $\mathbf{S}^*$
Output: The next possible event set $\mathbf{E} = \{E_1, E_2, \ldots, E_M\}$ where the
events are sorted according to their occurrence frequency
for each sequence $S_n$ in $\mathbf{S}^*$ do
    repeat
        Check the event $E_m$ subsequently occurs after the sequence
        $S_n$ in the ontology database
    until $E_m$ has relation with the last event $E_{I_n} \in S_n$
    Select $E_m$
    Refresh $\mathbf{E} = \mathbf{E} \cup E_m$
    Calculate the number of selection of $E_m$, sel_ $E_m$
end for
Sort $\mathbf{E}$ based on sel_ $E_m$
return $\mathbf{E}$
```

4 System Implementation

In the previous section, we show the components of the system architecture and describe the main functions that enable interoperable communication, context management and reasoning. In this section, we describe the implementation of a prototype of the IoT browsing system and demonstrate its main features in an IoT browsing test scenario

Regarding current trends of heterogeneous IoT devices, we employ various ZigBee and Bluetooth Low Energy 4.0 platforms to verify our system. We choose EcoBT [41] CC2540 BLE and OctopusN [42] CC2540 ZigBee PRO modules as the end devices that are equipped with temperature, humidity and illumination sensors. We use PandaBoard [43] network development platform to design our local gateways. The system is designed through Apache Axis2 [44] with Apache Tomcat [45] to provide web services, and uses semantic technologies through Apache Jena [46] and Protégé [47]. The implementation details of the proposed system are described in the following subsections.

4.1 A Smart Home Browsing Scenario

We present the system in a smart home browsing scenario, in which users browse/monitor the conditions (e.g., temperature, CO_2) in the smart home and manage all the network devices within the smart home. A user can access the smart home browsing system with an IoT web browser by real-world-linking technologies (e.g., quick response codes, near field communication, or visual recognition technology). The browser cannot only offer smart devices' information, but also offer a drag-down tool to construct the event flows. After logging in the system, the user will know all the IoT services provided in this house, such as network device management service that lists all devices installed in the building. The user can explore the hierarchical relationships among the devices (i.e., which cloud services manage the devices), and control them (if their accesses are allowed) through HTTP commands, know whether the devices belong to this home. Moreover, the context data of the smart home (e.g., temperature and CO_2 concentration) collected from various devices can be shared among multiple IoT services and interpreted for contextual events to achieve spatial-temporal-aware and context-aware services for the IoT browser user. For example, as different sensors collaboratively contributing to collection of the context data, emergency events can be interpreted from the context data, and the home environment monitor service can make an alarm when an emergency event occurs.

4.2 Devices Registration

In the above smart home browsing scenario, the IoT browsing system is activated by starting up local home gateways to manage the IoT devices within the smart home. When an IoT device connects to the LANs/PANs, it needs to send a registration message to the local gateway. Then, the local gateway forwards the registration message to the cloud. The registration message, which contains device identification, device capabilities and network address, can be encoded in a semantically annotated XML-based description file (e.g., SensorML profile). With the created device description files, the cloud server registers the device for interactive and browsing applications. The registration information is then stored in the resource ontology-based database and context ontology-based database, which are hosted by Apache Jena TDB [46] for RDF storage. The registration procedure for IoT devices is depicted in Fig. 8.

4.3 Semantic Markup

In order to realize automatic device configuration and context management, the semantic annotation techniques (e.g., XLink, RDFa, GRDDL, SAWSDL, etc. [7]) are incorporated in our IoT browsing system. When the cloud services receive

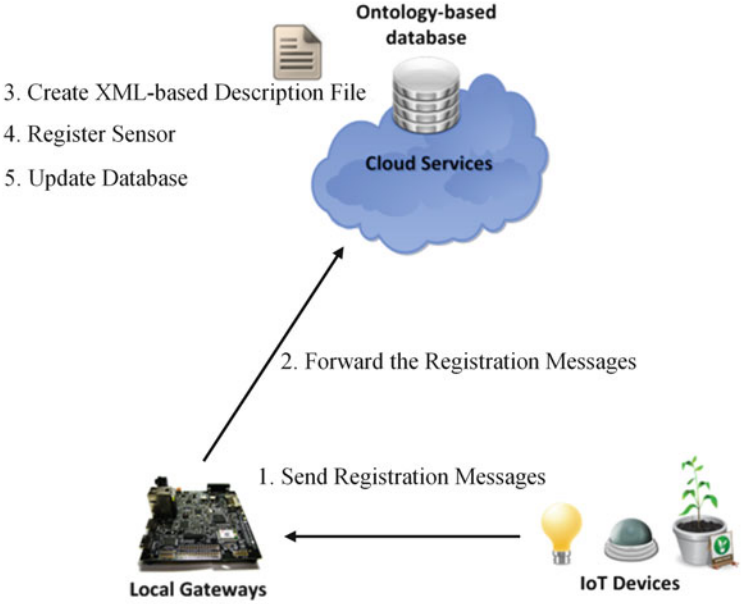

Ontology-based database
3. Create XML-based Description File
4. Register Sensor
5. Update Database
Cloud Services
2. Forward the Registration Messages
1. Send Registration Messages
Local Gateways
IoT Devices

Fig. 8 The registration procedure of IoT devices

the semantically-annotated description files or semantically-annotated metadata profiles, they extract the semantic information to construct the relations among devices or context data, and update the ontology-based databases automatically. In our system, we construct a context ontology-based database with five context concepts, that is, metadata of sensor, electrical equipment, location, furniture, and sensor platform which are denoted as “Sensor_type”, “Electrical_equipment”, “Location”, “Furniture”, and “Sensor_device” respectively as depicted in Fig. 9. For simplicity, many properties and relations are not shown in this figure. According to the RDF statements in this example, the related information about the CO_2 sensor, “CO2_01” is constructed. It shows that the CO_2 sensor is located in the Kitchen, “kitchen_01”.

To describe the relation among objects, the object property is introduced. Here, an object represents a semantic description (e.g. things, time, or locations etc.). In Fig. 10, for example, the term “hasLocatedIn” describes the relation between CO2_01 and Kitchen_01 meaning that CO2_01 is in the Kitchen_01. Moreover, data property is introduced to describe the information of each object. An example is shown in Fig. 11 which describes the MAC address of “Eco_01”.

Multiple IoT services can share the context concept with each other by matching multiple context ontologies (e.g., “Sensor Type” Ontology and “Location” Ontology illustrated in Fig. 9) and reusing existing ontology (e.g., “Electrical_equipment” Ontology reused by multiple smart home services). Thus they will have a common

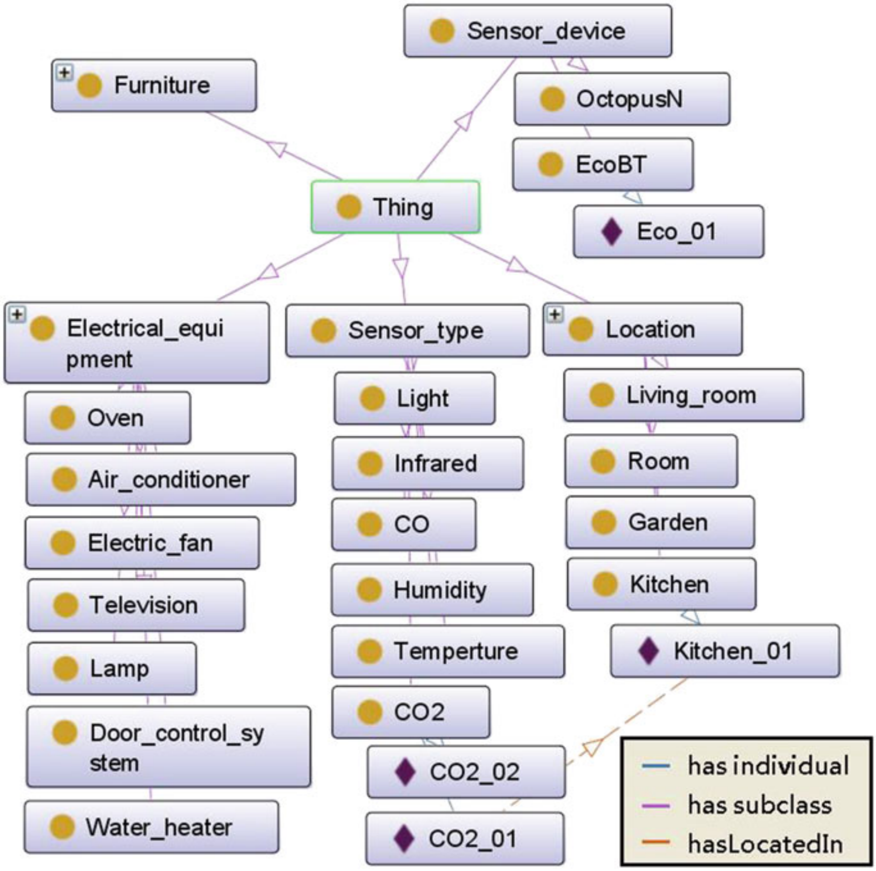
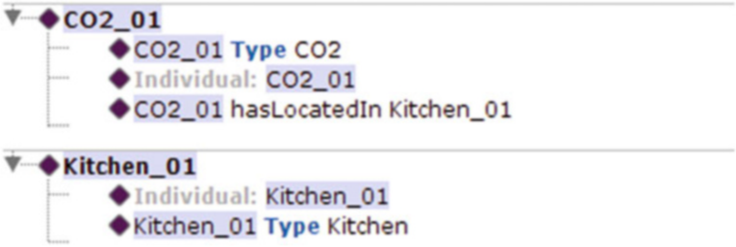


Fig. 9 An example of context ontology-based database

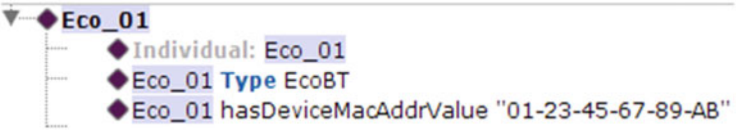


CO2_01
- CO2_01 Type CO2
- Individual: CO2_01
- CO2_01 hasLocatedIn Kitchen_01

Kitchen_01
- Individual: Kitchen_01
- Kitchen_01 Type Kitchen

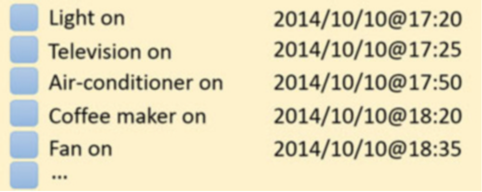
Fig. 10 An example of relation among objects

knowledge for context data processing and displaying (e.g., "Location" concept illustrated in Figs. 9 and 10). With the common knowledge, IoT services can be easily collaborated with others for IoT applications.



Eco_01
Individual: Eco_01
Eco_01 Type EcoBT
Eco_01 hasDeviceMacAddrValue "01-23-45-67-89-AB"

Fig. 11 An example of data property of objects



Light on	2014/10/10@17:20
Television on	2014/10/10@17:25
Air-conditioner on	2014/10/10@17:50
Coffee maker on	2014/10/10@18:20
Fan on	2014/10/10@18:35
...	

Fig. 12 A list of event information

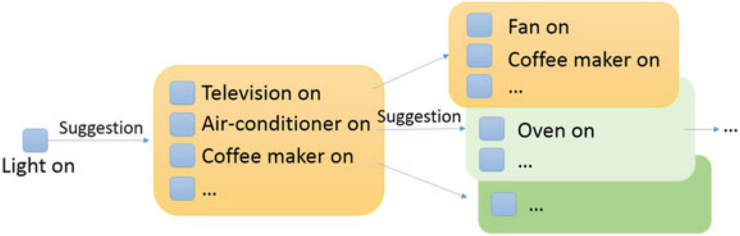


Fig. 13 An example of learning

4.4 Learning

We implement a LCS-based learning algorithm to predict the next possible events as described in Sect. 3.4. When the user builds their event flows on the IoT browser according to their habits and preferences, learning will provide the user suggestion of next possible events based on the current event flow the user has built. First, learning extracts the historical event information from the ontology database which can be seen as a string list as shown in Fig. 12. Then, learning uses the event flow that the user has built as the reference pattern to compare with the records to find the next possible events. An example is shown in Fig. 13. Initially, the user only sets up the event called “Light on” on the IoT browser. Then, learning suggests the user some next possible events, such as “Television on”, “Air-conditioner on”, and “Coffee maker on”, based on the current event flow (“Light on” only now). Next, if the user chooses “Television on” as the next event, learning will suggest the user “Fan on” and “Coffee maker on” based on the event flow order of “Light on” and “Television on”. With the help of learning, the user can build the event flow more accurately since the event flow that the user builds sometimes may be different from the events that actually occur.

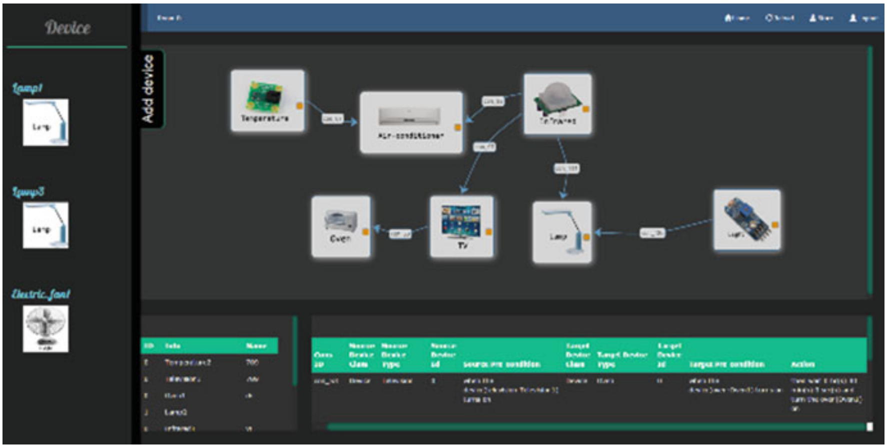
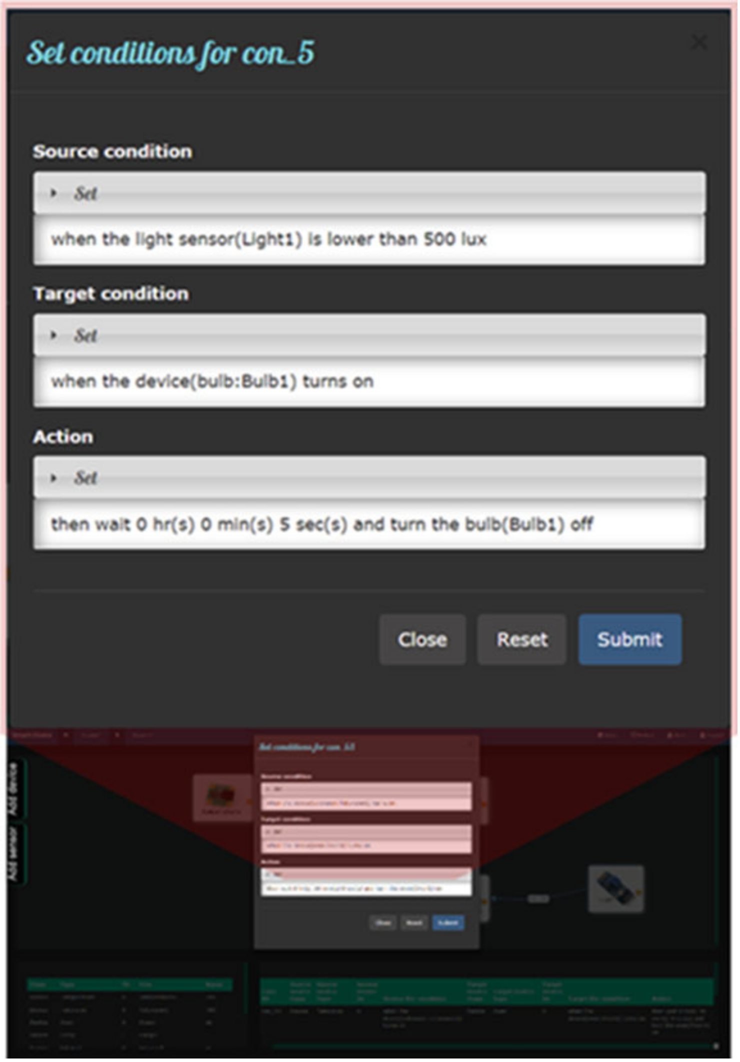


Fig. 14 Registered devices, event flow, and the information of devices

4.5 *IoT Browser User Interface*

Figure 14 presents a user interface for browsing and managing network devices within the smart home environment. The users can access the smart home browsing system with an IoT browser by indexing to a URL retrieved from search engines. After entering the system, the users will know all the IoT services in their home such as network device management service that lists all devices installed in the house as depicted, and can design their own event flows through the web interface as shown in Fig. 14 (the devices that have been registered are listed in the left column, the event flow in the middle column, and the detailed information of devices in the bottom column of the interface). Each block in the figure is an event associated with the device, and several continuous events connected by the directed lines comprise an event flow. When the event at the start of the arrow occurs, the event at the end of arrow will be triggered if all its preconditions are met. The detailed information of preconditions of a event can be set by clicking the incoming directed lines, and then the setting window will appear, as shown in Fig. 15. The precondition includes the condition of the previous event (source condition), the condition of the current event needed to be triggered (target condition), and the time needed to wait until the event is triggered (action). To automatically provide the user the services that they build, the system queries and searches the context ontology-based database by SPARQL [24] and uses Apache Jena API [46] to check device states to know whether the preconditions of an event are met. If all the preconditions are met, the event is triggered.



Set conditions for con_5

Source condition

Set

when the light sensor(Light1) is lower than 500 lux

Target condition

Set

when the device(bulb:Bulb1) turns on

Action

Set

then wait 0 hr(s) 0 min(s) 5 sec(s) and turn the bulb(Bulb1) off

Fig. 15 The setting window of events

4.6 Prototyping Experiences

The experiences of our prototype experiment shows that when users explore IoT, they care about what services are provided within the house/building rather than where a specific service is provided. Therefore, our system is designed based on this feature and manages the spatial-temporal information about devices and context data efficiently. Furthermore, the spatial-temporal concepts for IoT services are unified for easy discovery (e.g., finding all IoT services within the house/building) and control (e.g., building event flows based on the system suggestion). In our implementation, a large number of devices may co-exist in the same area, and it is highly probable that a certain device is required in multiple services. Take temperature sensors as an example, since emergency service (e.g. the fires) and air-conditioner service both require temperature information, it will cause the house to be awash with proprietary devices if they all intend to deploy the temperature sensors rather than reuse the sensors. Web services and semantic technologies are suitable for providing the IoT browsing environment with the above features. On the other hand, the context-aware services will refresh information automatically and, in turn, reduce the operational efforts of browsing IoT.

5 Conclusions and Future Works

In this chapter, we have pointed out the unique features of IoT browsing, which make it distinctive from traditional web browsing in four aspects: (1) the distribution of IoT information, (2) the importance of spatial-temporal characteristic, (3) the linking between virtual world and physical world, and (4) the emergence of novel ways of interaction. Besides, concerning the location-based browsing and the growing IoT services/devices in the future, we have found that the semantic interoperability between multiple IoT services is necessary for frequent IoT services integration and IoT devices reusing. Regarding these features and requirements, we have proposed an IoT browsing system with middleware with learning capability to construct a suitable browsing environment for IoT browser. The system provides semantic device interoperability through module-based middleware and web services. By employing semantic annotation technologies and ontology knowledge sharing, we have achieved device description self-reporting, context (especially spatial-temporal information) management, context concept integration, and resource reusability. Moreover, we design a learning algorithm based on historical context information to help users design their own event flows on the IoT browser. Thus, any IoT devices for various IoT services and applications can be easily positioned and accessed.

The future research line can be extended under the scope of the following aspects: Our implementation experiences reveal that web design techniques should be enhanced in some circumstances where IoT devices join/leave the network frequently. More specifically, in addition to automatic registration and configuration of IoT devices, the procedure of embedding IoT objects to web pages should be simplified, and it is desired to make displaying of IoT information easier. With respect to interaction between users and IoT devices, other real-world-linking technologies along with augmented reality can improve user interaction experience. For example, we can compare the user-captured IoT object images with the images provided by the manufacturer to recognize the IoT objects. There are still a lot of issues need to be addressed, such as a more powerful naming services for controlling

IoT devices and a more efficient way to manage and query graph/RDF-based data models. With the help of more powerful naming services, the control of IoT devices can be achieved more thoroughly such as indicating the routing path. For example, we can change the routing path by changing the hierarchical URL commands. Furthermore, with the growing IoT browsing services, the security issues such as privacy will become more important and should be considered in the system as well. We believe that the IoT browsing system can bring forth fundamental changes in human life, and more attention needs to be paid to the above issues in future works.

References

1. Garcia-Macias JA, Alvarez-Lozano J, Estrada-Martinez P, Aviles-Lopez E. Browsing the internet of things with sentient visors. IEEE Comput. 2011;44(5):46–52.
2. Atzori L, Iera A, Morabito G. The internet of things: a survey. Comput Netw. 2010;54(15):2780–805.
3. Song Z, Cárdenas AA, Masuoka R. Semantic middleware for the Internet of Things. In: Internet of Things (IOT). Tokyo: IEEE; 2010. p. 1–8.
4. Shylby Z. Embedded web services. IEEE Wirel Commun. 2010;17(6):52–7.
5. Noy NF, McGuinness DL. Ontology development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Technical Report; 2001.
6. Happel H-J, Seedorf S. Applications of ontologies in software engineering. In: 2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE 2006). Athens; 2006. http://km.aifb.kit.edu/ws/swese2006/.
7. Barnaghi P, et al. Semantic sensor network XG final report. W3C Incubator Group Report. Technical Report. Accessed 2011 June 28. http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/.
8. Wang X, Zhang D, Gu T, Pung H. Ontology based context modeling and reasoning using OWL. In: IEEE Annual Conference on Pervasive Computing and Communications Workshops; 2004.
9. Mockford K. Web services architecture. BT Technol J. 2004;22:19–26.
10. Wang W, Tse PW, Lee J. Remote machine maintenance system through internet and mobile communication. Int J Adv Manuf Technol. 2007;31(7–8):783–9.
11. Bröring A, et al. New generation sensor web enablement. Sensors. 2011;11(3):2652–99.
12. Botts M, Robins A. Bringing the sensor web together. Géosciences. 2007;6:46–53.
13. Fielding RT. Architectural styles and the design of network-based software architectures. Ph.D. Dissertation, University Of California, Irvine; 2000.
14. Dillon TS, Talevski A, Potdar V, Chang E. Web of things as a framework for ubiquitous intelligence and computing. In: UIC'09: The 6th International Conference on Ubiquitous Intelligence and Computing, ser. Berlin/Heidelberg: Springer; 2009.
15. O'Brien L, Merson P, Bass L. Quality attributes for service-oriented architectures. In: SDSOA'07: The International Workshop on Systems Development in SOA Environments, ser. Washington, DC: IEEE; 2007.
16. Amirian P, Mansourian A. Potential of using web services in distributed GIS applications. GIS Development, MIDDLE EAST, The Middle East Bi Mountly Magazine; November-December 2006.
17. Guinard D, Trifa V, Wilde E. A resource oriented architecture for the Web of Things. In: Internet of Things (IOT). Nov 29–Dec 1 2010; 2010.

18. Kamilaris A, Pitsillides A, Trifa V. The smart home meets the web of things. Int J Ad Hoc Ubiquitous Comput. 2011;7(3):145–54.
19. Gómez-Goiri A, de Ipiña DL. On the complementarity of triple spaces and the web of things. In: WoT'11: The Second ACM International Workshop on Web of Things. New York, NY; 2011. doi:10.1145/1993966.1993983.
20. Miller E. The semantic web: a web of machine processable data. Artificial Intelligence-A Guide to Intelligent Systems; 2004. http://www.w3.org/2004/Talks/0908-egov-em/.
21. Miller E, Swick R. An overview of w3c semantic web activity. Bull Am Soc Inf Sci Technol. 2003;29(4):8–11.
22. Bizer C, Heath T, Berners-Lee T. Linked data - the story so far. Int J Semantic Web Inf Syst. 2009;5(3):1–22.
23. Tiropanis T, Davis H, Millard D, Weal M. Semantic technologies for learning and teaching in the Web 2.0 era - a survey. In: WebSci'09: Society On-Line; 2009. http://eprints.soton.ac.uk/267106/.
24. Prud'hommeaux E, Seaborne A. SPARQL query language for RDF W3C recommendation. Technical Report; 2008. http://www.w3.org/TR/rdf-sparql-query/.
25. Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis. 1993;5(2):199–220.
26. Meditskos G, Dasiopoulou S, Efstathiou V, Kompatsiaris I. SP-ACT: a hybrid framework for complex activity recognition combining owl and SPARQL rules. In: PerCom Workshops. IEEE; 2013. p. 25–30.
27. Compton M, et al. The SSN ontology of the W3C semantic sensor network incubator group. Web Semant Sci Serv Agents World Wide Web 2012;17:25–32.
28. Pfisterer D, et al. SPITFIRE: toward a semantic web of things. IEEE Commun Mag. 2011;49(11):40–8.
29. Dey AK. Understanding and using context. Pers Ubiquit Comput. 2001;5(1):4–7.
30. Gutheim P. An ontology-based context inference service for mobile applications in next-generation networks. IEEE Commun Mag. 2011;49(1):60–6.
31. Strang T, Popien CL (2004) A context modeling survey. In: UbiComp'04: Workshop on Advanced Context Modelling, Reasoning and Management-The Sixth International Conference on Ubiquitous Computing; 2004.
32. Sheng QZ, Yu J, Dustdar S. Enabling context-aware web services: methods, architectures, and technologies. Boca Raton: CRC Press; 2010.
33. Morrison JP. Flow-based programming. In: A new approach to application development. 2nd ed. CreateSpace Independent Publishing Platform; 2010.
34. Guinard D, Floerkemeier C, Sarma S. Cloud computing, REST and Mashups to simplify RFID application development and deployment. In: WoT'11: Workshop on the Web of Things; 2011.
35. Reijers N, Lin K-J, Wang Y-C, Shih C-S, Hsu JY. Design of an intelligent middleware for flexible sensor configuration in M2M systems. In: SENSORNETS'13: International Conference on Sensor Networks; 2013.
36. Rashidi P, Cook DJ, Holder L, Schmitter-Edgecombe M. Discovering activities to recognize and track in a smart environment. IEEE Trans Knowl Data Eng. 2010;23(4):527–39.
37. Rashidi P, Cook DJ. Mining and monitoring patterns of daily routines for assisted living in real world settings. In: the 1st ACM International Health Informatics Symposium; 2010.
38. jsplumb. 2014. http://jsplumbtoolkit.com/doc/home.html.
39. Weiss C, Karras P, Bernstein A. Hexastore: sextuple indexing for semantic web data management. VLDB Endowment. 2008;1(1):1008–19.
40. Maué P. Semantic annotations in OGC standards, OGC 08-167. Open Geospatial Consortium Discussion Paper. Technical Report; 2008.
41. Chou PH, et al. EcoBT sensor platform; 2012. http://eco.epl.tw/
42. Sheu JP. Octopus_N wiki;2012. http://hscc.cs.nthu.edu.tw/~sheujp/wiki/index.php/Octopus_N.

43. PandaBoard. 2013. http://www.pandaboard.org/.
44. Apache Axis2/java. The Apache software foundation. 2012. http://axis.apache.org/axis2/java/core/.
45. Apache Tomcat. The Apache software foundation. 2014. http://tomcat.apache.org/.
46. Apache Jena. The Apache software foundation. 2014. http://incubator.apache.org/jena/.
47. Protégé Ontology Editor. Stanford center for biomedical informatics research. 2014. http://protege.stanford.edu/.

Toward Social Services Based on Cyber Physical Systems

Rin-ichiro Taniguchi, Kauzaki Murakami, Atsushi Shimada, Shigeru Takano, Akira Fukuda, and Hiroto Yasuura

Abstract Cyber physical system (CPS) is a general computation concept, in which "Computers (Cyber world)" and "Real world" are integrated via computer networks. In a cyber physical system, there is a loop structure of "observation," "processing" and "feedback" in the real world: (i) various kinds of data are acquired from our real world using various sensors; (ii) then those data are transferred to computers, or cyber world, and are processed and analyzed; (iii) the analyzed results are fed back to the real world and the real world are modified according to the feedback. Based on this loop structure, the real world is changed, or adjusted. The concept of cyber physical system is well suited for the framework of various IT-based social services, and, in this chapter, we present our research project applying the CPS to social services, especially to an energy management problem, which is one of the most crucial issues for our future society.

1 Introduction

It is one of our most important problems in the world to realize a "sustainable society," in which we should seriously consider such issues as follows:

- Energy consumption reduction
- Environmental burden reduction
- Food and water supply
- Health care

We should note that we have to maintain high social activities in the sustainable society. Of course, if we reduced our activities it would be easy to achieve the sustainability. However, essentially, we can not, or we should not, reduce our social activities, and, then, it becomes a difficult problem to achieve the sustainability. In this situation, ICT is a key technology to solve this issue, because, in principle,

R. Taniguchi (✉) • K. Murakami • A. Shimada • S. Takano • A. Fukuda • H. Yasuura
Kyushu University, 744 Motooka Nishi-ku, Fukuoka, Japan
e-mail: rin@kyudai.jp; murakami@ait.kyushu-u.ac.jp; atsushi@limu.ait.kyushu-u.ac.jp; takano@inf.kyushu-u.ac.jp; fukuda@ait.kyushu-u.ac.jp; yasuura.hiroto.117@m.kyushu-u.ac.jp

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_17

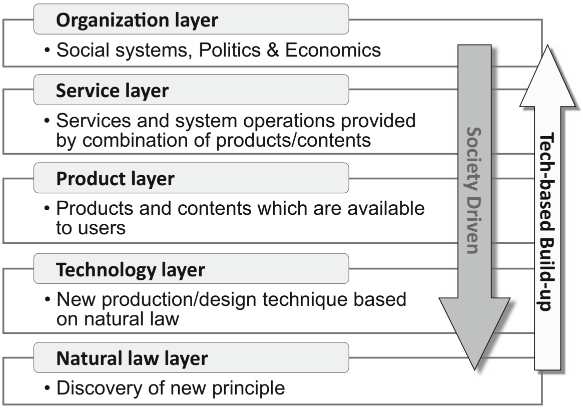

Organization layer
• Social systems, Politics & Economics
Service layer
• Services and system operations provided by combination of products/contents
Product layer
• Products and contents which are available to users
Technology layer
• New production/design technique based on natural law
Natural law layer
• Discovery of new principle
Society Driven
Tech-based Build-up

Fig. 1 Society-driven system development

cost and environmental load required for our activities can be reduced if we can acquire proper information at proper timing. Proper information at proper timing can optimize our social activities in many aspects.

The most important issue when we develop new social services is that we should start from the discussion of what a desirable society is, or what the desirable coming future is. It includes the design of social systems, social policies, etc. Then, we design actual services and system operations realizing the above, and, next, in a backward manner, we design detailed products or applications, etc. Of course, there are a lot of aspects for the desirable society/future, but we should start from the definition of our social goal. This scheme is called "Society-driven System Development," and it is the reverse direction of "Technology Build-up Development" starting from the seeds of technologies (Fig. 1) in a bottom-up manner. Both directions are important, but, society-driven system development, or "backcasting," is quite important especially when we develop new IT systems/services. Without this procedure, developed systems/services are often mismatched to coming social situations, and they are not used at all.

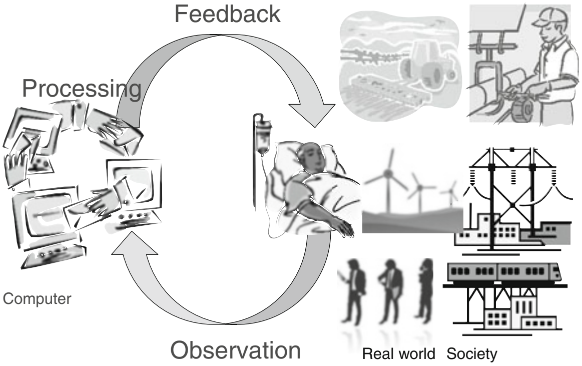
In this chapter, considering the above point, we describe our project to establish sophisticated social services based on ICT, especially based on the idea of "cyber physical system." The topic here is our attempt to a new energy management scheme.

2 Cyber Physical System as the Social Infrastructure

Cyber physical system (CPS) is a general computation concept, in which Computers (Cyber world) and Real world are integrated via computer networks.[1] In a cyber physical system, there is a loop structure of "observation," "processing" and "feedback" in the real world as shown in Fig. 2. Various kinds of data are acquired

[1] Here, we can suppose every kind of networks. Of course, *Internet* can be used for integration.

Fig. 2 Cyber physical system



from our real world using various sensors. Then those data are transferred to computers, or the cyber world, and are processed and analyzed. Analyzed results are fed back to the real world and, then, the real world are modified according to the feedback. Based on this loop structure, the real world is changed, or adjusted. The concept of cyber physical system is general, and it has been used in a wide range of systems/applications, such as small-sized embedded systems, ITS (Intelligent Transportation System), aircrafts and airports, medical equipments, factory automation, smart grids, smart agriculture, smart cities, etc.

In the original concept of cyber physical system, feedback to the real world is given from the cyber world. However, in complex social structures, not all the processed or analyzed results are directly fed to the real world. Rather, it is often adequate that the processed or analyzed results are given to human, and that, then, the human makes feed back to the real world according to "their interpretation" of the processed results. This framework is necessary in the case that high-level decision making should be done in our social systems.

Therefore, we extend the original CPS so that human can be included in the system. Feedbacks are reflected indirectly to the real world, i.e., feedbacks are given to human by visualization or other interaction processes, and, then, human control the real world. We call this paradigm as "Social" CPS, which becomes quite important for sophisticated social infrastructures.

Recently, we are establishing the concept of "Urban Operating System," OS of the society, in a large-scaled project "Co-Evolutional Social Systems,"[1] according to the idea of CPS (Fig. 3). Basically, "Operating System" of computers provides us abstraction of hardware, access interface to abstracted hardware (or hardware resources), and to resource management. Application programs use the computer resources through the interface and realize their own programmed functionalities. In a similar way, "Urban Operating System," or UOS for short, provides us unified access interface to sensors, displays and actuators, computation resources and networks distributed in the real world, and we can develop various social service systems on the platform of UOS.

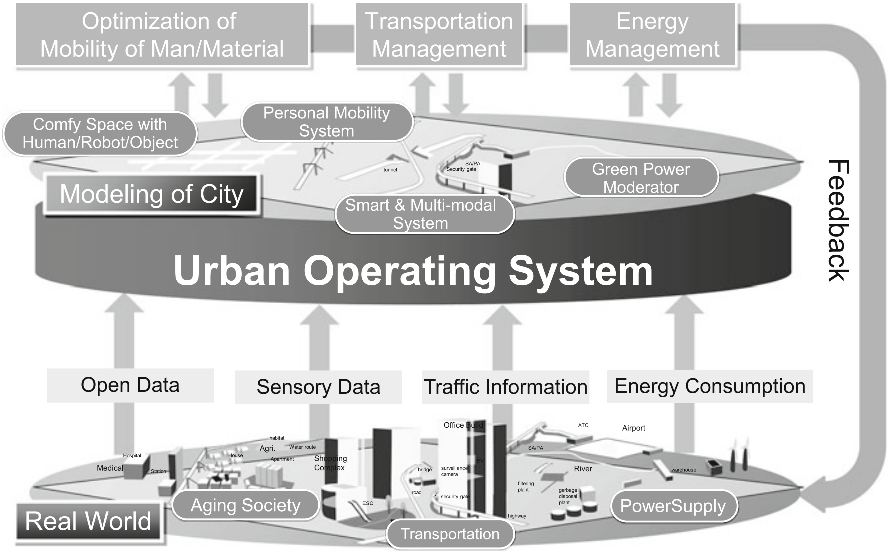


Fig. 3 Urban operating system

Although we have already developed various social service systems, they are developed independently, and, thus, it is not easy to combine them due to lack of a common platform. If we develop applications on UOS, or on the unified framework, we can combine, or mashup, them into new interesting services, which potentially create new social values. It is quite important that the basis of the social services, or, in technical words, API (Application Programming Interface), should be open as much as possible.[2] Many people can construct and provide social services depending on their viewpoints, which might be a right way to cope with the social diversity: young and elder, female and male, handicapped and non-handicapped, foreigner and native, etc.

3 Application to Energy Personalization

3.1 Problem Analysis

As mentioned in the introduction, it is quite an important problem for the sustainable society to solve the energy issue, or to reduce energy consumption. There are a wide variety of approaches to this problem ranging from large-scaled systems to relatively small-scaled systems such as Smart Grid, BEMS (Building Energy Management Systems), HEMS (Home Energy Management System), etc. Smart Grid is an

[2] It is necessary to design a good mechanism to control access privilege for each API. Then, many people can share the unified platform efficiently.

intelligent electric power supply network, which are dynamically controlled to balance electric power consumption with production referring to information of consumers' and suppliers' behavior and other environmental measurement. BEMS is to make energy consumption in buildings efficient by measuring activities of people in the buildings, environmental information such as temperature, brightness of rooms and corridors and by controlling air conditioners, lights and other appliances. HEMS is similar to BEMS, and it is small-scaled and adapted to individual houses. Those are based on typical cyber physical systems having a loop structure of sensing, processing and feedback.

Although the above approaches are effective to some extent, what can we do to achieve further energy efficient society? Of course, as mentioned, the social activities should not be reduced. In addition, we do not want to sacrifice our personal comfort, in other words, personal comfort and social efficiency should be well balanced, which is not an easy problem.

Our basic idea is quite simple—"personalization." If we can observe how each person consumes electricity, i.e., if we can know characteristics of personal electricity consumption with contextual or environmental information, we can realize further efficient electric power savings as follows.

- We can control electric facilities/appliances more efficiently. More over, we can design power-efficient facilities based on the observed features of power consumption.
- By feedback to each person, we expect improvement of action/mind toward energy consumption reduction
- We can expect to discover personal behaviors which lead to reduction of energy consumption

Personalization of the electric power usage is, in other words, fine grained observation of electric power usage situation, but, in general, it is not easy, because electric power usage is usually measured by space, i.e., by building, by house, or by room in the finest case. We do not have a mechanism to observe characteristics of personal power usage precisely. Of course, we can estimate average power consumption by dividing the total amount of power usage by the number of people, but it does not mean anything. Some use a lot, and some use less depending on the situation or the context. We would like to know how each person consumes the electricity more accurately. Therefore, at the first step of our project, correlation of personal action or behavior and observable electricity consumption status (Fig. 4).

3.2 Estimation of Personal Electric Power Consumption

We have global and local sensors which can measure electric power consumption in rooms and by appliances, respectively. By using the global sensors, the energy power consumption $W_r(h)$ in a room r are measured every 30 min.[3] The local

[3] Currently, in some rooms, the consumption are measured every minute to achieve higher accuracy.

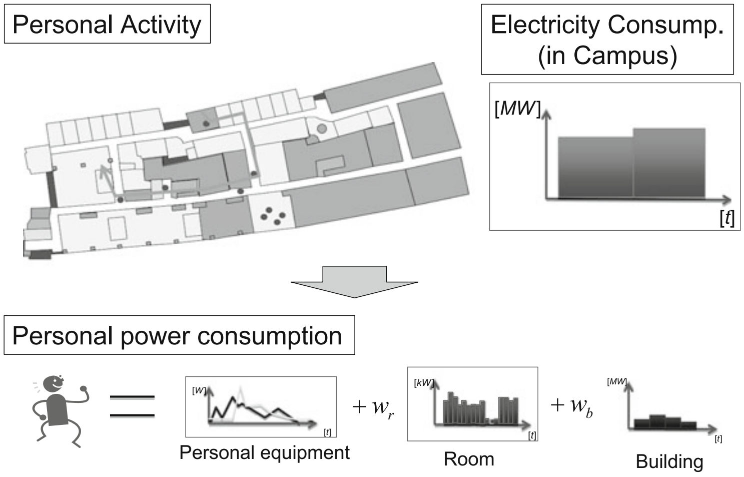


Fig. 4 Personal power consumption

power consumption is measured by smart taps: $p_i(t)$ is electric power consumption by a person i's appliances at time t; $q_j(t)$ is that of a shared appliance j at time t. $p_i(t)$ and $q_j(t)$ can be measured every 5 s. Please note that all the local power consumption represented in $p_i(t)$ and $q_j(t)$ are also accumulated in the measurement by the global sensor, i.e., the power consumption by the room.

Here, $W_r(h)$ can be expressed by the total power consumption of $p_i(t)$, $q_i(t)$ and common power usage $C_r(h)$ such as ceiling lights.

$$W_r(h) = \sum_{t=0}^{T_h-1} \left(\sum_{i=1}^{N_{r,p}} p_i(t) + \sum_{j=1}^{N_{r,q}} q_j(t) \right) + C_r(h), \tag{1}$$

where $N_{r,p}$ is the number of people in the room r, $N_{r,q}$ is the number of shared appliances such as printers and refrigerators in that room, and T_h is the number of samples per 30 min. The common power consumption $C_r(h)$ can be estimated as the difference between the power consumption obtained by the global and local sensors.

$$C_r(h) = W_r(h) - \sum_{t} \left(\sum_{i=1}^{N_{r,p}} p_i(t) + \sum_{j=1}^{N_{r,q}} q_j(t) \right). \tag{2}$$

We have modeled individual electric power consumption as follows:

$$P_i(t) = p_i(t) + Q_i(t) + \hat{C}_i(t), \tag{3}$$

where $Q_i(t)$ and $\hat{C}_i(t)$ are the estimated power consumption per person i at time t by the use of the shared and common appliances, respectively. They are estimated from measured power consumption, W_r, p_i, q_j, and the number and the actions of the people in the room, which is acquired by people activity recognition. It will be described in the next section.

We estimate the shared power usage per person as follows:

$$Q_i(t) = \sum_{j=1}^{N_{r,q}} \alpha_{i,j}(t) q_j(t). \tag{4}$$

Here we introduce an weight function $\alpha_{i,j}(t)$, which is calculated from the activity of the person i and the property of the shared appliance j. For example, in case of refrigerators, their consumption are divided by the number of the people accessed the refrigerator in a certain period. In case of shared printers, a similar idea can be applied because a person who uses a printer usually accesses the printer to get printed papers. The access to such appliances can be, once again, recognized by our people activity recognition scheme. Currently, we judge the person i uses the appliance j, when the person stays near to the appliance (less than τ_0) for a certain period (larger than τ_1).

In this scheme, we should be careful that the timing of access to the appliance is not exactly the same as the timing of power usage. For example, a printer uses electric power when it prints papers, but the timing of the user's access to the printer is after the printing is finished. Therefore, there is some delay, and we have to find correspondence between those.

The common power consumption per person i at time t is estimated as the average of common power consumption (2) for the people within the room as follows:

$$\hat{C}_i(t) = \begin{cases} C_r(h)/(T_h N_{r,p}(t)), & \text{if person } i \text{ is in room } r, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

By substituting the estimated (4) and (5) into (3), we can compute the individual power consumption which takes into account the shared and the common power consumption per the person.

Our current estimation method of personal electric power consumption is not yet matured. We can establish more sophisticated algorithm when people's actions are recognized or categorized more precisely. To realize precise personalization, person action recognition is the key issue.

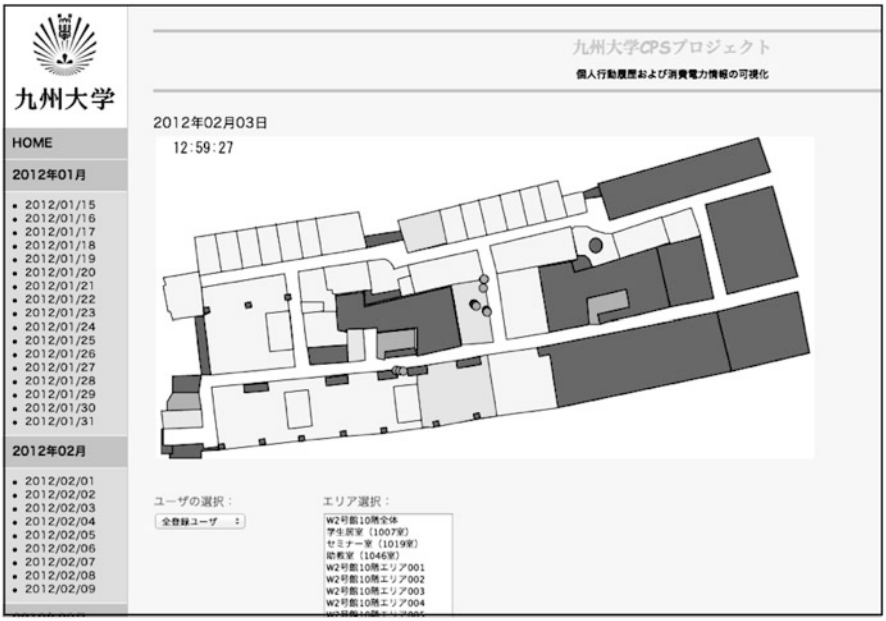


Fig. 5 Visualization of people’s activities

3.3 Experimental Results

As a feasibility study, we acquired the location and track log data of 20 people during 1 month in an experimental field. The acquired log data obtained in the experimental field can be seen on the web browser as shown in Fig. 5.

In Fig. 6a, we show a simple example of a individual electric power consumption obtained by our proposed method, where the horizontal axis is the time scale (sec.), and the vertical one is the electric power consumption (Watt) of the person. We also have a visualization tool (Fig. 6b) which shows person tracking result with his/her status of electric power consumption. It is quite useful to analyze people’s behavior from the viewpoint of energy saving effect.

We describe the details of the result shown in Fig. 6a as follows. First we observed a sudden increase of power consumption since the target person has used a printer. When the energy consumption suddenly decreases at the center of Fig. 6a, the target person has moved from his room to another meeting room. This observation is the reason why the common energy consumption were changed by moving the room. Finally since his laptop PC on his room has changed to sleep mode, we observed the decrease of power consumption.

To realize balancing personal comfort and organizational efficiency based on personalization of energy consumption, we have to solve several difficult issues. The most important issues are clarifying “evaluation figure” including measurement

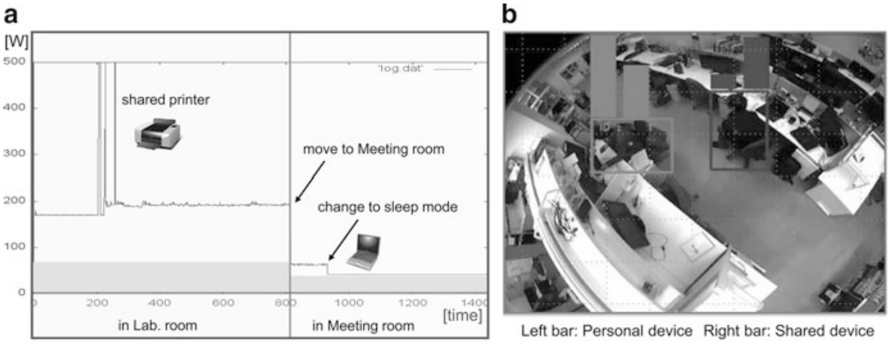


Fig. 6 Visualization of personal electric power consumption. (**a**) Timeline of personal electric power consumption. (**b**) Video-based visualization of personal electric power consumption

of “comfort.” Many schemes to measure “comfort” have been proposed, but there is no standard, or good, scheme. It depends on activities which person does, and we should establish a “context-dependent” measure for comfort. Based on the evaluation figure, we have to establish incentive or penalty, which is one of the most challenging topic in this project.

4 Person Activity Recognition

In our project, person activity recognition is the key component. Here, we explain our method to observe person activity. Of course, the person activity recognition is applicable to wide variety of personalized social services, which are becoming more important to realize human-centered, or co-evolutional, society.

We currently use three kinds of sensors in this project:

- cameras for automatic visual surveillance in which people are automatically tracked in distributed camera views.
- access points of wireless network (WLAN APs), which are used to estimate the users’ positions based on “received signal strength indication (RSSI)” of the mobile terminals,
- IC card readers to identify the users’ ID,

From them, the sequences of the people’s locations are acquired and are integrated into consistent loci for the people. The reason we use three kinds of sensors is that each type of sensors has merits and demerits, that the characteristics of the information provided are slightly different, and that integrating the three methods gives us more accurate location information. Here, each of localization methods is described in the following subsections, and, then, the integration scheme will be discussed in Sect. 4.4.

4.1 Camera-Based Person Localization

Surveillance cameras provide us relatively accurate positional information when the target object in the camera views. However, the sensing areas, or the camera views, are restricted, and we need a certain number of cameras, or multi-view system, if we want to observe a wide area environment. Camera-based person localization, or tracking, can be used many social application such as traffic analysis, purchasing behavior analysis in large-scaled shopping malls, etc. The biggest issue of the multi-view system is to reduce the number of cameras as much as possible, which decreases the cost of system construction.

4.1.1 Basic Concept

We have made the following assumptions.

- Geometrical relationship among the sensors are not known.

  When we construct relatively large-scaled person tracking system, several people can install sensors independently, and, therefore, when we use those sensors we can not expect to have precise geometrical relationship among all of those sensors.
- The sensors are installed sparsely.

  Their views do not overlap with each other. This is mainly due to the cost problem of system construction.
- Not only cameras but sensors acquiring object positions can be used.

  In practical systems, we expect to use not only ordinary video cameras but other types sensors, such as laser range sensors. The laser range sensors do not acquire the texture information of the objects.[4]

To realize object tracking in circumstances where the geometrical configuration among sensors are not known and where their views are not overlapped, we have developed a system having the following features.

- Each object is individually tracked across multiple sensor views. Object identification is realized referring to the topology of the sensor views, or the abstract structure of object flows among sensor views.
- The topology of sensor views are automatically acquired on-line. We do not have to explicitly prepare an off-line training phase. Of course, at beginning, the tracking accuracy is relatively low, but when time goes the accuracy becomes higher. The important point is that when the pattern of object movement changes, we expect the system can adapt the changes according to the on-line estimation algorithm.

[4] It is sometimes preferable in case of protection of personal privacy is strongly required.

4.1.2 Sensor-View Topology

Our tracking method relies on "sensor-view topology," or how sensor-views are related to one another, and, first, we present how the sensor-view topology is represented and acquired in our system.

4.1.3 Definition of Sensor-View Topology

We suppose that an object disappearing at the position, or "exit-point," p_k in a sensor S_i re-appears at the position, or "entry-point," p_l in S_j. If there is a road, or a walk, between p_k and p_l, a certain number of objects disappearing at p_k re-appear at p_l, and we can expect that an object disappearing at p_k will re-appear at p_l with a certain probability. However, it is not easy to make the correspondence between the disappeared object and the re-appeared object, since the appearances of an object change in different sensors. To solve this issue, we suppose that the moving speed of objects does not change largely in the same pair of the exit-point p_k and the entry-point p_l. This means that we can predict the time duration t_{p_k,p_l} required for objects to move from p_k to p_l, and, based on t_{p_k,p_l}, we can make the correspondence. To handle not only the position but also the time duration, we define "exit-information" and "entry-information" of objects. The exit-information of an object consists of the position p_k in S_i and the time t_{p_k}, while the entry-information consists of the position p_l in S_j and the time t_{p_l}.

According to the above consideration, here, we define sensor-view connection as tuple of $\{\boldsymbol{p}_k, \boldsymbol{p}_l, t\}$, that an object disappearing at $\boldsymbol{p}_k = (x_k, y_k)$ in a sensor S_i re-appears at $\boldsymbol{p}_l = (x_l, y_l)$ in S_j after a time duration t. The sensor-view topology is defined as a set of sensor-view connections, observed in a given set of sensor views. When views of sensors S_i and S_j overlap, we can handle such situation by supposing the transit time can be a negative value.

For example, in Fig. 7a, the rectangle represents an observing view of a video sensor S_1, the half circle represents that of a laser range sensor S_2, and arrows represent trajectories of objects which should be tracked. In this case, $\text{OUT}_1 \rightarrow \text{IN}_1$ (5 s) is a sensor-view connection between S_1 and S_2.

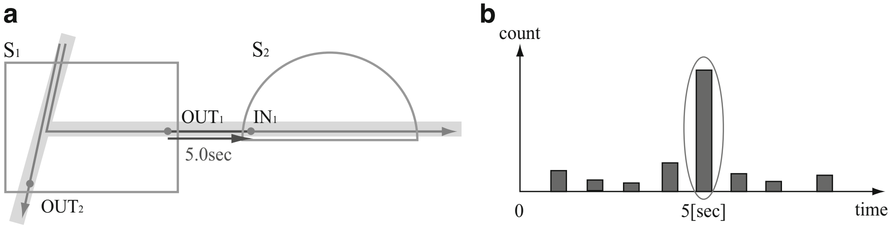


Fig. 7 Example of estimation of a sensor-view connection. (**a**) An example of sensor-view connection. (**b**) A distribution of temporal pairs in estimating $S_1 \rightarrow S_2$

4.1.4 Estimation of Sensor-View Topology

We define a correct pair of exit/entry-information as true correspondence. Meanwhile, the other pairs are defined as false correspondence. When a number of objects are observed simultaneously, there are a lot of pairs of exit/entry-information. We have to extract the true correspondence from them. However, the object identification has not been achieved yet when the topology is being estimated. Therefore, we temporarily make exit-entry point pairs in all combinations, regardless of their correctness, and, we estimate the accurate topology based on the following uniformity relating to true correspondence.

Spatial uniformity Exit and entry-points are observed around some specific points.

Temporal uniformity The moving speed of objects does not change largely from an exit-point to its corresponding entry-point.

For example, we suppose a simple example shown in Fig. 7a, where the connection $S_1 \rightarrow S_2$ should be estimated. When all of the "exit/entry-information" are paired temporarily, a correct pair $\mathrm{OUT}_1 \rightarrow \mathrm{IN}_1$ and incorrect pairs, $\mathrm{OUT}_2 \rightarrow \mathrm{IN}_1$ etc, are obtained. Figure 7b shows a typical histogram of the temporal pairs under the assumption of constant moving speed. The horizontal axis is the transit time of temporal pairs, and the vertical axis is the number of observation. We can find a peak of the distribution at 5 s, which corresponds to correct sensor-view connection, i.e., $\mathrm{OUT}_1 \rightarrow \mathrm{IN}_1$ with the transit time of 5 s.

The issue here is how to correctly estimate transition time when there are ambiguous correspondences having similar transit times. Therefore, to improve the accuracy, we can use additional features acquired from the observing scene if available. For example, in the case of video cameras, appearance features of objects, such as color histogram, becomes helpful information to estimate the topology.

When a connection of an exit-point in S_i and an entry-point in S_j is estimated, temporal connections are voted in a voting space $\boldsymbol{V}^{ij}$. A voted temporal connection is five-dimensional vector: $\boldsymbol{B}^{ij} = (x^{\mathrm{OUT}}, y^{\mathrm{OUT}}, x^{\mathrm{IN}}, y^{\mathrm{IN}}, t)$, where $(x^{\mathrm{OUT}}, y^{\mathrm{OUT}})$, $(x^{\mathrm{IN}}, y^{\mathrm{IN}})$ are the coordinates of exit-point and entry-point, and t is transit time. We can give a weight w to each vote. When the weight is large value, it is highly possible that the exit/entry-information comes from the same object. In our experiment, we have used color histogram of objects if available, and we have implemented a hand-over mechanism of the histogram (Fig. 8). In this case, the w is calculated by histogram intersection.

Sensor-view connections are estimated whenever an object appears. Actually, temporal connections are classified by the Nearest Neighbor method, and each class corresponds to a sensor view connection. Mean μ_{tl} and variance σ_{tl} of transit times of each class, which corresponds to each sensor view connection, is used to identify objects in different sensor views.

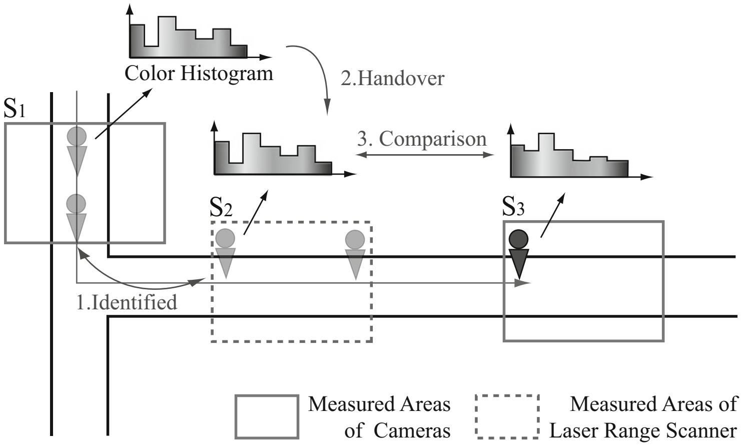

Color Histogram
2.Handover
3. Comparison
S1
S2
S3
1.Identified
Measured Areas of Cameras
Measured Areas of Laser Range Scanner

Fig. 8 Hand over of color information

4.1.5 Object Tracking Across Multiple Sensors

Tracked objects in each sensor[5] are identified based on estimated sensor-view connections. To identify an object which passes in multiple observing views of sensors, we introduce a DoC (Degree of Confidence) of exit/entry-information pairs. A DoC is calculated in the following formula:

$$L(\mathrm{IN}, \mathrm{OUT}) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(\frac{(t-\mu_t)^2}{2\sigma_t^2}\right) \tag{6}$$

where μ_t and σ_t are mean and variance of transit time of a sensor-view connection which corresponds to the exit/entry-information pair. The processing flow of object identification is as follows:

Step1 When an object O_k appears and when its appearing point $(x^{\mathrm{IN}}, y^{\mathrm{IN}})$ is not close to any of entry-points of sensor-view connections, temporal sensor-view connections, which are acquired by combining the entry-information and all the recent exit-information, are calculated and put in the voting space. Then, the sensor view connection estimation process is executed.

Step2 When the appearing point is close to an entry-point of a sensor-view connection, we get a DoC corresponding to the sensor-view connection. Then we search for an disappeared object which maximizes $L(\mathrm{IN}_{O_k}, \mathrm{OUT}_m)$ $(m = 1, \ldots, M)$, where OUT_m represents the exit-information of recent disappeared objects and M is the number of objects.

[5] Object tracking in each sensor is realized by combining basic image analysis methods [2–4].

Step3 When the appearing object is identified, sensor-view connection which consists of the pair of corresponding exit/entry-information is voted to the voting space V^{ij}. The amount of vote is $w = L(\mathrm{IN}_{O_k}, \mathrm{OUT}_m) \times W$ (W is constant when we only use exit/entry information.). Finally, the sensor-view connections are updated.

In our approach, sensor-view connections are estimated every time an object is observed. Therefore, estimation of connections and object tracking are achieved automatically without calibration of sensors' positions.

4.1.6 Performance Evaluation of Wide Area Object Tracking

We performed experiments of people tracking in a campus building using both of video sensors and laser range sensors. We have used 10 cameras and 2 laser range sensors, and Fig. 9 shows a part of the sensor arrangement, i.e., sensors installed in the ground floor. In Fig. 9, solid rectangles represent observing views of the video cameras, and dotted rectangles represent observing views of laser range sensors. In this experiment, we observed the circumstance for about 5 h, and more than 400 people were observed.

Figure 10a shows the result of sensor-view topology estimation, which was acquired after observing the circumstance for 10–15 min. Sensor-view topology was

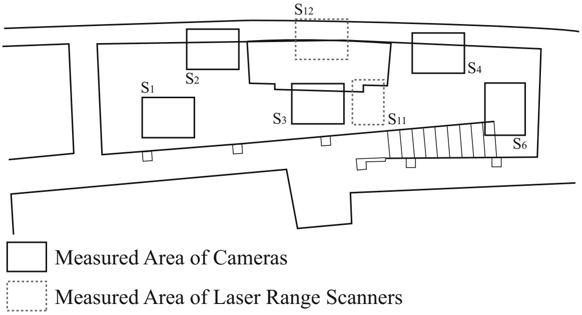


Fig. 9 Configuration of floor and sensor installation

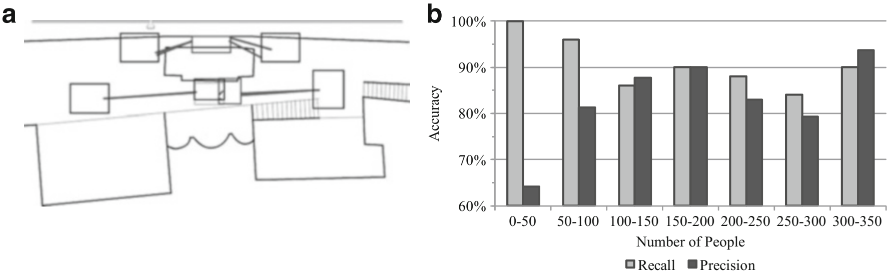


Fig. 10 Experimental results of people tracking. (**a**) Estimation of sensor-view topology. (**b**) Accuracy of people tracking

correctly estimated, and lines connecting the sensor-views in the figure indicate estimated sensor-view connections. Then, we have evaluated the object tracking performance. Here, we used data collected from the sensors in the ground floor, i.e., from 5 video cameras and 2 laser range sensors, where about 350 people are observed. The performance was evaluated in terms of *recall* and *precision*.

Figure 10b shows how the performance changed as the number of observed people increased, where the recall and the precision in each observing period is illustrated. Here, the observing periods are represented in terms of the number of observed people, and, hence, their actual physical times are not the same. At beginning, the performance is not good, but it becomes better as the number of observations increases. Please note that color histogram information was used to evaluate object correspondence when object tracking results were available from video cameras.[6] When several people walking together disappear at the same time and when they re-appear at another entry point, those re-appeared people are possibly mis-identified. It sometimes happens especially when only exit and entry information is used to identify them. However, if we use information reflecting peoples appearance such as color histogram, this problem can largely relaxed.

4.2 *Person Localization based on Wireless Network*

We have built our sensor network based on PicoMESH [5]. The PicoMESH is a mesh-type ad-hoc network, in which a large number of access points can be installed quite easily thanks to its sophisticated wireless backhaul mechanism (Fig. 11). Its network throughput is quite high due to its original efficient packet transmission protocol, which is adequate for sensor networks with many sensor devices. On this PicoMESH network, we have built a localization mechanism of mobile devices referring to the RSSIs received by the access points.

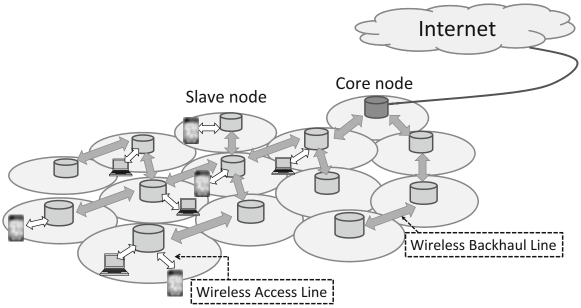


Fig. 11 Sensor network based on PicoMESH technology

[6] The similarity of the histograms was calculated based on histogram intersection.

There are several services in which localization applications are installed in mobile terminals. On the contrary, in our system, the localization algorithm is installed on the server in the network, which refers to the RSSI of terminals measured at the access points. The merit of our scheme is that we do not have to request users to install a localization program on their terminals. It is easy for everyone to use because everything can be prepared in the network system, which simplifies to introduce location-aware services for various kinds of users. In addition, because everything is done in the server, it is possible to estimate the positions of multiple terminals at arbitrary timings, which is quite useful for various applications requiring the distribution of people in a certain area, such as our "personalization of energy usage" project.

The outline of the position estimation algorithm is as follows. For more details, please refer to [6].

1. When a terminal position is estimated, a radio wave emitted from the same terminal is received in different access points, and, thus, we expect there is strong correlation among the signal strengths received at those access points. The correlation is the cue for position estimation.
2. The basic method is "fingerprinting," which consists of two steps: in the first step, or learning step, the positions in terms of a set of predetermined sample positions and the RSSIs at a set of access points are accumulated in the database (primary database); in the second step, or estimation step, the database is searched by the RSSIs at the access points, and possible position candidates are retrieved.
3. The above basic method is, sometimes, not very accurate, and the error correction based on user interaction is provided. If a user finds the estimated position is not correct and if the user knows the correct position, the user provides the correct position to the system. At the same time, the situation, or the context, which consists of time, temperature, humidity, etc, is also stored in the correction database.
4. When the position estimated in the basic mode, and when there is correction information corresponding to the estimated position, it is checked whether the current situation is similar to the situation stored in the correction database. If the both situations are similar, the estimated position is corrected. Of course, the corrected position information can be checked by the users.

Figure 12b shows an example of the position estimation, which shows the method provides relatively accurate results.

4.3 IC Card Readers to Identify the People's ID

In our campus we have various IC card based services such as entering the gates, shopping, bus card and library's services. In this research, we use IC card based management system for people's entry to and exit from buildings and rooms. This original IC card system was developed in our university [7]. Since our management

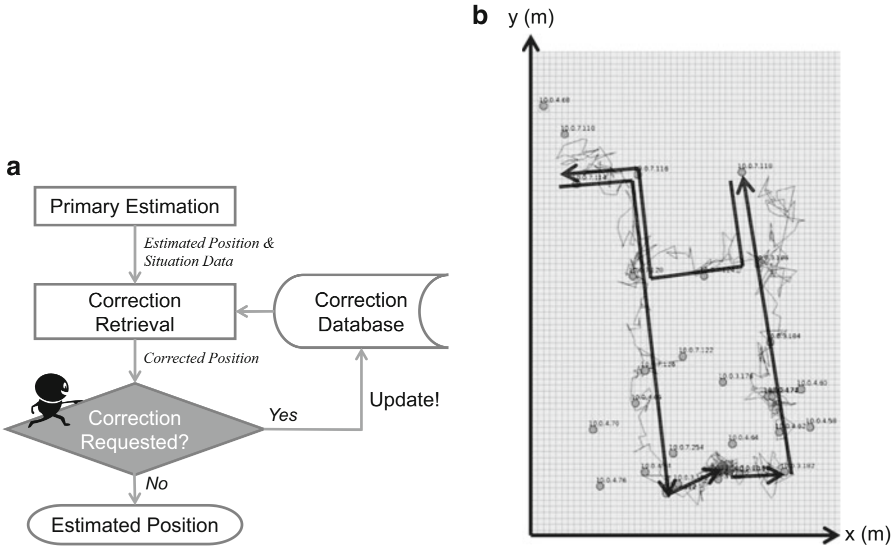


Fig. 12 Position estimation based on WLAN APs. (**a**) Position estimation algorithm. (**b**) Position estimation example

system can store a log of entry and exit information, we can know in real time who is in the buildings or the rooms. It is useful to compute the common power consumption in the rooms. However, since our system can not prevent unauthorized entry to or exit from the rooms by tailgating, we have to develop localization algorithm in cooperation with the other sensors.

4.4 Integration of the Three Localization Methods

The characteristics of the three localization methods used here are summarized as follows:

1. Surveillance cameras provide relatively precise positions of the people in the camera views but it is difficult to provide information about personal ID, i.e., about who they are. In addition, when the people are not in the camera views, their position are not known. We can estimate their positions when they are out of camera views (described in Sect. 4.1), but we can not have very high accuracy.
2. Localization by WLAN APs has relatively large coverage although the accuracy is not very high compared with the camera-based localization. This is because the received strengths of radio waves are sometimes affected by the environment, such as the number of objects in the space. However, when the people are out of the camera views, the position information provided by this scheme is more reliable. In addition, it provides additional information of the mobile terminal IDs, which can be usually bound to personal IDs,

3. IC card readers provide not only personal IDs but also the positions of the people. When a personal ID is sensed at an IC card reader, the person having this personal ID is quite near to the position where the IC card reader is installed. The personal IDs can be used to correct person tracking errors or to disambiguate person tracking results when the personal IDs are matched with extracted loci of the people by the camera-based localization and the WLAN-based localization. However, IC cards are not always used when entering a room. When a group of people enter a room, not all the people use their IC cards, but one of the people uses his/her IC card. Therefore, localization and identification by WLAN APs should be used to make up for this insufficiency.

Considering these characteristics, we integrate sensing results provided by the three types of sensors to acquire person localization results as accurate as possible. The basic scheme is as follows (Fig. 13):

1. IC card IDs are matched with mobile terminal IDs. This is realized by analyzing record of in-room times by IC card IDs and the position information with mobile terminal IDs acquired by WLAN AP localization.
2. IC card IDs and loci of tracked people are matched, under the following assumptions:

    - The spaces around the IC card readers can be observed by the cameras, i.e., the actions of touching IC cards with the readers can be observed.
    - IC card readers are out of the camera views but quite near to the views, and the spatial proximity of the loci to the readers' positions can be strong cue for matching.

If there is any ambiguity in matching, we hold multiple matching candidates and, once the ambiguity is resolved, the previous ambiguity can be resolved by a backtracking process.

For the positions of the people out side of the camera views, the position information acquired by WLAN APs has higher priority unless it is very far from the position estimated by the camera-based person localization described in Sect.4.1. This is a simple algorithm, and more sophisticated fusion mechanism is a future research topic.

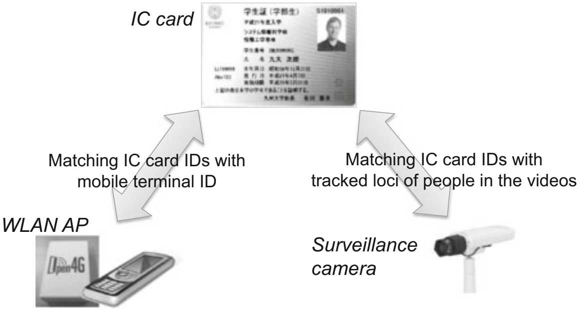

IC card
Matching IC card IDs with mobile terminal ID
Matching IC card IDs with tracked loci of people in the videos
WLAN AP
Surveillance camera

Fig. 13 Integration of three kinds of sensing scheme

5 Conclusion

In this chapter, we have outlined our research on cyber physical system (CPS) and its application to personalization of energy usage based on person activity recognition. This is on-going project and not yet matured. However, fine-grained control of energy is indispensable to reduction of the energy consumption without reducing people's social activities and to balancing personal comfort with social efficiency. To achieve personalization of energy consumption, there are several problems remaining. One of the biggest is how to measure "comfort." Many schemes to measure "comfort" have been proposed, but there is no standard, or good, scheme. We should establish a good measure for "comfort."

As cyber physical systems are becoming important social infrastructures, we would like to develop many social services on this framework. Support for transportation is one of the next possible applications. Especially, adaptive transportation support, or mobility support, for various kinds of people, including elderly people, handicapped people in various environments. Outing, or moving, seems to be one of the essential wishes of human beings, and, thus, support of transportation is very important to make the world vigorous. Also, we are engaged in the application to agriculture, or Smart Agriculture [8]. It should be noted that, in agricultural application, pronouncement of safe and secure products, as well as economic efficiency, is a key issue.

Finally, we should mention the cyber security issues. In the CPS-based society, cyber security is a crucial issue. If cyber attacks destroy social systems, the damage to human life becomes quite huge and it becomes difficult to make a safe and secure life. To make social services flexible and diverse, IT systems supporting the services are constructed on a certain kind of open architecture using *Internet*, and, in such situations, cyber security issues become crucial. We should investigate how to maintain the security of cyber physical systems.

Acknowledgements The research work presented in this chapter has been partly supported by "A Research Promotion program for national level challenges 'Research and Development for the realization of next-generation IT platforms'" by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Center for Co-Evolutional Social Systems. http://coi.kyushu-u.ac.jp/en/.
2. Tanaka T, Shimada A, Arita D, Taniguchi R. Non-parametric background and shadow modeling for object detection. In: Proceedings of 8th Asian conference on computer vision; 2007. p. 159–68.
3. Zhao H, Chen Y, Shao X, Katabira K, Shibasaki R. Monitoring a populated environment using single-row laser range scanners from a mobile platform. In: Proceedings of IEEE international conference on robotics and automation; 2007. p. 4739–45.

4. Isard M, Blake A. CONDENSATION-Conditional density propagation for visual tracking. Int J Comput Vision. 1998;29(1):5–28.
5. Furukawa H. PicoMESH, a W-LAN system enabled by expansive backhaul with wireless multihop capability. The intelligent buildings and smart homes conference 2009.
6. Tagashira S, Kanekiyo Y, Arakawa Y, Kitasuka T, Fukuda A. Collaborative filtering for position estimation error correction in WLAN positioning systems. IEICE Trans Commun. 2011; E94-B(3):649–57.
7. Yasuura H. Towards the digitally named world challenges for new social infrastructures based on information technologies. In: Proceedings of euromicro symposium on digital system design architectures, methods and tools; 2003. p. 17–22.
8. Arita D, Okayasu T, Nugroho AP, Yoshinaga T, Doi N, Shimada A, Taniguchi R. Agricultural information sensing and visualization for farmer-consumer communication. In: 10th Joint workshop on machine perception and robotics; 2014.

Portable Health Clinic: A Telehealthcare System for UnReached Communities

Ashir Ahmed, Andrew Rebeiro-Hargrave, Yasunobu Nohara, Rafiqul Islam Maruf, Partha Pratim Ghosh, Naoki Nakashima, and Hiroto Yasuura

Abstract One billion people (15 % of the world population) are unreached in terms of access to quality healthcare services largely as a result of the paucity of healthcare facilities and medical experts in rural areas. We have prototyped "portable health clinic (PHC), a compact telehealth system with diagnostic equipment and GramHealth software for archiving and searching patients' past health records. The back-end of the system consists of data servers and a medical call center. The front-end has the instances of portable briefcase consisting of medical sensors and measuring equipment operated by healthcare workers living in unreached communities. The front-end data transmission system and Skype telemedicine calls connect with the back-end using mobile network coverage and Internet. Doctors at the medical call center access GramHealth data cloud through the Internet or have a copy of the database in the call center server. Upon receiving a multimedia call from a patient, the doctor can find that patient's previous EHR record and then create and send an e-Prescription. The healthcare worker's PHC briefcase is designed to be low cost and portable. It is envisioned as costing less than US$300 (an amount an entrepreneur can borrow from micro-finance institutions such as Grameen Bank in Bangladesh) and light enough to be carried by a female health assistant. The PHC briefcase will be owned and operated by a village health assistant. This will be a sustainable business model as the health assistant can build a professional relationship with her local clientele. We carried out experiments in three remote villages and in two commercial organizations in Bangladesh by collaborating with local organizations to observe the local adoption of the technology. We are looking at the applicability of our PHC system for aging societies in developed countries.

A. Ahmed (✉) • A. Rebeiro-Hargrave • Y. Nohara • N. Nakashima • H. Yasuura
Kyushu University, Fukuoka, Japan
e-mail: ashir@ait.kyushu-u.ac.jp

R.I. Maruf • P.P. Ghosh
Global Communication Center, Grameen Communications, Dhaka, Bangladesh

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_18

1 Background of Telehealth Systems

Telehealth is the delivery of health-related services and information via telecommunications technologies. Health-related services are delivered by healthcare workers and supported by remote doctors in medical institutions [1]. Telehealth is normally used to keep and treat patients at home and out of hospitals [2]. It is used to remotely monitor chronically ill patients [3]. Telehealth is not used for medical screening programs that test for chronic non-communicable diseases such as hypertension, diabetes, dyslipidemia, obesity, kidney disease and liver dysfunction in individuals who do not show symptoms. Medical screening programs are good for identifying morbidity at an early and treatable stage, and for exposing individuals who normally ignore the disease symptoms [4]. Once identified at risk, a patient is referred to a clinic for treatment. A portable telehealth service at the screening center can shorten the time between the patient being referred and receiving a medical intervention.

Telehealth system can be technically complicated Information Communication Technologies (ICT) and lead to resistance to adopt by all parties involved—healthcare worker, the patient—the remote doctor, and the ICT system. Healthcare workers may reluctantly use a telehealth system that is different from their traditional working practices [5]. Patients may face linguistic and cultural differences when dealing with sensor and video equipment and ICT oriented service providers. Doctors may be concerned with legal considerations and medical liability if there is malfunction in the ICT system or a misunderstanding during a remote patient consultation [6]. The telehealth system can be poorly designed so that technical updates and new measuring elements cannot be accommodated by the technology, thus leading to reinvestment into replacement equipment. However, telehealth systems do not need to be technically complicated or quickly outdated if a simple logic is found that meets the demands of all parties.

Telehealth has received many positive reviews in advanced economies [7]. In the UK, a 3-year telehealth randomized control trial of 6,000 people resulted in a healthcare reductions of emergency admissions (20 %), A&E attendance (15 %), and mortality rates (45 %) [8]. However, not all reviews are so conclusive: a 12 month trial of 1,500 patients with lung disease, diabetes or heart failure found that telehealth had no impact on generic quality of life anxiety or depressive symptoms [9]. In low-income economies, it is difficult to sustain telehealth systems. Low-cost telehealth applications have been piloted and proven to be feasible, clinically useful, sustainable, and scalable in such settings and underserved communities but these applications are not being adopted on a significant scale once the initial seed funding has ended [10]. However, it is possible to build an affordable end-to-end telehealth system for low-income patients that produces Electronic Health Records and is sustainable.

Research Questions and Our Approach This research addresses three questions associated with telehealth:

1. Can a telehealth system supplement medical screening and disease management program?
2. Can a telehealth system be simplified so that a low-literate villager and remote doctor can comfortably talk at the same level?
3. Can a compact telehealth system reduce morbidity?

We discuss two telehealth case studies in Bangladesh. First, a traditional telehealth service provided by Grameen Phone that enables affordable remote medical consultancy to low income people across Bangladesh. Second, a sensor based telehealth system called the Portable Health Clinic which is carried by community health workers and allows an efficient medical screening system and a synchronous telehealth service. Discussing the Portable Health Clinic, we will argue that a colour coded triage application is the most efficient information method for healthcare workers, patients and doctors to understand the severity of morbidity. The triage logic allows ICT system to evolve and accommodate new technology. We presented the results and analysis of two PHC field campaigns in Bangladesh.

1.1 Telehealth by Mobile Phone in Bangladesh

Health advice by mobile phone is the most effective way to reach and serve low-income people with poor access to medical facilities in countries, such as Bangladesh. In a typical developing country scenario, the doctor is located in an urban area in a call center or in his/her house. A patient or a representative from the patient side places a call to a hot-line number of a call center. The call is usually routed to a doctor in a round-robin fashion. There are exceptions when a female patient prefers a female doctor or a follow-up patient looks for the previously consulted doctor. The consultancy contains three major phases: (a) *Introduction phase*: the doctor introduces him/herself, and then asks for patient basic information (name, age, sex, location etc.). (b) *Diagnosis phase*: the patient explains the symptom and then the doctor asks few questions to better understand the cause of the symptom. (c) *Advice phase*: the doctor then either prescribes medicine (over the counter medicine only because of the medical policy issue), or suggests a nearby hospital for further checkup and consultancy. An advanced healthcare service provider keeps the patient–doctor conversation records in a CDR (call details record) and uses special software tool to keep the patient profile details including the list of prescribed medicines (Fig. 1).

We studied two major mobile-phone based healthcare facilities in Bangladesh: (a) 789 health-line service, offered by GrameenPhone, a mobile operator and (b) Tele-health 10600, offered by Japan-Bangla Friendship Hospital. The 789 health-line-service offers medical advice, doctor and medical facility information, drug information, laboratory test interpretation and medical emergency number at

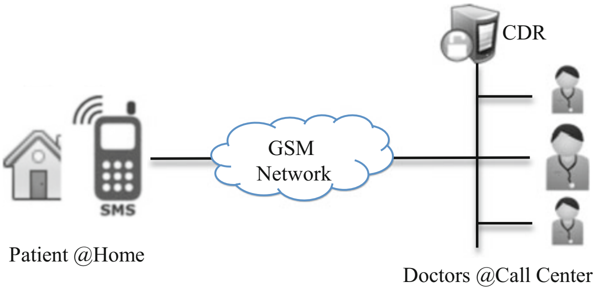


Fig. 1 A typical mobile-phone based remote healthcare system

Table 1 Case study result: what's inside the patient–doctor conversation log

Observed item	Results (n = 400)
(a) Caller	Patient: 60 %, Relatives: 40 %
(b) Age distribution of the patient	0–10 years: 29 %, 11–20 years: 15 %21–30 years: 24 %, 31–40 years: 17 %41–50 years: 9 %, 50+ years: 7 %
(c) Sex	Male: 67 %, Female: 33 %
(d) Location	Rural: 30 %, Urban: 70 %
(e) Call completion	Complete: 68 %, Incomplete: 32 %
(f) Time of call	Day (8:00–15:30): 57 %Evening (15:30–23:00): 18 %Night (23:00–8:00): 25 %
(g) Time occupancy of a single call	Introduction phase: 8 %, Diagnosis phase: 27 %, Advice phase: 67 %
(h) Consultancy about	Disease related: 79 %, Preventive healthcare related: 21 %
(i) Type of advices	Prescribed medicine: 54 %, Advice: 28 %, Referred to specialist/hospital: 17 %
(j) Patients	Follow up: 17 %, New: 83 %
(k) Patients' satisfaction	Fully satisfied: 71 %, unsatisfied: 21 %, average: 8 %
(l) Major diseases consulted	Gastro-intestinal: 22 %, Respiratory: 17 %, Reproductive: 10 %, skin: 10 %

a price of 5 taka (about 6 cents) per minute. Around 200 registered doctors are available in three shifts to serve more than 15,000 calls in 24 h a day. The Telehealth 10600 is relatively small, received 500 calls per day in their starting period in 2008–2009, offered free of charge consultancy to the farmers for their farming related diseases. We have gathered and analyzed conversation records of patient and doctors archived for a month in December 2009. There were more than 10,000 audio call records. We have clustered the records in 100 groups and randomly selected 400 audio records for our case study. In the following table (Table 1), we summarize our observation. It was interesting to observe that there are a good number of female patients making calls (33 %) by themselves. This is quite amazing because usually female patients are always attended by her husband or parent and many times they feel shy to discuss private diseases with a male doctor. Over a mobile phone, the female patients do not see the face of the doctor and do not hesitate to share their private diseases with an unseen doctor. This is an amazing advantage of remote consultancy over mobile phone.

In the followings, we describe the major findings:

(a) Caller: It is not the patients who make the call. There are cases when someone else places the call on the patient's behalf. Some low-literate patients do not feel confident to directly talk to a doctor. Obviously, the kids (<12 years) the older people or very sick patients always depend on a relative to make the call on their behalf (40 %).

(d) Location: The service is not only targeted for the rural patients, the urban poor and also the elite class people also makes calls during emergency period or at unusual time (after midnight) when hospitals are not open. In our observation, 70 % of the calls were made from urban areas.

(e) Call Completion Rate: Not all the calls are complete calls i.e. the doctor–patient conversation do not have a happy ending. There are more than 32 % incomplete calls. The reasons assumed are: due to the poor voice quality, test calls made by curious people to check whether the service really work, sometimes the patients get irritated when the doctor speaks in a very formal way which they are not really accustomed with and they drop the call from their side.

(f) Time of call: 57 % of the calls were made in the day time (8:00–15:30). Normally the clinics in the urban areas start their services in the evening hours. Female patients find this time suitable when they have less household works.

(g) Time occupancy of a single call: We measured the conversation phases to see the pattern how much time is spent for which purpose in a single conversation. The result shows that 67 % of the conversation time is spent to advice a medicine and explain the usage.

(h) Consultancy about: 21 % people make calls for healthcare issues i.e. to know about the preventive healthcare or seeking for explanation of clinical records.

(j) Follow-up Patients: There was no simple way to identify a follow-up patient from the call audio call records. Only one identifier was the caller phone number. We made assumption from the content of the conversation and found that 17 % of the total consulting patients were follow up patients who consulted for the same disease case.

(k) Patients' satisfaction: 71 % people are satisfied with the service. Twenty-one percent of the patients were not satisfied. The reasons assumed due to the voice quality, some people feel irritated when the doctor asks very common or repeating questions-people think that they do it for making more money as the fees increases with the duration of the call.

Technical Challenges Although our study shows that 71 % people are satisfied with the present consultancy service. There is however, a big room for improving the service by introducing simple additional functions into the present system without making any substantial changes in the infrastructure. In this section, we discuss the technical challenges followed by our ideas to address these issues.

Maintaining a Patient ID A patient ID is a key element to keep and maintain individual healthcare records. The present system does not offer a unique ID to their patients. A CDR keeps the mobile phone number of the caller, however there are cases when a patient calls from relatives' phone or uses a family-owned share phone. Therefore, the phone number cannot be a unique ID.

Disease Diagnosis Process In the present system, there is no diagnostic tool at the patient side. The doctors are afraid of making inaccurate assumptions from the symptoms expressed by the caller. A physical measurement is necessary to better understand the degree of a symptom and to make a better clinical decision.

Patient Profile Archive The doctors at the call center are offered and trained to insert the patient profile during the conversation. Many doctors do not feel comfortable to use a computer during the conversation. Also it will take extra time to insert the patient profile which will again irritate the patient. As a result, the patient profile never gets sufficiently stored. Without past records, it is difficult to take care of the follow-up patients.

Patient's Location Currently the call center has to ask a series of questions to identify the geographical location of a patient. A doctor cannot accurately refer a patient to a hospital or to doctor if patient's location is not known.

Prescription The medicines in developing countries have English names. The low-literate patients have difficulties to understand the names prescribed by the doctor and take a memo. Some providers use SMS service to send the medicine names. There is a policy that the doctor can only prescribe OTC medicine. Therefore, the doctors have limitations to treat the patients.

Health Data Portability Some patients have the past clinical records in hard paper format. It is difficult to read out the clinical data for the remote doctor. Some hospitals keep the past records in digital format. Currently there is no scheme to transfer the digital data from one hospital to another. The same is true for the developed countries too.

The technical challenges show that health advice by mobile phone service provides relief but is limited in is effectiveness and cannot supplement a healthcare service. In the next section, we introduce a Portable Health Clinic which uses sensors and telemedicine to provide life-saving healthcare to low-income people in unreached communities in Bangladesh.

2 Portable Health Clinic System

The Portable Health Clinic (PHC) system is an e-health system with a telehealth component. The PHC was designed by Kyushu University and Grameen Communication's Global Communication Center (GCC) to provide affordable e-Health service to low-income subjects living in unreached communities [11]. It consists

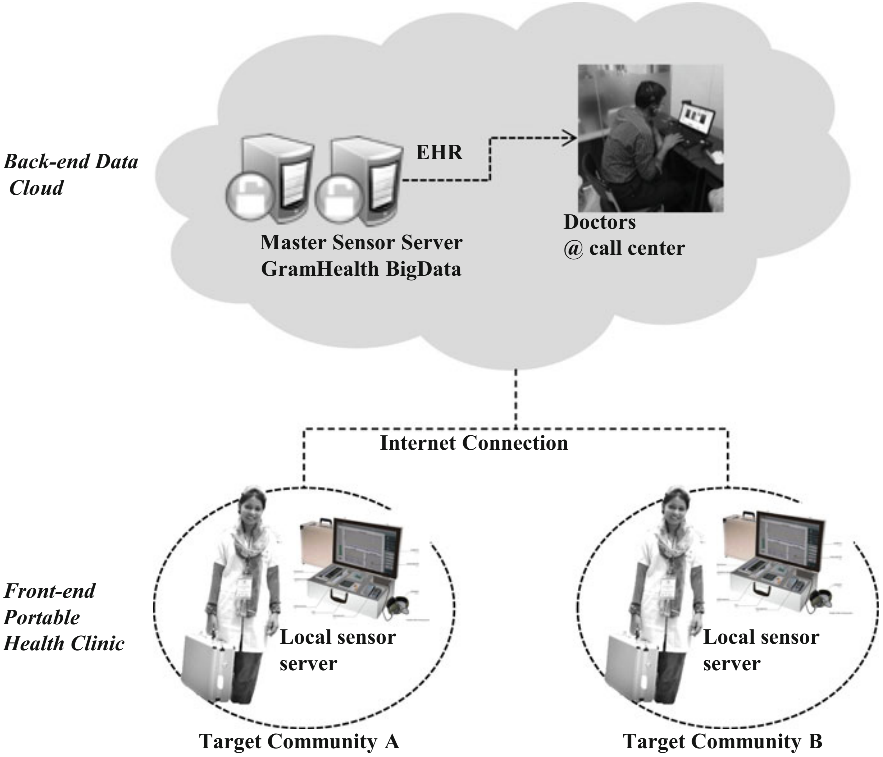


Fig. 2 Portable health clinic architecture

of back-end of data servers and a medical call center, and inexpensive front-end instances of portable briefcase consisting of medical sensors and measuring equipment. The front-end communicates with the back-end using mobile network coverage and Internet (Fig. 2).

The PHC back-end comprises GramHealth software applications, database, and medical call center. GramHealth software applications process patients' Electronic Health Records (EHR) and doctor's e-Prescriptions, and store them in a database. Doctors at the medical call center access GramHealth data cloud through the Internet or have a copy of the database in the call center server. Upon receiving a multimedia call from a patient, the doctor can find patient's previous EHR, can create, and send an e-Prescription [12]. This saves time and effort as the doctor does not need to ask questions about the patients' personal profile (basic attributes and medical history) but can focus on the immediate health inquiry.

The PHC front-end instances consist of a network of medical sensors and devices medical sensors packed into the healthcare worker's briefcase (Fig. 3). The medical sensors are used to identify non-communicable diseases (NCDs) and have a Body Area Network' (BAN) interface to transmit patient data to tablet PC (local sensor

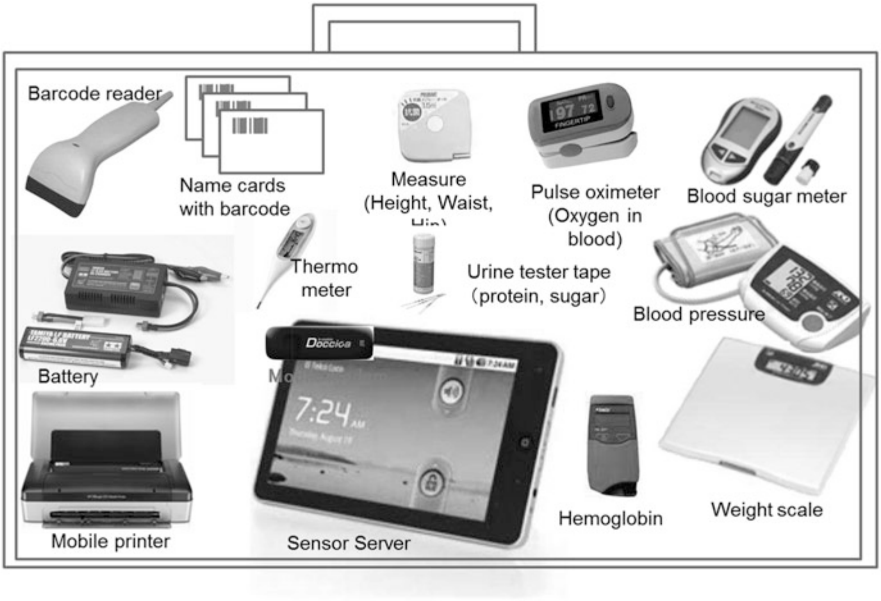


Fig. 3 Sensors and devices product detail in the healthcare worker's briefcase

server within the briefcase). The local sensor server synchronizes its cache with the master sensor server when an Internet connection is available. The master sensor server in the back-end data cloud stores all sensor data and provides data to the GramHealth database and doctors in the call center (Fig. 2). The interface of the local sensor server is the same of that of the master sensor server; therefore sensor boxes can directly connect to the master sensor server by changing the configuration address.

Portable Health Clinic Tools The tools of the Portable Health Clinic are based on Systems Health Care model. Each tool is a process step containing a service component with self-contained functionality:

- **Index**: used for quantification and measurement value such as blood pressure, weight, height, body mass index (BMI). Measurement results are affected by target conditions, measurement method, sensor device, and environment.
- **Criterion**: used for classification and identification. For example, hypertension by blood pressure is diagnosed by 140 mmHg for systolic blood pressure (SBP) according to the guideline [12].
- **Causality**: used for diagnosis and prognosis. Causality is structured by cause and effect. Diagnosis is realized by the direction from effect to cause and prognosis is reverse one, from cause to effect.

The PHC tools are realized using nonintrusive medical sensing devices and ICT framework.

Table 2 Sensor and equipment for a typical portable clinic briefcase

Sensor	Transmission	Weight (gm)	Sensor	Transmission	Weight (gm)
Weight scale	BAN	2500	Tape measure(For Height)	Manual	100
Blood pressure	BAN	300	Tape measure (For waist & hip)	BAN	300
Pulse oxymeter	BAN	60	Mobile printer	Bluetooth	2500
Blood glucose	RFID	50	Mobile scanner	USB	350
Body temperature	RFID	27	Web camera	in-built	-
RFID reader	USB	35	Hemoglobin meter	Manual	350

Index Devices Inside the briefcase Clinical data and quantification of personal health indices are ascertained using medical sensors packed into the healthcare workers briefcase (Fig. 3 and Table 2).

Criterion Using Morbidity Stratification Algorithm Non-communicable morbidity is identified and classified using a triage stratification algorithm. Local sensor data of each health check-up items is compared against risk stratification matrix based on International diagnosis standards (WHO). The results are categorized and graded according to a triage: green (healthy), yellow (caution), orange (affected), and red (emergency). The current triage risk stratification is parameterized against a "B-logic" (Bangladesh logic) [13] and an example of the categories is shown in Fig. 8.

Identifying Causality Using Telemedicine Remote diagnosis is achieved using telemedicine and teleprescription procedures. The healthcare worker sets up a telemedicine session for orange and red subjects using pervasive mobile network coverage to connect the patient to the medical call center at the back-end. In the call center, male and female doctors are available to provide telemedicine. Doctors access the electronic results of subject health check-up and provide advice for the disease and an e-Prescription for the patient to access medicine via the network.

2.1 Healthcare Data Collection and Turnaround Time

Compactness is the most important attribute of the Portable Health Clinic. The briefcase is light enough for a female healthcare worker to carry to a patient's home; to screening point in a rural village and suburban factory; or to a disaster area. All the sensors and devices can be to set up on a table. The sensors are intuitive enough for a low-literate patient to use; and the sensors readings are sent wirelessly to the tablet sensor server.

Telehealth workflow is the next important attribute of the Portable Health Clinic. The healthcare worker follows a repeatable sequence of events to ensure an efficient patient turnaround for group health checkup. When screening a queue of people, the most efficient workflow is:

(a) **Registration**. A patient registers his/her vital information such as name, age, sex, location and disease complaints. A data entry operator inputs the data into GramHealth DB. A patient ID is given to the patient. The patient pays for the service in advance.

(b) **Index**. A healthcare worker conducts the patient's physical check up (body temperature, weight, height, BMI, Waist, Hip, Blood test, Urine test) and data is automatically sent to GramHealth server.

(c) **Criterion**. The sensor server grades the patient according to the color-coded risk stratification: green (healthy), yellow (caution), orange (affected) and red (emergency). The "green" patients are given the health checkup results. The "yellow" marked patients are given a health guidance booklet. The "orange" and "red" marked patients consults with a call center doctor.

(d) **Causality**. Tele health consultancy. Color-coded "yellow" and "red" marked patients talk to the remote doctor for further investigations of their disease and explanation of their medical records. Tele health consultancy is over voice and video. The audio record is archived in GramHealth database.

(e) **Prescription and Suggestion**. The remote doctor identifies the disease after checking the clinical data, discussing with the patient for their symptom analysis and his/her past health records, if any. The doctor then fills up the prescription and a technical assistant helps the doctor to insert the necessary information into the database and sends to the Healthcare worker

(f) **Sign Off**. The Health Assistant prints and gives a copy of the Electronic Health Record and prescription to the patient and schedules a follow-up health checkup within 2 months.

The PHC patient process cycle differs from the fixed health clinic process cycle in the sense that the Index process, (such as measuring blood glucose levels) comes before criterion process (triage segregation), and before the causality process (Doctor consultancy). Therefore in the PHC, every patient is clinically measured and assessed locally before setting up a telemedicine session. This means that the remote doctor is connected only to morbid patients. The differences in the order of process components and patient turnaround time between and fixed health clinic system approach and the portable health clinic approach are shown in Fig. 4.

GramHealth Database We studied the characteristics of GramHealth BigData (a healthcare database archived during our remote healthcare consultancy pilot program in Bangladesh). We assume that careful analysis of these records can produce invaluable medical information that has never been explored. The complex nature (multi-lingual, multi-modal, poly-structured) of these records (Fig. 5) is very difficult to analyze by the current data mining tools. In this work, we attempt to manually analyze these data and produce new medical information to determine analyzing methodology. Table 3 shows the characteristics of our GramHealth BigData. The database archives four different data that has been collected in four different steps. (1) *Registration data* (*Vital data*) containing demographic informa-

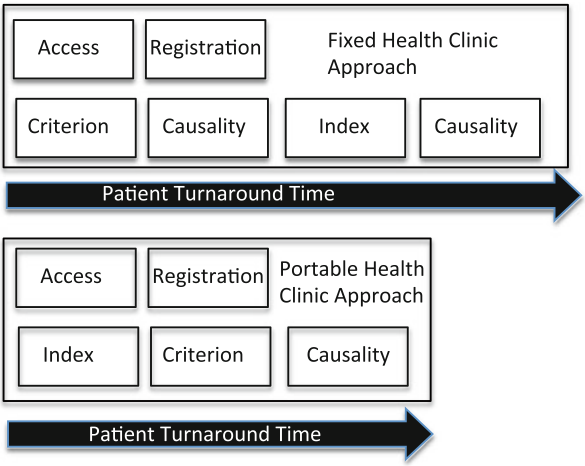
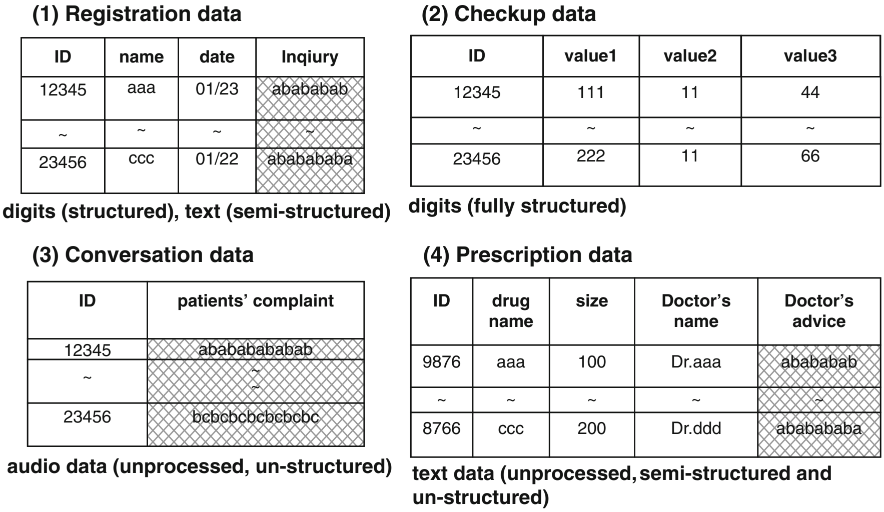


Fig. 4 Process components against patient turnaround time



(1) Registration data

ID	name	date	Inqiury
12345	aaa	01/23	abababab
~	~	~	~
23456	ccc	01/22	ababababa

digits (structured), text (semi-structured)

(2) Checkup data

ID	value1	value2	value3
12345	111	11	44
~	~	~	~
23456	222	11	66

digits (fully structured)

(3) Conversation data

ID	patients' complaint
12345	abababababab
~	~ ~
23456	bcbcbcbcbcbcbc

audio data (unprocessed, un-structured)

(4) Prescription data

ID	drug name	size	Doctor's name	Doctor's advice
9876	aaa	100	Dr.aaa	abababab
~	~	~	~	~
8766	ccc	200	Dr.ddd	ababababa

text data (unprocessed, semi-structured and un-structured)

Fig. 5 Properties of GramHealth database

tion (2) *Checkup data* (Clinical Data) collected by using healthcare measurement tools, PHC (Portable Health Clinic) (3) *Conversation data* between a patient and a doctor, and (4) *Prescription data* that has been suggested by the doctor for that patient. Figure 6 shows when the data is captured during the process cycle.

Table 3 Characteristics of GramHealth DB

Types of data	Data contents	Data structure
(1) Registration data	Patient profile, registration date, past disease inquiry	Structured (patient profile), semi-structured (inquiry)
(2) Check-up data	Physical status (BMI, Blood pressure, SpO2 etc.)	Structured (as the devices produce only digits)
(3) Conversation data	Between the patient and the doctor to identify a disease	Un-structured
(4) Prescription data	Disease name, prescribed medicine, clinical suggestion	Semi-structured and/or un-structured

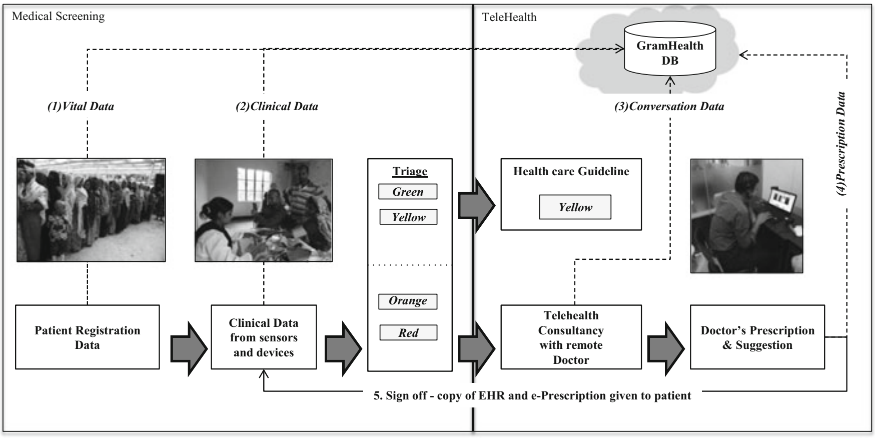
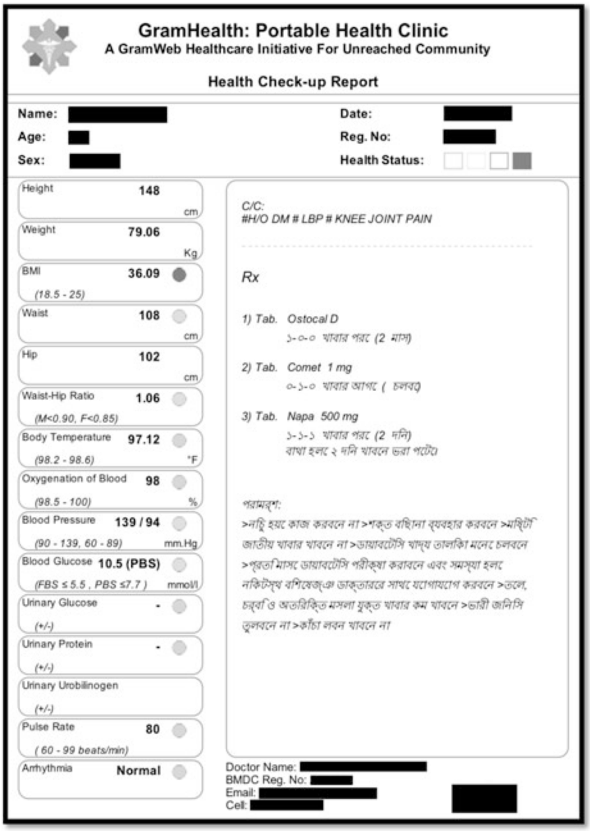

Medical Screening
TeleHealth
GramHealth DB
(1)Vital Data
(2)Clinical Data
(3)Conversation Data
(4)Prescription Data
Triage
Green
Yellow
Orange
Red
Health care Guideline
Yellow
Patient Registration Data
Clinical Data from sensors and devices
Telehealth Consultancy with remote Doctor
Doctor's Prescription & Suggestion
5. Sign off - copy of EHR and e-Prescription given to patient

Fig. 6 PHC medical screening and telehealth process cycle

2.2 Portable Health Clinic Human–Computer Interface

The human–computer interface of the Portable Health Clinic is based upon a color-coded triage decision making algorithm. Color coded risk categorization is very convenient method to convey complex medical information to the healthcare worker and for the low-literate clients in rural villages. When taken at first sight, the individual sensor and devices reading and results are unfamiliar to a semiskilled healthcare worker and her patients. The combination of all the clinical data is often meaningless to both parties. In contrast, when the sensor results are converted to colors of green (safe), yellow (caution), orange (affected) and red (emergency) and shown to the healthcare worker, the healthcare worker can set up a telehealth consultation and the patient will understand which reading is indicating risk to their health and its severity.

The doctor at the call center benefits from the color-coded risk categorization. The distant patients' basic data and clinical data is transmitted by the sensor server to doctor's call center screen and she can immediately see if the next patient is red



GramHealth: Portable Health Clinic
A GramWeb Healthcare Initiative For Unreached Community

Health Check-up Report

Name:	[redacted]	**Date:**	[redacted]
Age:	[redacted]	**Reg. No:**	[redacted]
Sex:	[redacted]	**Health Status:**	□ □ □ ■

Height	148			cm
Weight	79.06			Kg
BMI	36.09	●	(18.5 - 25)	
Waist	108	○		cm
Hip	102			cm
Waist-Hip Ratio	1.06	○	(M<0.90, F<0.85)	
Body Temperature	97.12	○	(98.2 - 98.6)	°F
Oxygenation of Blood	98	○	(98.5 - 100)	%
Blood Pressure	139 / 94	○	(90 - 139, 60 - 89)	mm.Hg
Blood Glucose	10.5 (PBS)	○	(FBS ≤ 5.5 , PBS ≤7.7)	mmol/l
Urinary Glucose	-	○	(+/-)	
Urinary Protein	-	○	(+/-)	
Urinary Urobilinogen			(+/-)	
Pulse Rate	80	○	(60 - 99 beats/min)	
Arrhythmia	Normal	○		

C/C:
#H/O DM # LBP # KNEE JOINT PAIN

Rx

1) Tab. Ostocal D
১-০-০ খাবার পরে (2 মাস)

2) Tab. Comet 1 mg
০-১-০ খাবার আগে (চলবে)

3) Tab. Napa 500 mg
১-১-১ খাবার পরে (2 দনি)
ব্যথা হলে ২ দনি খাবনে ভরা পটে

পরামর্শ:
>নিচু হয়ে কাজ করবনে না >শক্ত বছিানা ব্যবহার করবনে >মিষ্টি জাতীয় খাবার খাবনে না >ডায়াবটেসি খাদ্য তালকিা মনেে চলবনে >প্রতিমাসে ডায়াবটেসি পরীক্ষা করাবনে এবং সমস্যা হলে নকিটস্থ বশিষেজ্ঞ ডাক্তাররে সাথে যোগাযোগ করবনে >তলে, চর্বি ও অতরিক্তি মসলা যুক্ত খাবার কম খাবনে >ভারী জনিসি তুলবনে না >কাঁচা লবন খাবনে না

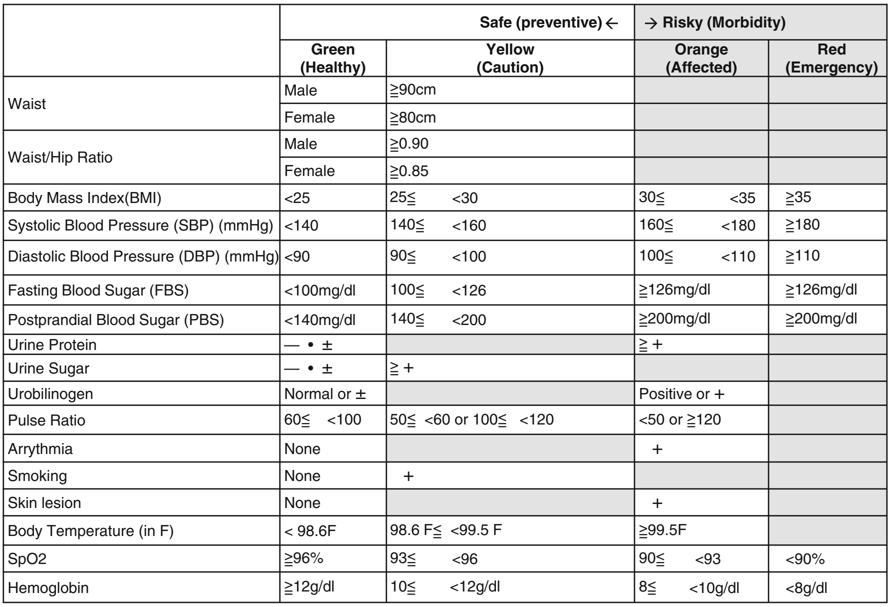
Doctor Name: [redacted]
BMDC Reg. No: [redacted]
Email: [redacted]
Cell: [redacted]
[redacted]

Fig. 7 PHC electronic health record and e-prescription

(emergency) or orange (affected). Without opening any files, the doctor can focus on a clinical readings highlighted with a red circle and make an inference by looking at readings of the other check-up items. The doctor will have a mental image of the patient condition prior to the consultation. During the doctor/patient video telephony consultation, the doctor can refer to triage colors and the low-literate patient will have good idea of what is normal and what is abnormal.

The patient Electronic Health Record and e-Prescription are based on the color-coded risk categorization (Fig. 7). The patient's health status is marked in one color. If one reading is red then the health status are shown as red, even though all other readings are green. The results of each clinical measurement are shown by value and by triage color. This is to help the villager to understand which readings are safe and which readings show risk.

The Portable Health Clinic color-coded triage is a software application located sensor server carried by the healthcare work. The software application is a risk stratification matrix and is based on International diagnosis standards (WHO).



	Safe (preventive)←		→ Risky (Morbidity)	
	Green (Healthy)	Yellow (Caution)	Orange (Affected)	Red (Emergency)
Waist	Male	≧90cm		
	Female	≧80cm		
Waist/Hip Ratio	Male	≧0.90		
	Female	≧0.85		
Body Mass Index(BMI)	<25	25≦ <30	30≦ <35	≧35
Systolic Blood Pressure (SBP) (mmHg)	<140	140≦ <160	160≦ <180	≧180
Diastolic Blood Pressure (DBP) (mmHg)	<90	90≦ <100	100≦ <110	≧110
Fasting Blood Sugar (FBS)	<100mg/dl	100≦ <126	≧126mg/dl	≧126mg/dl
Postprandial Blood Sugar (PBS)	<140mg/dl	140≦ <200	≧200mg/dl	≧200mg/dl
Urine Protein	— • ±		≧ +	
Urine Sugar	— • ±	≧ +		
Urobilinogen	Normal or ±		Positive or +	
Pulse Ratio	60≦ <100	50≦ <60 or 100≦ <120	<50 or ≧120	
Arrythmia	None		+	
Smoking	None	+		
Skin lesion	None		+	
Body Temperature (in F)	< 98.6F	98.6 F≦ <99.5 F	≧99.5F	
SpO2	≧96%	93≦ <96	90≦ <93	<90%
Hemoglobin	≧12g/dl	10≦ <12g/dl	8≦ <10g/dl	<8g/dl

Fig. 8 PHC color-coded triage categories for check-up items (the detail list can be found in [13])

The current PHC triage categories used to access patient morbidity is shown in Fig. 8. The threshold between each color category has been parameterized to reflect standard Bangladesh medical values. The ability to re-parameterized the risk stratification matrix and hence the color-coded risk triage make the Portable Health Clinic very flexible when deployed against different medical settings.

A future objective of the Portable Health Clinic is to personalize the color-coded risk triage to reflect age groups, regional differences, and eventually an individual genetic characteristic. It is not difficult to automatically adjust the triage boundaries and change the system's decision making policies but the chosen threshold need to be medically sound and supported by data.

3 Testing the Portable Health Clinic System

Portable Health Clinic service has been tested against a sample population from unreached communities in Bangladesh. The experiment design involved conducting validity experiments in urban, sub-urban and rural areas in Chandpur district and Shariatpur district between September 2012 and November 2013. The experiment environment consisted of the following facilities:

- Small call center in Dhaka (the capital city of Bangladesh) with two female and two male doctors, and one transcript writer;

- A portable health clinic briefcase with 12 diagnostic tools;
- Mobile health check-up team comprised two health assistants, three program assistants and one quality check officer. Patient data was captured from diagnostic tools wireless BAN and by manually inserted into our GramHealth database through a user-friendly web interface;
- Off-line version of GramHealth to store the patient's health profile and synchronize with the central server when the sufficient network bandwidth is available;
- On-line GramHealth software tools to process and store patient electronic health records.

The mobile healthcare team visited rural villages and urban factories, and set up a PHC health camp service for 1 week. There was a marketing campaign 2 weeks in advance and subjects could preregister to the event and book a time for a health checkup. The healthcare team revisited the areas every 2 months to repeat the measures. Patients who were previously measured as risky (orange and red) were asked to join the next health camp for a new check-up.

3.1 Results of the Mass NCD Mass Screening

The Portable Health Clinic medically screened 8,527 subjects in 2012 and 10,575 subjects in 2013 for the incidence of non-communicable diseases. There were 2,361 repeat subjects who were measured multiple times between 2012 and 2013. The Electronic Health Records are stored in the GramHealth database.

The median age of the sample population was 31 years. The average age was 36 years. The minimum age was 16 and maximum was 106. The relatively youthful average resulted in high prevalence of green and yellow results in all check-items. Out of the total sample population, 11,098 were identified not a risk for and 5,643 patients were identified as affected by for NCDs. The distribution of morbidity was strongly related to age (Fig. 9). The prevalence of orange (NCD affected) and red patients (NCD emergency) increased with age cohort. At the age 50 years nearly half of all tested within this cohort required a telehealth consultancy.

The increase of morbidity (defined as requiring a telehealth consultancy) with age cohort has profound impact on the future productivity of the unreached communities. It is expected that aging population will survive longer [9], however with over 50 % affected with NCDs, their quality of life will be impaired and there will be a increase demand on the poor resourced public healthcare system.

3.2 Results of the Telehealth Consultations

The primary research objective to use PHC screening and telehealth to identity low income people at risk and improve their medical condition. The results for

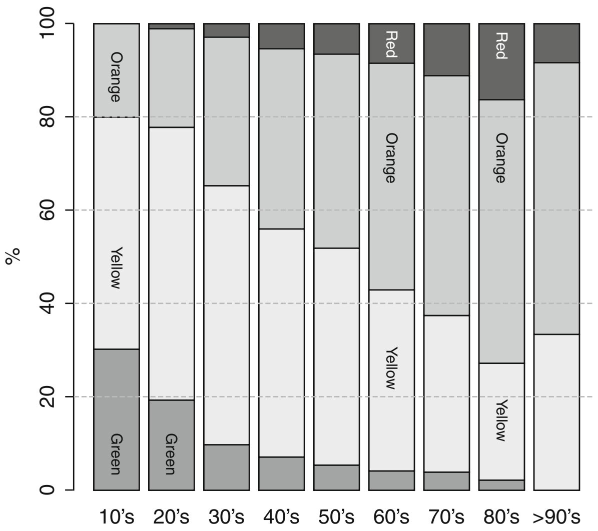


Fig. 9 Color-coded triage results against age distribution (N = 16,741)

2012 confirmed identification of NCDs and telehealth intervention led to a marked decrease in morbidity. In the first health checkup, 8,527 subjects were medically screened. PHC triage categorized 5,917 subjects as low risk (green 1,182 and yellow 4,735). Caution (yellow) subjects were given health guideline booklets. PHC triage identified 2,610 patients as high risk and the same amount of patient–doctor telehealth consultancies were established (orange 2,399 and red 211). The high-risk patients were prescribed medicine and were informed to attend the next PHC health care camp for a follow-up health check-up, scheduled 2 months later.

In the follow-up phase (after 2 months) 1,003 patients were re-examined using the same PHC workflow procedure. The purpose was to measure the effectiveness of remote consultation and the suitability of the prescribed drugs on the patient. The follow-up triage measurements showed a marked improvement in the patient previously categorized as orange or red 355 (35 %) patients had moved to the yellow category and 63 (6 %) patients had moved to the green zone. This left 585 (59 %) patients remaining in the orange and red categories (Fig. 10).

The results for 2013 also show a marked decrease in non-communicable morbidity following intervention by the portable health clinic service. In the first health checkup, 8,214 subjects were screened. PHC triage categorized 5,405 subjects as low risk (green 1,083 and yellow 4,322). Caution (yellow) subjects were given health guideline booklets. PHC triage identified 2,809 patients as high risk and the same amount of patient–doctor telehealth consultancies were established (orange 2,548 and red 261). The high-risk patients were prescribed medicine and were informed to attend the next PHC health care camp for a follow-up health check-up, scheduled 2 months later.

In the follow-up phase (after 2 months) 1,019 patients (follow-up data analysis in ongoing) were re-examined using the same PHC workflow procedure.

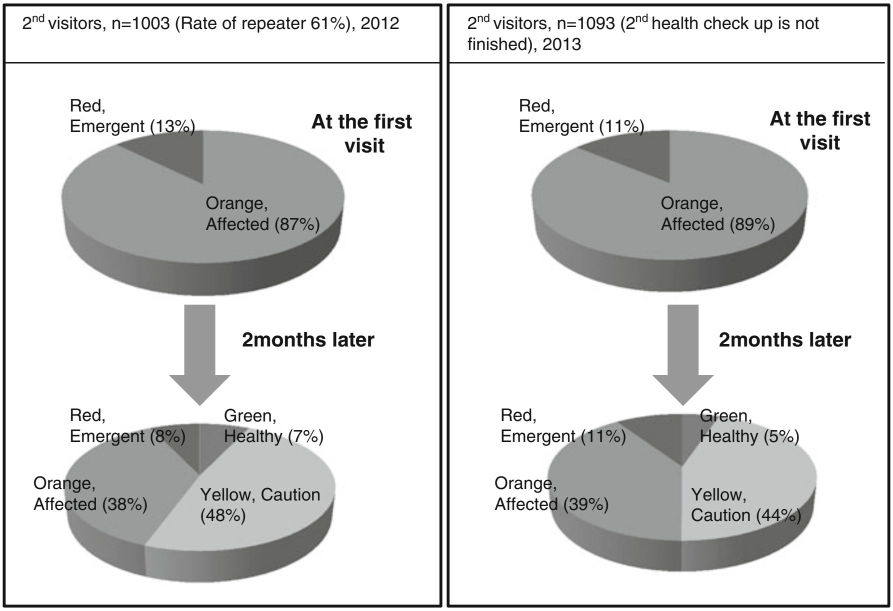


Fig. 10 Portable health clinic follow-up visitors—who were prescribed medicine and were tested 2 months later

The follow-up triage measurements showed a marked improvement in the patient previously categorized as orange or red 451 (44 %) patients had moved to the yellow category and 63 (6 %) patients had moved to the green zone. This left 505 (50 %) patients remaining in the orange and red categories (Fig. 10).

The PHC turnaround procedure appeared to accommodate large queues of patients within a short period of time. According to measurements taken by mobile healthcare team, the average turnaround time of a patient from taking a token to receiving a report and health booklet could be 10.25 min. The average turnaround time of a patient from taking a token to receiving a prescription including the telemedicine session with a remote doctor could be 18.25 min (see Fig. 11). These finding are very good particularly when the alternative for rural villagers is a long travel to the nearest hospital in a city and a long treatment wait. However, the turnaround times do not include queuing before registration or disruptions to Internet coverage.

The quality of experience is a subjective measure of the customer's experience with the PHC service. To ascertain the patient's perspective of service, several patients were surveyed. The survey categorized the opinion into what patients liked, disliked and what they thought needed to be improved. The general finding was that patients were satisfied that the service was brought to their village; it was affordable and based on advanced technology. The patients were not satisfied about queues before the entering the turnaround procedure. The patients expected a full healthcare service that included identifying infectious diseases and face-to-face consultancy

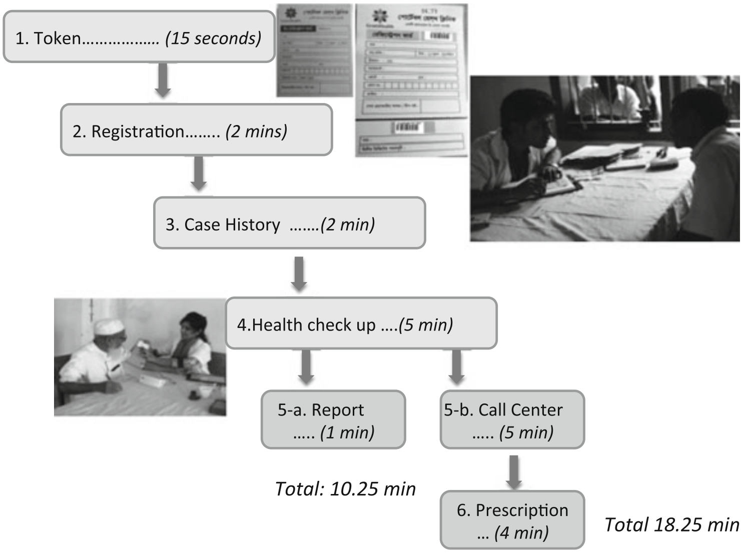

1. Token................. (15 seconds)
2. Registration........ (2 mins)
3. Case History ....(2 min)
4.Health check up ....(5 min)
5-a. Report ..... (1 min)
5-b. Call Center ..... (5 min)
Total: 10.25 min
6. Prescription ... (4 min)
Total 18.25 min

Fig. 11 Portable health clinic patient turnaround time

with a doctor on site even if they were not ill. This was not within the scope of the PHC which focuses on non-communicable diseases and uses telemedicine approach. The service improvements were based on what patients disliked (Table 4). For future PHC campaigns, screening the service desks and providing more privacy to the patients will be considered.

The doctors at the remote call center in Dhaka were also surveyed and their opinions recorded using the same criteria. The doctors appreciated the cost effectiveness and reachability of the PHC service but had reservations concerning the limited scope of the clinical investigation (only non-communicable measurements) and that the examination was remote and subject to connection interference. There improvements were based on adding more vital measurements and improving coverage (see Table 5).

4 Conclusions

The Portable Health Clinic is affordable technology that meets the requirements for aging population and unreached communities afflicted with non-communicable diseases. It is a service-oriented technology that brings eHealth with an optimized turnaround time to the customer. Each patient is measured locally with advanced medical devices and the system's stratification algorithm determines whether a remote doctor consultancy is required. Field experiments showed a decrease in morbidity after being diagnosed, prescribed medicine and monitored. The Portable

Table 4 Patients QoE of the PHC service

Patients liked	Patients disliked	What needs to be improved
This opportunity is very close to my location	Long queue for health check up	The total process is very complex. A simpler and faster check up method is required
Price is cheap	Take long time to get prescription after doctor consultancy	Waiting time should be reduced
I can get my health status immediately and get some instruction how to keep good health	PHC has limited number of sensors	More tests (checkup items) can be included
Time saving	Privacy is not maintained properly	Need privacy in all process
It is based on advance technology	Voice quality is not clear always during doctor consultancy	Instant/ faster report and prescription
Friendly and homely environment	No medicine facility with this system	Medicine facility
		One stop health service solution can be introduced (like.. referral face to face doctor, additional health checkup in other

Table 5 Doctors QoE of the PHC service

Doctors liked	Doctors disliked	What needs to be improved
Cost-effective for the rural people	Lack of some vital investigations e.g. S.creatinine, ECG, Hb% for all	Addition of vital investigations
Available at door step	Patients cannot be examined thoroughly (100 % as of face to face), so possibility of incomplete diagnosis	Addition and review of needed drugs and their availability in that locality that are to be prescribed
Good team work at site	Most of the time unavailability of good internet connection. So, doctors cannot hear and see patients and vice versa	Improvement of internet
Instant and correct investigations and heath case history	Sometimes difficult to get patients compliance due to local language, very old aged patient, and education	
Follow up can be done	Dependency on electronic equipments sometimes hampers total procedure in limited internet and electricity locations	

Health Clinic addresses three telehealth areas: it identifies risk patients and treats them in their own environment; it enables the healthcare worker and patient to overcome their resistance to technology by applying 'easy-to-understand' color-coded risk triage stratification human–computer interface; and it reduces morbidity and keeps patients out of hospital. Portable Health Clinic is attractive for low resource underserved areas such as unreached communities in Bangladesh. It is compact solution and the healthcare worker can carry the medical equipment to any village or doorstep of any house, and set up the clinic in a matter of minutes. The triage is adaptable and can be personalized according to the patient or disease requirements. New sensors can measuring logics can be added the sensor server. Two years of testing in Bangladesh has identified the need for NCD screening in underserved areas and results show a decrease in morbidity once the remote doctor has diagnosed and prescribe treatment. However, the Portable Health Clinic is a small-scale system, it does not identify infectious or respiratory diseases, and EHRs are not interoperable with other health systems. These issues will need to be addressed in the future versions.

Acknowledgments The authors would like to thank the KDDI Foundation and Kyushu University P&P program for their financial support for the research activities.

References

1. World Health Organization. Telemedicine, opportunities and developments in member states: report of the second global survey of eHealth. Geneva: World Health Organization; 2009.
2. Craig J, Patterson V. Introduction to the practice of telemedicine. J Telemed Telecare. 2005;11(1):3–9.
3. Cafazzo JA, Leonard K, Easty AC, Rossos PG, Chan CT. Bridging the self-care deficit gap: remote patient monitoring and hospital at home. In: Electronic Healthcare First International Conference, eHealth 2008; 2009 Feb 14.
4. Wilson J, Junger G. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.
5. Currell R, et al. Telemedicine versus face-to-face patient care: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2000; (2):CD002098.
6. Stanberry B. Legal and ethical aspects of telemedicine. J Telemed Telecare. 2006;12(4): 166–75.
7. Telehealth Report 2014. A dedicated study of telehealth that provides detailed analysis of the world market. www.ihs.com.
8. UK department of health whole systems demonstrators: an overview of telecare and telehealth. 11.06.2009; 2013. http://www.theguardian.com/healthcare-network/2013/feb/27/telehealth-not-effective-long-term-conditions-study.
9. Wootton R. Telemedicine support for the developing world. J Telemed Telecare. 2005;11(8):384–90.
10. Ahmed A, Inoue S, Kai E, Nakashima N, Nohara Y. Portable health clinic: a pervasive way to serve the unreached community for preventive healthcare. In: Proceedings of the 15th International Conference on Human-Computer Interaction (HCI 2013); 2013.
11. Kato S. A study on implementing a portable clinic based on social needs. Undergraduate Thesis, Kyushu University; 2012.

12. Naoki N, Nohara Y, Ahmed A, Kuroda M, Inoue S, Ghosh P, Islam R, Hiramatsu T, Kobayashi K, Inoguchi T, Kitsuregawa M. An affordable, usable and sustainable preventive healthcare system for unreached people in Bangladesh. Report; 2013.
13. Nohara Y, Kai E, Ghosh P, Islam R, Ahmed A, Kuroda M, Inoue S, Hiramatsu T, Kimura M, Shimizu S, Kobayashi K, Baba Y, Kashima H, Tsuda K, Sugiyama M, Blondel M, Ueda N, Kitsuregawa M, Nakashima N. Health checkup and telemedical intervention program for preventive medicine in developing countries: verification study. J Med Internet Res. 2015;17(1).