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

Smart Sensors and Systems

Innovations for Medical, Environmental, and IoT Applications

Springer

Smart Sensors and Systems

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

Smart Sensors and Systems

Innovations for Medical, Environmental, and IoT Applications

Springer

Editors
Chong-Min Kyung
#310 IT Convergence Building (N1)
Center for Integrated Smart Sensors
Yuseong-gu, Daejeon
Korea (Republic of)

Yongpan Liu
Circuits and Systems Division
Tsinghua University
Beijing, China

Hiroto Yasuura
System LSI Research Center
Kyushu University
Fukuoka, Japan

Youn-Long Lin
National Tsing Hua University
Hsinchu, Taiwan

ISBN 978-3-319-33200-0 ISBN 978-3-319-33201-7 (eBook)
DOI 10.1007/978-3-319-33201-7

Library of Congress Control Number: 2016955366

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

There are some technical keywords pointing to the future such as IoT (Internet of Things), virtual reality, big data/deep learning, 3D printing, energy harvesting, wireless transfer, etc. Smart sensor being responsible for scavenging data underneath has some overlap with each of these future keywords. For more than 3–4 decades, the growth of the semiconductor industry has been described roughly according to Moore's law. As the demand from the market and the way it is responded to become more and more diverse, the relative portion of CPU, memory, PCs and tablets dwindles, while growth is expected in IoT, including vehicular, embedded medical, and consumer IoT based on AR/VR. Led by CIS (CMOS image sensor), a huge variety of sensors are emerging. Smart sensor has many challenges such as reducing power, size, and cost while improving the performance. The future of IoT in the vehicular, game, biomedical, and environmental market depends on how these challenges are met in the development of smart sensors.

This book is the second in a series of a collection of selected papers presented in the AWSSS (Asian Workshop on Smart Sensors and Systems) where important research results on smart sensors and systems in Asia are annually reported. The previous AWSSS workshops were held in Jeju (2013), Hualien (2014), Karatsu (2015), and Beijing (2016). Smart sensor research encompasses materials/devices, analog and digital circuit platform, systems, applications and services and, therefore, requires a close interaction among researchers in different disciplinary areas. This book reports on 18 up-to-date research activities being made in the design of smart sensors and systems in Asia.

The book is composed of five parts. Part I consists of two chapters and covers materials and structural platforms for smart sensors. Chapter 1 discusses on the current state of the biomimetic materials based on detecting mechanical stimulation, liquid flow, tactile, light, olfactory, chemical, and cellulose-, collagen-, and virus-based structures as sensor structural platforms. Chapter 2 describes a CMOS lab-on-a-chip as a structural platform for personalized biomedical diagnosis.

Part II describes various circuit platforms and consists of four chapters. Chapter 3 describes the design of a variety of extremely low-power, mm-scale

IoT devices mainly for biomedical applications from the circuit and systems design perspective. Chapter 4 explains various smart sensor microsystems. Chapter 5 describes an energy-efficient approximate computing applied for smart cameras using RRAM crossbars. Chapter 6 is on energy-aware management of flash memory which is very important for implantable sensors where battery lifetime is critical.

Part III has two special chapters; Chapter 7 is on biomimetic camera design based on the intuition from the compound eyes of insects and arthropods. Recently with the growing interest in unmanned driving, ADAS (Advanced Driver Assistance System) has opened a huge market, and we have Chap. 8 reporting on the recent research progress on the ADAS.

Part IV consists of five chapters on biomedical and health monitoring using various sensor systems. The first one (Chap. 9) reports on implantable optical neural interface as a strong candidate for replacing electrical interface schemes. Chapter 10 describes a real-time monitoring of brain and cardiac activity using DBS (deep brain stimulation) and ECG (electrocardiogram), respectively. Chapter 11 covers a wide area including various SoCs for micro-gas chromatography SoC for selectively measuring each VOC (volatile organic compound), release-on-demand drug delivery SoC, a batteryless remote-controlled locomotive SoC, etc. Chapter 12 is on flexible materials using UV sensor as an example. Chapter 13 describes a urine sensing system based on SoC.

Finally, we have five chapters dedicated to big data, which has become a "big" word recently. The big data will certainly become very critical in the future and prominent in various areas through deriving important conclusions from all sensor data captured. Using indoor GPS (global positioning system) as an example, Chapter 14 reports on how the accurate (indoor position) data can be captured. Chapter 15 is a theoretical framework for dealing with various smart objects. Final two chapters report on the application of such big data for agriculture, i.e., cattle management and hospital and patient management.

We sincerely thank each contributor for summarizing their wonderful research as a part of this volume and truly hope this volume contributes to the booming of research, industry, and economy related with smart sensors. Voila! Smart sensors are coming but in different costumes.

Daejeon, Korea — Chong-Min Kyung
Fukuoka, Japan — Hiroto Yasuura
Beijing, China — Yongpan Liu
Hsinchu, Taiwan — Youn-Long Lin
March 2016

Contents

Part I
Materials and Structural Platform for Smart Sensors

Chapter 1
Biomimetic Materials and Structures for Sensor Applications

Do Hoon Lee, Wonbin Song, and Byung Yang Lee

Abstract Diverse biological tissues and structures that often exhibit remarkable physical and chemical properties can be found throughout nature. Starting from very few and simple building blocks such as collagen fibrils, nature effortlessly makes hierarchical and complex structures which are often hard to imitate with the current top-down microfabrication techniques. With the recent development of diverse assembly methods of nanobiomaterials, we have started to build biomimetic structures with diverse optical, mechanical, and electrical properties using bottom-up approaches. The properties of such biomimetic materials, when exposed to certain physical or chemical stimuli, sometimes change enough and may be utilized for sensing applications. For example, some filamentous viruses can be assembled into colorful films on solid substrates, the colors of which can change when exposed to organic solvents and volatile organic compounds. These same films, when applied with mechanical pressure, can exhibit piezoelectric properties, where mechanical pressure can be transduced to electrical signals, allowing the utilization of these structures as mechanical force sensors. In this chapter, we will discuss the current state of the biomimetic materials and structures for sensor applications, giving emphasis on hierarchical structures based on fiber building blocks.

Keywords Biomimetic sensor • Olfactory sensor • Cilia • Self-assembly • Structural color • Compound eye • E-nose • E-tongue • Cellulose • Collagen • Bacteriophage • Magnetoreception • Electroreception

1.1 Introduction

Living creatures such as insects, animals, and plants show functions and materials within their bodies that are directly related to their survival and prosperity. Most natural materials are complex composites that are usually built from very basic building blocks. These building blocks are generally secreted as proteins or fibers

D.H. Lee • W. Song • B.Y. Lee (✉)
School of Mechanical Engineering, Korea University, Anam-ro 145,
Seoul 02841, South Korea
e-mail: blee@korea.ac.kr

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_1

and then self-assembled to build diverse structures in hierarchical integrations. These complex structures, which have advanced throughout the years of evolution, are inspiring scientists and engineers in the design of novel materials [1]. The field of biomimetics is directly aimed at learning or being inspired from nature and engineering new materials and structures with novel or advanced functionalities compared to nature. Thus, it has enabled the development of material science in this field [2].

However, the field of biomimetic materials or structures for sensor applications is relatively in its early stage. This can be attributed to several reasons. First, utilizing biomimetic tissues as sensors requires further development of responsive structures and transducers in order to detect the external stimuli. Another aspect is that the mechanism of sensory systems such as electroreception or magnetoreception observed in nature is not yet fully understood, and thus mimicking them is difficult. Even if the mechanism is well known, mimicking the sensory systems can be a difficult engineering task. The usual case is that engineers have difficulty mimicking something that is so simply done in nature. Needless to say, nature has developed and optimized an incredible variety of sensors for navigation, spatial orientation, prey, object detection, etc. providing engineers with new ideas for improvement in current technology, new sensor technology, and potential sensor miniaturization. Indeed, several reviews have been published in regard to biomimetic sensors [3–6]. The discussion here will be different to previous reviews on biomimetic materials in that it will be focused on the biomimetic structures and materials as sensors or sensor platforms instead of the different kinds of biomimetic sensors available in the literature. The goal of this chapter is to present the recent research trends in the field of biomimetic structures and materials for sensor applications.

For this purpose, we will follow a different classification of biomimetic sensors than before. For example, Stroble et al., in their review of biomimetic sensor technology, classified the biomimetic sensors according to the external stimuli (acoustic, biological, chemical, electric, optical, magnetic, mechanical, etc.) or according to the mimicked sensory extroreceptor types (chemoreceptor, electroreceptor, magnetoreceptor, etc.) [7]. These two categories would serve well if we were interested in only the stimulus and interaction that we want to mimic. However in this chapter, we will classify first according to functional structures and then present the recently reported strategies and different engineering materials and approaches to enable a peculiar sensing capability. This will enable us to incorporate a wider scope of biomimetic materials and structures for sensors.

1.2 Structures for the Detection of Mechanical Stimuli

Among the many biomimetic sensors, mechanical sensors that can detect mechanical stimuli such as air and liquid flow, strain, and vibration are the most abundant. This is natural considering that a living creature's ability to detect prey and predators in the environment is directly related to its chance of survival.

The physiological study of animals and insects gives us extensive understanding of the mechanism by which the sensory systems of these living creatures survive. In nature, many insects and arthropods exhibit structures optimized for the detection of air flow. The ability to detect minute perturbance in the surrounding environment is directly related to its survival against predators and enemies. These structures have usually small hair-like structures, and as such, we will discuss hair-like structures for flow sensors, various tactile sensors, and then do a brief discussion on e-skins.

1.2.1 Hair-Like Structures for Flow Detection

Many animals such as fish and spiders are able to detect the change in the flow of air or water in their surroundings in order to survive within challenging environments. Also, many insects or arthropods have hairs on their bodies that are often packed at high density. In the case of wood crickets, the density of airflow-sensing hairs can reach values higher than 400 hairs mm^{-2} [8, 9]. These hairs are connected to hair cell sensors which are responsive to flow, vibration, touch, acoustic vibration, and gravitational force.

This fascinating ability of animals to detect minute vibrations and forces has attracted the engineering community to imitate those functions and structures [8, 10]. The most direct method of imitating this structure is by utilizing microfabrication techniques such as the microelectromechanical system (MEMS) technology. As a matter of fact, biomimetic sensors with MEMS structures is one of the most developed fields of biomimetic sensors. The materials used in MEMS-based biomimetic sensors include polymers, inorganic materials, and composite structures. The structures are usually arrays of high-aspect-ratio pillar-like structures combined with electronic transducers. This structure tries to imitate the sensory organs that many insects or arthropods have as flow-sensitive hair structures and nerve receptors generating the nerve signal with the mechanical stimuli. Several reviews can be found on artificial hair cell based sensors. For example, an early work by Gijs et al. shows arrays of MEMS fabricated flow sensors inspired by the acoustic flow-sensitive hairs found on the cerci of crickets. The hairs consist of up to 1 mm long SU-8 structures mounted on suspended membranes with normal translational and rotational degrees of freedom. Electrodes on the membrane and on the substrate form variable capacitors, allowing capacitive read-out [11].

A recent MEMS-based biomimetic structure was reported by Hein et al., where arrays of magnetic cilia were shown for the low-power detection of vibration and flow. They utilized an anodized aluminum oxide template to electrodeposit cobalt cilia. The motion of arrays of Co cilia was then detected using magnetic sensors. The flow sensors were tested in a microfluidic channel. They showed the ability to detect flows from 0.5 to 6 ml/min with a signal to noise ratio of 44 using only 140 μW of power and no amplification. The vibration sensors were tested using a shake table in the low earthquake-like frequency range of 1–5 Hz. The vibration

response was a mW signal at twice the frequency of the shake table [12]. Polymers can also be utilized to build cilia-like structures for flow detection. Polymers such as polypyrrole have recently been demonstrated to be good materials for flow sensors [13]. Polymeric nanowires of polypyrrole have been implemented as artificial cilia on giant-magneto-resistive multilayer sensors. The arrays were tagged with a magnetic material, the stray field of which changes relative to the underlying sensor as a consequence of mechanical stimuli delivered by a piezoactuator.

Another kind of sensor would be those imitating flow sensors when immersed in liquids. This imitates the sensory organs of aquatic creatures such as fish. Most fish have the capability of sensing flows and nearby movements even in dark or murky conditions by using the lateral line organs. This enables them to perform a variety of underwater activities such as localizing prey, avoiding predators, navigating in narrow spaces, and schooling. Nguyen et al. demonstrated a MEMS-based artificial lateral line using an array of MEMS flow sensors [14]. The signals collected via the artificial lateral line were processed by an adaptive beam forming algorithm developed from Capon's method. The system produced 3D images of source locations for different hydrodynamic activities including the vibration of a dipole source and the movement of a tail-flicking crayfish.

Meanwhile, animals like bullfrogs can adjust the stiffness of the hair cells with the adaptive mechanism of mechanical relaxation of hair bundles. In imitation of this function, MEMS structures with adaptive hair cell stiffness have been proposed by using electrostatic spring softening and hardening techniques. [15] This enables the adjustment of the sensor's responsivity, bandwidth, threshold, and (thermal) noise level. In addition, the modulation of the torsional stiffness of the sensory system enhances selectivity to arbitrary flow frequencies and simultaneously achieves significant amplification of the sensor response.

1.2.2 Structures for Tactile Sensing

The sense of touch is related to the detection of diverse physical phenomena such as pressure, shear force, and temperature. The biomimetic approach towards tactile (touch) sensors and strain sensors shows diverse pathways. Other than traditional MEMS-based sensors, several new strategies are being examined recently. Tee et al. demonstrated the utilization of flexible organic transistor circuits that translates pressure directly into digital frequency signals. The output frequency ranges between 0 and 200 Hz, with a sublinear response to increasing force stimuli that mimics the slow-adapting skin mechanoreceptors. The output of the sensors was further used to stimulate optogenetically engineered mouse somatosensory neurons of mouse cortex in vitro, achieving stimulated pulses in accordance with pressure levels. This work shows that the interfacing of smart electronics with real nerve systems is in continuous progress [16]. Another pressure sensor was shown, where reduced graphene oxide was used to implement mechanosensitive papers. Sheng et al. showed the formation of bubble-decorated honey-comb structured papers for

the detection of pressure with a sensitivity of 161.6 kPa^{-1} at a strain less than 4 %. The paper showed enhanced sensitivity compared to the structures without the bubbles [17]. The pressure detection capability of plant leaves was demonstrated by Su et al. They mimicked mimosa, a plant that can close their leaves under external stimuli. They utilized real mimosa leaves as templates from which corrugated polymer surfaces were cured. These polymer leaves were deposited with metal films. The sensor showed a sensitivity of 50.17 kPa^{-1}, quick responding time (<20 ms), and durable stability (negligible loading–unloading signal changes over 10,000 cycles) [18]. A polymer based tactile sensor was demonstrated to measure normal pressure and shear and torsional force, and thus they showed high sensitivity and a wide dynamic range (Fig. 1.1) [19]. The device was based on two interlocked arrays of high-aspect-ratio Pt-coated polymeric nanofibers that are supported on thin polydimethylsiloxane layers. When different sensing stimuli were applied, the degree of interconnection and the electrical resistance of the sensor changed in a reversible, directional manner with specific, discernible strain-gauge factors. Also, the sensor was used to monitor human heartbeat and the bouncing of water droplets on a superhydrophobic surface. Interlocked ZnO nanowire arrays were proposed by Ha et al. A hierarchical micro- and nanostructured ZnO nanowire (NW) array in an interlocked geometry was used to detect both static and dynamic tactile stimuli through piezoresistive and piezoelectric transduction modes, respectively [20].

Inspired by insects such as spiders, which detect vibration through crack-shaped slit organs, engineers are trying to imitate these slit structures for the detection of small strains and vibrations [21]. Kang et al. demonstrated these strain sensors based on nanoscale crack junctions (Fig. 1.2) [22]. The sensors are sensitive to strain (with a gauge factor of over 2000 in the 0–2 % strain range) and vibration (with the ability to detect amplitudes of approximately 10 nm). The device is reversible, reproducible, durable, mechanically flexible, and can thus be easily mounted on human skin as an electronic multipixel array. The ultrahigh mechanosensitivity is attributed to the disconnection–reconnection process undergone by the zip-like nanoscale crack junctions under strain or vibration.

Meanwhile, many attempts to imitate the skin of humans have been reported [23, 24]. Human skin is a remarkable organ that consists of a network of sensors that relay information regarding tactile and thermal stimuli to the brain. Among many applications, the electronic sensor networks and devices inspired by the human skin will be applied to diverse fields such as robotics and biomimetic prosthetics [24]. A recent study of the human skin-inspired e-skin was demonstrated by Park et al. [25]. They demonstrated multimodal e-skins based on flexible and microstructured ferroelectric films, which enhance the detection and discrimination of multiple spatiotemporal tactile stimuli such as static and dynamic pressure, temperature, and vibration. The piezoelectric and pyroelectric properties that detect dynamic touch and temperature were realized by using piezoelectric and pyroelectric materials of ferroelectric polymer composites composed of poly (vinylidene fluoride) (PVDF) and reduced graphene oxide (rGO).

One of the problems that flexible tactile sensors have is that the sensing property decays gradually with time. This is in contrast to real human skin, where the functions

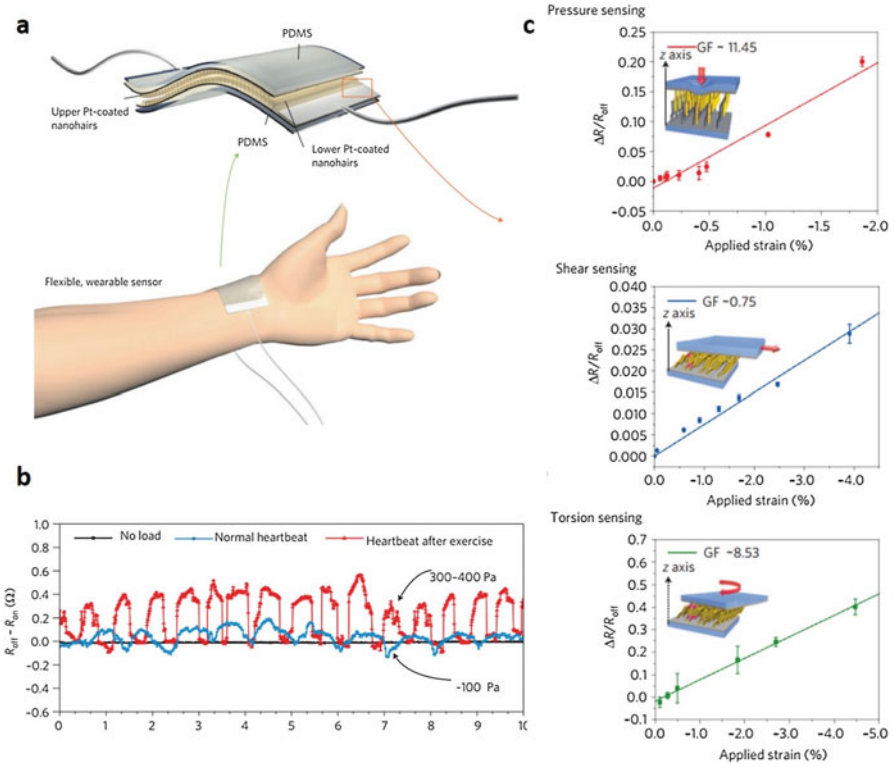


Fig. 1.1 Strain sensor based on interlocking nanofibers. (**a**) Schematic of the assembly and operation of a flexible sensor layer sandwiched between thin PDMS supports (~500 μm thickness each). (**b**) Measurement of the physical force of a heartbeat under normal (~60 beats min^{-1} with an average intensity of ~100 Pa) and exercise conditions (~100 beats min^{-1} with an average intensity of 300–400 Pa). (**c**) Ratio ($\Delta R/R_{off}$) vs applied strain, the slope of which yields the corresponding piezoresistive GF, defined as $GF = (\Delta R/R)/\varepsilon$: ~11.45 for pressure, ~0.75 for shear and ~8.53 for torsion. (Reprinted with permission from [19])

and mechanical and sensing properties are retained and restored in case of mild damage. Tee et al. proposed a composite material composed of a supramolecular organic polymer with embedded nickel nanostructured microparticles, which shows mechanical and electrical self-healing properties at ambient conditions [26]. They also showed that the material is pressure and flexion-sensitive, and therefore suitable for electronic skin applications. The electrical conductivity was tuned by varying the amount of nickel particles up to values of 40 S cm^{-1}. On rupture, the initial conductivity is repeatedly restored with ~90 % efficiency after 15 s healing time, and the mechanical properties are completely restored after ~10 min. Further attempt for self-healing was demonstrated by using a polymer and reduced graphene oxide laminate structures for sensing pressure [27]. The structure consisted of piezoelectric polymer layers sandwiched between two self-healing electrodes composed of poly(N,N-dimethylacrylamide)-poly(vinyl alcohol)/reduced graphene oxide (PDMAA-PVA/rGO) hybrid as the electrode film. Real skin can detect mechanical

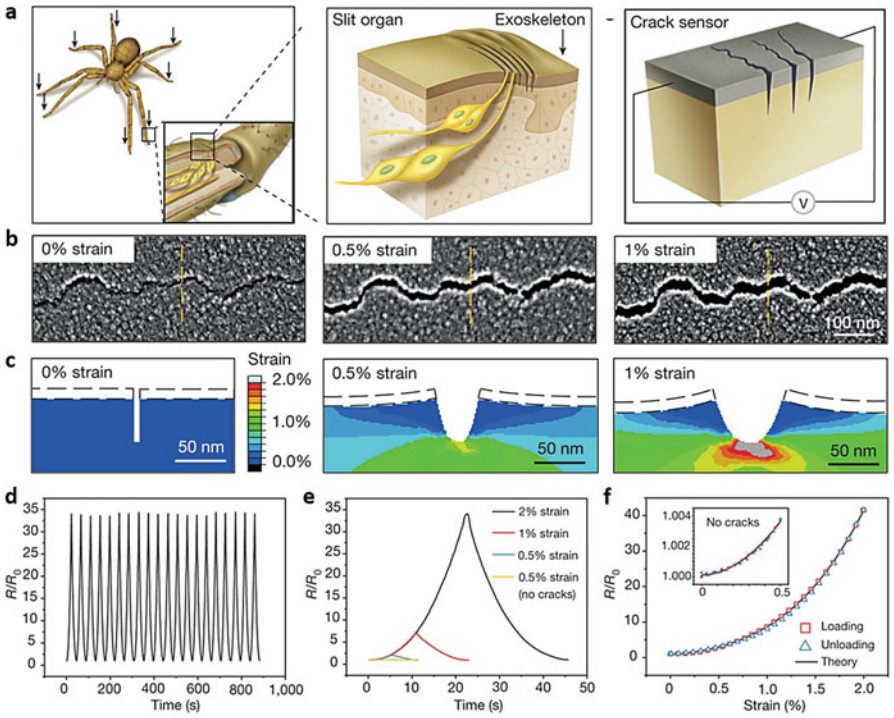


Fig. 1.2 Spider-inspired strain sensor. (**a**) Spider's slits connected to the nervous system to monitor vibrations. Right side of a scheme is illustration of the crack-based sensor and its measurement. (**b**) SEM images of the zip-like crack junctions for different applied strains: 0 % (*left*), 0.5 % (*middle*) and 1 % (*right*). (**c**) Finite-element method modelling results of crack interfacial deformation by 0 % (*left*), 0.5 % (*middle*) and 1 % (*right*) strain. (**d**) The normalized resistance measured at a strain sweep rate of 1 mm min^{-1}. (**e**) Reversible loading–unloading behavior for various final strains. (**f**) Resistance at the slowest loading–unloading rate of 0.1 mm min^{-1}, compared with the theoretical fit. Inset, results for no cracks. (Reprinted with permission from [22])

and temperature change. Kim et al. recently demonstrated an e-skin, where they integrated various sensors with flexible electronic strategies to develop a smart prosthetic skin based on single crystalline silicon nanoribbons for the detection of strain, pressure, and temperature (Fig. 1.3) [28]. Other humidity sensors, electroresistive heaters, and stretchable multi-electrode arrays for nerve stimulation were integrated together.

Other aspect of the human skin is the perception of "texture." In humans, the tactile perception of fine textures (spatial scale <200 μm) is mediated by skin vibrations generated as the finger scans the surface. When the sensor surface is patterned with parallel ridges mimicking the fingerprints, the spectrum of vibrations elicited by randomly textured substrates is dominated by one frequency set by the ratio of the scanning speed to the inter-ridge distance. For the human touch, this frequency falls within the optimal range of sensitivity of Pacinian afferents, which

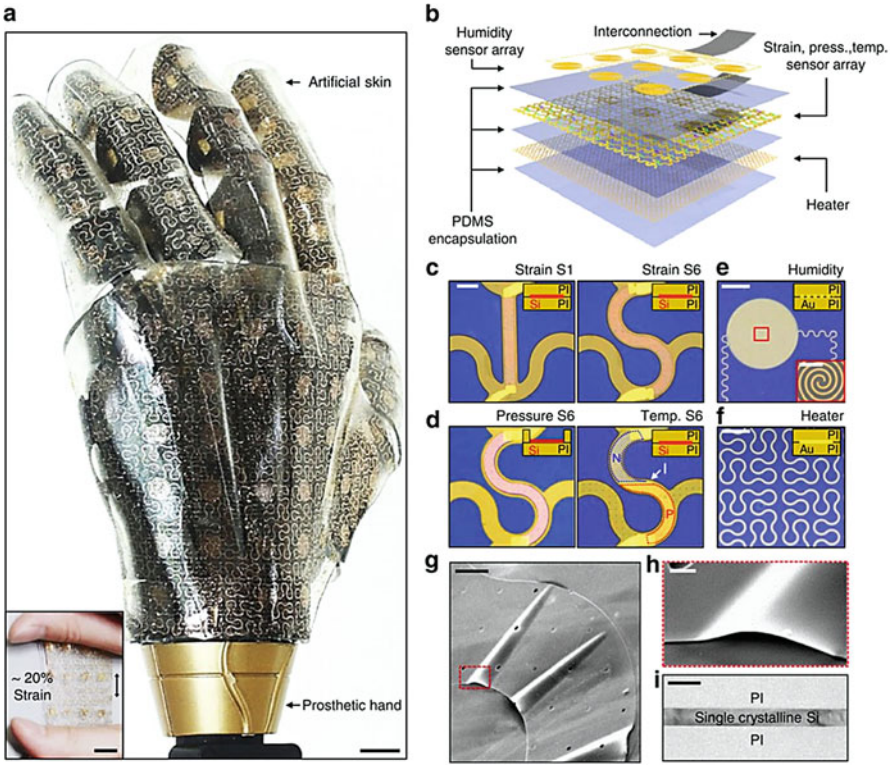


Fig. 1.3 Smart artificial skin with integrated stretchable sensors and actuators. (**a**) Smart skin covering a prosthetic hand. Scale bar, 1 cm. The inset shows the artificial skin stretched by 20 %. Scale bar, 1 cm. (**b**) Artificial skin comprised of six stacked layers. (**c**) Representative microscopic images of SiNR strain gauge. (**d**) SiNR pressure sensor and temperature sensor. (**e**) Humidity sensor. Scale bar, 2 mm. *Bottom right*: magnified view of the central area. Scale bar, 0.5 mm. (**f**) Electroresistive heater. Scale bar, 4 mm. (**g**) SEM image of the SiNR transferred on the silicon oxide substrate. Scale bar, 20 mm. (**h**) The magnified view of wrinkled SiNR. Scale bar, 2 mm. (**i**) A cross-sectional TEM image of the strain gauge, showing that the SiNR encapsulated with PI layers is located at the neutral mechanical plane. Scale bar, 200 nm. (Reprinted with permission from [28])

mediate the coding of fine textures. Thus, fingerprints may perform spectral selection and amplification of tactile information that facilitate its process by specific mechanoreceptors [29]. Zhang et al. demonstrated artificial epidermal ridges made of polydimethylsiloxane (PDMS) combined with micro-fabricated metal strain gauge arrays. The aspect ratio of the artificial epidermal ridges was optimized using material stability calculations and finite-element method (FEM) simulations, and thus the optimal structure obtained was 400 μm in width and 110 μm in height. Experiments showed that the strain gauges were 1.8 times more sensitive than those of a tactile sensor without ridges [30].

1.2.3 Structures for Light Detection

The eyes of animals and insects have been extensively studied by the engineering community. The light-sensing organs are one of the most studied sensory systems in animals and insects. The eyes of animals and humans are usually camera-type eyes, having the structure of a single lens that focuses images onto a light-sensitive retina. In the case of insects, we can find compound eyes, which are composed of multiple light-sensing elements. Furthermore, the compound eyes can be classified into two types: superposition and apposition. In the case of apposition compound eyes, each light-sensing element or lens is optically isolated from one another, with each providing part of the total scene. In case of superposition compound eyes, the individual light-sensing elements are not optically isolated [31]. In terms of biomimicking, the apposition compound structure is much easier to implement, and as such, most of the reports are concentrated on those kinds of eyes. Recently, Song et al. demonstrated an easy way of fabricating hemispherical, compound apposition layouts of arthropod eyes [32]. The devices combined elastomeric compound optical elements with deformable arrays of thin silicon photodetectors into integrated sheets. These sheets can be elastically transformed from the planar geometries to hemispherical shapes for the integration into apposition cameras. The surface of the hemisphere is populated by imaging elements (artificial ommatidia), which are comparable in numbers (180) to those of the eyes of fire ants (*Solenopsis fugax*) and bark beetles (*Hylastes nigrinus*).

Floreano et al. demonstrated a fruit fly-inspired compound eye structure [33]. The structure consists of three planar layers of separately produced arrays, a microlens array, a neuromorphic photodetector array, and a flexible printed circuit board that are stacked, cut, and curved to produce a mechanically flexible imager. They demonstrated an artificial compound eye with a hemispherical field of view with an embedded and programmable low-power signal processing, high temporal resolution, and local adaptation to illumination. The prototyped artificial compound eye possesses several characteristics similar to the eye of the fruit fly Drosophila and other arthropod species.

1.3 Biomimetic Structures for Biochemical Sensing

Through millions of years of evolution, animals and insects have developed sensing organs to smell and taste substances related to food, enemy, and surrounding information. The sensory systems related with smell and taste are basically receptor-based detectors connected to nerve systems, where the receptor binding event is translated to nerve impulse that is finally relayed to the brain, where complex data processing is performed for specific taste and smell recognition. Receptors in these sensors are not of high specificity, but of broad response. However, the animal senses utilize a combination of different receptors to allow

combinatorial processing of information to discriminate thousands of targets. From these principles, diverse biomimetic structures have been proposed by the engineering society. In particular, the concepts of the electronic nose (e-nose) and electronic tongue (e-tongue) are driving engineers and scientists to develop new materials and structures for biochemical sensors [23]. Here, we will briefly discuss the recent research trends of biomimetic structures for the detection of biochemical target molecules.

1.3.1 Structures for Biomimetic Olfactory Senses

Biomimetic olfactory sensors including e-noses have received considerable attention in the field of sensor technology. Recent applications of electronic nose technologies have come through advances in sensor design, material improvements, software innovations, and progress in microcircuitry design and systems integration. Electronic noses have provided a plethora of benefits to a variety of commercial industries, including the agricultural, biomedical, cosmetics, environmental, food, manufacturing, military, pharmaceutical, regulatory, and various scientific research fields [34].

Inspired by the structure of the sensing system of some moth species that can detect single pheromone molecules, Spitzer et al. demonstrated a microcantilever-based sensor of ppt-level detection limit to trinitrotoluene [35]. The moth antenna was mimicked by a first preparing highly ordered TiO_2-NT arrays on cantilevers with a micrometric surface area. To achieve these architectures, we developed a two-step procedure consisting of the physical vapor deposition (PVD) of a dense layer of Ti metal onto a silicon microcantilever followed by its anodization in a fluoride-containing electrolyte. Another example of bio-templated structure for sensing applications was demonstrated by Zhang et al. using the natural bristles of a butterfly (*Papilio maacki*) to fabricate gas-sensing materials consisted of single porous SnO_2 microtubes (SPSMs) [36]. Electrodes were contacted on the two ends of the sensing layer. The sensor was highly sensitive to around ppm-level concentrations of ammonia, formaldehyde, and ethanol at room temperature. The average response and recovery times in the dynamic performance were only about 3 and 30 s.

Peptides can be self-assembled into nanofibers with sensing capabilities. Wang et al. reported the utilization of metalized peptide nanofibers as sensing materials [37]. Artificial peptide nanofibers were created with a special designed peptide molecule that contains complex motif sequences and then further metallized to synthesize nanofiber-based silver nanowires. A hybrid nanomaterial was obtained by assembling the prepared silver nanowires on graphene nanosheets. These sheets were utilized in non-enzymatic electrochemical detection of hydrogen peroxide.

Meanwhile, surface plasmon techniques combined with odorant binding proteins (OBPs) were investigated by Zhang et al. [38]. They prepared a sensor based on localized surface plasmon resonance (LSPR) to monitor binding of small odorant molecules to OBPs from honeybees. Other than floral odorants, this

sensor also showed response to nitro-compounds such as 2,4,6-trinitrotoluene, 2,4-dinitrotoluene, and 3-mononitrotoluene.

The majority of olfactory sensors have some kind of transducer combined with olfactory receptor (OR) proteins. Lee et al. mimicked the human olfaction mechanism by using carboxylated polypyrrole nanotubes (CPNTs) functionalized with human OR protein [39]. The e-nose was able to detect gaseous odorants at a concentration as low as 0.02 parts-per-trillion (ppt), which was comparable to a highly trained, human expert's nose. Goldsmith et al. coupled ORs with carbon nanotube transistors [40]. The resulting devices transduced signals associated with odorant binding to ORs in the gas phase under ambient conditions and showed responses that are in excellent agreement with results from established assays for OR–ligand binding. This work shows that the combination of natural OR with artificial electronic transducer such as carbon nanotube transistors can result in effective bioelectronic noses comparable to biological olfactory systems. Another significant application of carbon nanotube transistors was shown by Kim et al. [41]. They demonstrated a sensor for the detection of an explosive chemical, trinitrotoluene (TNT). The receptors were identified by an evolutionary selection method, phage display, and the receptors were covalently linked to a polydiacetylene (PDA) polymer layer integrated with the single-walled carbon nanotube field-effect transistors. Selective binding events between the TNT molecules and phage display-derived TNT receptors were effectively transduced to sensitive SWNT-FET conductance sensors through the PDA coating layers. The resulting sensors exhibited 1 fM sensitivity toward TNT in real time, with selectivity over various similar aromatic compounds. Kwon et al. showed an artificial multiplexed superbioelectronic nose (MSB-nose) that mimics the human olfactory sensory system, using highly uniform graphene micropatterns (GMs) that were conjugated with two different ORs [42]. The graphene patterns served as liquid-ion gated field-effect transistors. Field-induced signals from the MSB-nose showed minimum detectable level of 0.1 fM towards the target odorants.

Guo et al. demonstrated recently an effective way of building a free-standing biomimetic sensor by covalently bonding RGD-peptide on the surface of pyrenebutyric acid-functionalized graphene film to allow real-time detection of nitric oxide, an important signal yet short-life molecule released from the attached human endothelial cells under drug stimulations [43].

1.3.2 Structures for Biomimetic Tasting Senses

Song et al. developed a human taste receptor protein hTAS2R38, and then they functionalized a carboxylated polypyrrole nanotube (CPNT)-field effect transistor with the receptor protein to build an e-tongue with high sensitivity and selectivity [44]. Taster type (PAV) and nontaster type (AVI) hTAS2R38s were expressed in Escherichia coli and immobilized on a CPNT-FET sensor platform. Among the various tastants examined, PAV-CPNT-FET exclusively responded to target

bitterness compounds, phenylthiocarbamide (PTC) and propylthiouracil (PROP), with high sensitivity at concentrations as low as 1 fM. However, no significant changes were observed in the AVI-CPNT-FET in response to the target bitter tastants. This e-tongue exhibited different bitter-taste perception of compounds containing thiourea (N–C═S) moieties such as PTC, PROP, and antithyroid toxin in vegetables, which corresponded to the haplotype of hTAS2R38 immobilized on CPNTs. This correlation with the type of receptor is very similar to the human taste system. Thus, the artificial taste sensor developed in this study allowed for the efficient detection of target tastants in mixture and real food sample with a human-like performance and high sensitivity. Meanwhile, Lee et al. demonstrated a floating electrode-based bioelectronic tongue mimicking insect taste systems for the detection and discrimination of umami substances [45]. Here, carbon nanotube field-effect transistors with floating electrodes were hybridized with nanovesicles containing honeybee umami taste receptor, gustatory receptor 10 of *Apis mellifera* (AmGr10). This sensor differentiated between l-monosodium glutamate (MSG), best-known umami tastant, and non-umami substances with a high sensitivity and selectivity.

1.4 Optical Sensors

According to Ge et al., we can find many examples of structural colors in nature like the wings of a butterfly, the cuticles of beetles, fish, and the feathers of a bird. [46] These colors have attracted considerable attention in various research fields due to its beauty and potential for utilization as optical sensing materials and structures. Structural color is based on the interaction of light with periodic microstructures. In butterfly wings for example, several optical phenomena such as multilayer interference and photonic crystal effect combine to make the colors [47, 48]. Inspired by these natural structures, many efforts have been given to replicate the high quality nanostructures and the corresponding colors [49]. Here, the focus will be on the recent reports and advances in regard to color-based structures and materials for sensing applications. Several examples of direct mimicking of the photonic structures of butterflies and insects can be found in recent reports. For instance, Kolle et al. fabricated photonic structures that mimicked the color mixing effect found on the wings of the Indonesian butterfly (*Papilio blumei*). The bright green colored areas on the wings result from a juxtaposition of blue and yellow-green lights reflected from different microscopic regions on the wing scales. This was done by utilizing a combination of layer deposition techniques, including colloidal self-assembly, sputtering, and atomic layer deposition [50]. However, reports on the utilization of these photonic structures as colorimetric sensors are limited in number.

Kim et al. demonstrated a biomimetic humidity sensor inspired by the humidity-dependent color change observed in the cuticle of the Hercules beetle (Fig. 1.4) [51]. A thin-film-type humidity sensor with nanoporous structures

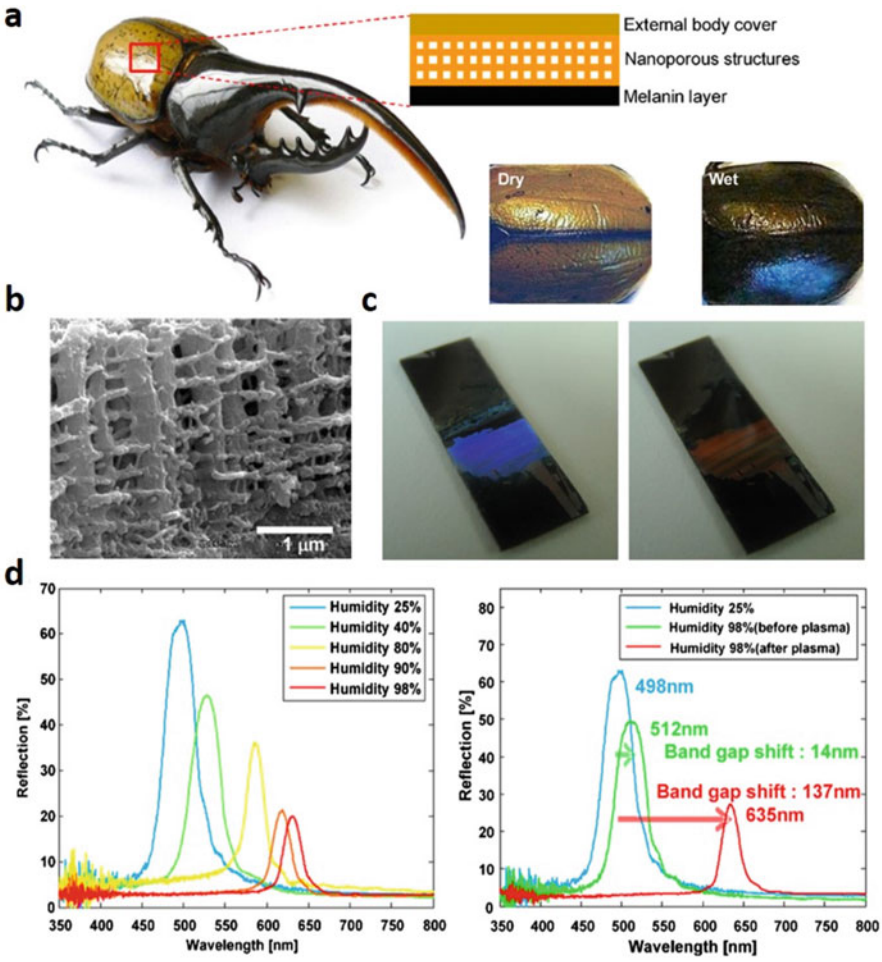


Fig. 1.4 Humidity sensor inspired by a Hercules beetle. (**a**) The exoskeleton of Hercules beetle changes from *khaki-green* (dry) atmosphere to *black* (high humidity). (**b**) SEM image of the cuticle of Hercules beetle. (**c**) Color change in the biomimetic sensors with relative humidity. (**d**) Corresponding reflectance spectra of the humidity sensor under various humidity conditions (*left*). Plasma surface treatment effect (*right*): the photonic bandgap shift increased dramatically from 14 to 137 nm in reflectance spectra of the humidity sensor after surface treatment, when the sensor is exposed to high humidity 98 %. (Reprinted with permission from [51])

(three-dimensional photonic crystals) mimicking the spongy multilayer in the beetles was designed and fabricated using the colloidal templating method and a hydrophilic surface treatment. The visible color of the fabricated humidity sensor changes from blue-green to red as the environmental humidity increases. Bai et al. recently demonstrated a facile, fast, and cost-effective technique for the fabrication of responsive colloidal photonic crystals (CPCs) with multicolor shifting properties by inkjet printing mesoporous colloidal nanoparticle ink on

both rigid and soft substrates. By adjusting the size and mesopore proportion of nanoparticles, they were able to control the original color and vapor-responsive color shift extent of the mesoporous CPC. As a consequence, multicolor mesoporous CPCs patterns with complex vapor-responsive color shifts or vapor-revealed implicit images were achieved. The complicated and reversible multicolor shifts of mesoporous CPC patterns are expected to be favorable for immediate recognition by the naked eye [52]. Inspired by the fog-collecting structure on Stenocara beetle's back, Hou et al. showed a photonic-crystal microchip with hydrophilic–hydrophobic micropatterns fabricated by inkjet printing. This device was used to realize high-sensitive ultratrace detection of fluorescence analytes and fluorophore-based assays. Ultratrace detections of fluorescence analyte (R6G) and molecular fluorophore-based assays (cocaine) down to 10–16 mol L^{-1} were achieved. This can be combined with biophotonic devices for the detection of drugs, diseases, and pollution of the ecosystem [53]. Zhang et al. utilized the wing scales from the sunset moth (*Chrysiridia rhipheus*) by incorporating it with an interpenetrating polymer network of chitosan and poly(vinyl alcohol) to generate novel pH sensors based on the "biological physical dye." The immobilized wing scales' visible reflectance was responsive to the pH conditions, owing to their inner microstructural changes induced by the change in the polymer network volume within the swelling/de-swelling process of the hydrogel. This is an interesting strategy for utilizing nature-inspired structures in visualizing pH conditions. This strategy is expected to serve as a new kind of Litmus paper without using any chemical-responsive pigments [54].

Another trend is mimicking the environment-responsive color-changing animals. A recent mimicking of a color-changing chameleon was performed by Chou et al. This chameleon-inspired stretchable electronic skin (e-skin) can easily be controlled by varying the applied pressure along with the applied pressure duration. The e-skin's color change can also be in turn utilized to distinguish the pressure applied [55]. Yu et al. demonstrated the biomimetic skin of cephalopods: animals such as octopus, squid, and cuttlefish can change its skin color by adapting to the coloration and texture of their surroundings for the purpose of concealment, communication, predation, and reproduction [56]. The color-changing skin combined multiplexed arrays of actuators and photodetectors in laminated, multilayer configurations on flexible substrates, with overlaid arrangements of pixelated, color-changing elements. This is a good example that shows how the development of flexible and stretchable electronics can be of help to the development of biomimetic structures.

1.5 Self-Assembled Structures and Materials for Sensors

It is worthy to note that there is a considerable amount of effort towards utilizing biomaterials or biomimetic synthetic materials as sensing materials or as components of sensors. Cellulose, for example, is being actively studied for sensing materials, being one of the most abundant nature-supplied materials found in the

skins and organs of animals. In this section, we will briefly discuss nature-extracted materials such as cellulose and collagen as engineering materials for sensor applications. Finally, recent studies in utilizing a microorganism itself, namely a bacteriophage, as a sensor material will be discussed.

1.5.1 Cellulose-Based Structures for Sensors

Cellulose is one of the most abundant natural polymers as it can be found easily in plants and trees. It is also considered in a wide range of fields, especially textile, paper, pharmaceuticals, the cosmetics industry, and so forth. Cellulose has a polysaccharide structure consisting of a linear chain of D-glucose units [57]. Due to the rich amount of hydroxyl groups on the molecular structure, cellulose is basically hygroscopic and polar. Therefore, the rich hydrogen bonding between cellulose chains makes them insoluble in many solvents including water. For this reason, cellulose has been widely used in the research field of sensors as a supporting matrix onto which other sensing materials can be integrated. Furthermore, the flexibility of cellulose is advantageous for the fabrication of flexible sensing devices [58].

The recent trend is the integration of cellulose with nanomaterials such as carbon nanotubes or graphene to impart interesting electronic and sensing properties while taking advantage of the flexibility and low cost of cellulose. For example, Yun et al. demonstrated a multi-walled carbon nanotube–cellulose paper as a chemical vapor sensor. The cellulose extracted from cotton pulp was covalently linked with the nanotubes. By stretching and applying mechanical strain to the paper, they were able to enhance the alignment between the nanotubes and cellulose fibers. Interestingly, the paper acquired an electronic conductivity with initial resistance. This resistance changed when the paper was exposed to gases such as methanol, ethanol, 1-butanol, and 1-propanol [59]. Another recent utilization of carbon nanotube-cellulose composites was to make a conductive CNT-cellulose paper for the detection of water. They tried to overcome the limit of conventional conductive polymer composites (CPCs) which are usually not appropriate for water detection because of their non-polar character of the polymer matrices [60]. Instead of paper, Qi et al. proposed a three-dimensional conductive porous aerogel composed of nanotubes and cellulose for gas sensing [61]. The porous structure enhances the surface-volume ratio of the sensing material and increases the overall sensitivity of the sensor.

Another trend of utilizing cellulose as a sensing material is in hybrid structures of cellulose with polymer-based materials. Mahadeva et al. demonstrated the application of cellulose as a flexible humidity and temperature sensor. A nanoscale polypyrrole polymer layer was combined with cellulose to fabricate a humidity and temperature sensor [62]. Hu et al. demonstrated a formaldehyde sensor based on polyethyleneimine nanofibers over bacterial cellulose membranes. This composite layer was used as the binding receptor layer when coated on a quartz crystal

microbalance (QCM). The sensor showed high sensitivity with good linearity and exhibited a good reversibility and repeatability towards formaldehyde in the concentration range of 1–100 ppm at room temperature. The sensing properties were mainly affected by the content of PEI components in the nanofibrous membranes, the concentration of formaldehyde, and relative humidity [63]. One last example of the cellulose-based sensor will be the composite structure of cellulose with inorganic metal oxide materials. Maniruzzaman et al. showed a composite structure of titanium dioxide and cellulose for the conductometric detection of glucose. Titanium dioxide nanoparticles were mixed with the cellulose solution. The enzyme glucose oxidase was immobilized into this hybrid nanocomposite via the physical adsorption method. They obtained a linear response to glucose in the concentration range of 1–10 mM [64].

As can be seen above, although cellulose is one of the most abundant nature-extracted materials, current research is focused on the resistive/conductive measurement of composite structures using cellulose. The main reason for this limited use of collagen as a sensing material is mainly due to the difficulty in assembling them into highly ordered hierarchical structures. Organized hierarchical structures is the principal way by which nature acquires its mechanical, optical, and sensing properties. Recently, a wide interest in nanocellulose-based structures is growing, where nanoscale cellulose is prepared by hydrolysis of cellulose fibers [65]. A further assembly and integration of these nanocellulose fibers have the potential of making nanocellulose an important sensing material in the near future.

1.5.2 Collagen-Based Structures for Sensors

Collagen is the main component of animal tissues such as skin, bones, and tendons [66]. Using its relatively well-ordered microstructure as a building block, they are assembled into diverse hierarchical structures resulting in tissues with diverse mechanical and optical properties.

An interesting application of collagen was demonstrated by Xu et al. by integrating collagen with Au nanoclusters [67]. They showed a visual sensor array based on collagen and enzyme coated Au nanoparticle clusters. They were able to discriminate eight different binding proteins using pattern analysis techniques such as linear discriminant analysis.

Collagen is generally used as sensor surrounding materials for the stable and intimate contact with electrochemical sensors. Ju et al. demonstrated implantable glucose sensors surrounded by collagen scaffolds for biocompatibility and sensor stability [68]. Another example was demonstrated by Wang et al., where the surface of a Si-based sensor was coated with extra cellular matrix collagen IV, which served as a cell-trapping structure. The sensor chip with cell-trapping patterns easily trapped cells [69].

Other biomimetic materials such as peptides are being examined as sensing materials. A recent report by Farrar et al. shows that we can combine a-helical

peptides with electrospinning techniques to obtain dipole-aligned structures with biomimetic materials [70].This was achieved by electrospinning poly(γ-benzyl α,L-glutamate), a liquid crystalline, and α-helical poly(α-amino acid) with macroscopic dipoles prealigned in the direction of helical axis. The dipoles became aligned and fixed during the process of electrospinning and they reported a piezoelectric coefficient of 25 pC N^{-1}, which is quite a high value compared to other biological materials that have values under 1 pC N^{-1}. These biomimetic synthetic polymers are expected to become major candidates of engineering materials for sensors and actuators.

1.5.3 Virus-Based Structures for Sensors

An interesting and novel biological material for sensors is the M13 bacteriophage. M13, a filamentous virus that can be genetically engineered, has been reported to be a strong candidate for building diverse sensors to detect external mechanical and biochemical stimuli. The bacteriophage is benign to humans and only infects bacteria. Due to the filamentous structure similar to collagen, cellulose, and chitin, they can assemble into diverse hierarchically organized structures. In addition, the chemical structure of the coat proteins can be easily modified to express specific binding receptors to target molecules by genetically engineering the DNA and expressing the coat proteins. Moreover, bacteriophages can be replicated in large quantities by simply incubating the host bacteria. Due to these properties, recent studies show that they can be used as engineering materials for sensor applications. Chung et al. showed that filamentous bacteriophage can serve as basic building blocks that can self-assemble into diverse hierarchical structures [71]. They utilized a simple pulling method where a solid substrate is dipped into the bacteriophage solution and then slowly pulled up at a given velocity. When the substrate is pulled, the individual phages self-organize on the substrate into supramolecular hierarchical structures. By controlling several factors such as pulling speed, substrate surface chemistry, bacteriophage concentration, and ionic concentration, Chung et al. showed that the assembled structures can be controlled, resulting in periodic structures with characteristically optical and mechanical properties.

By utilizing this assembly method, Lee et al. showed that bacteriophages can be assembled into hierarchical structures with piezoelectric properties, which enables the conversion of external mechanical stimuli into electrical signals (Fig. 1.5) [72]. Using piezoresponse force microscopy, they showed that individual bacteriophages have structure-dependent piezoelectric properties at the molecular level. The assembled bacteriophage film displayed piezoelectric coefficients around 7.8 pm V^{-1}. The dipole strength and piezoelectric response of the films can be easily tuned by changing the peptide structure of the coat proteins through genetic engineering. The piezoelectric device was responsive to external pressure, and a typical device of 1 cm^2 active area device produced up to 6 nA of current and 400 mV of potential, which was enough to turn on a liquid-crystal display.

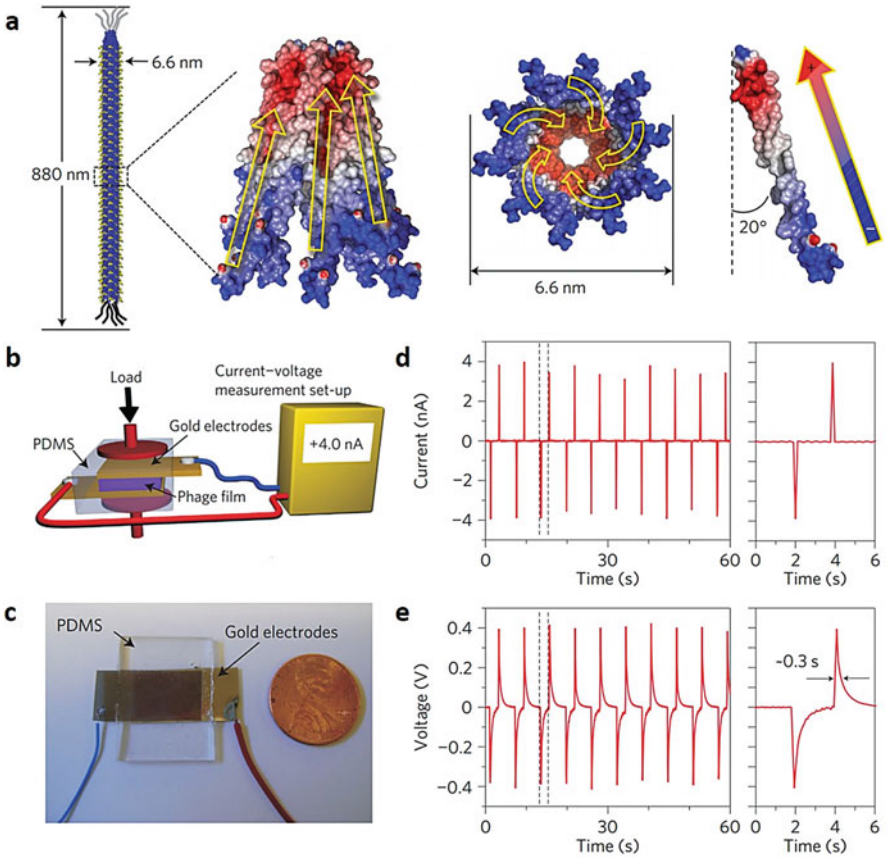


Fig. 1.5 Virus-based mechanical sensor. (**a**) Schematic of piezoelectric M13 phage structure. (**b**) Schematic of piezoelectric electrical energy generation measurement setup. A mechanical load was applied to the device while monitoring the voltage and current. (**c**) Photograph of a phage-based generator. (**d**) Short-circuit current signal from the phage-based generator. (**e**) Open-circuit voltage signal from the phage-based generator. (Reprinted with permission from [72])

Using the self-assembly of bacteriophages into diverse hierarchical structures, Oh et al. demonstrated a rapidly responsive (on the order of seconds) and reversible bacteriophage color film for the detection of gas-phase chemicals (Fig. 1.6) [73]. This mimics the color-changing properties of Turkey skin, where the color of the tissue is dependent on the physicochemical condition of the collagen bundles. The structural change during target molecule binding was observed with grazing incident small-angle X-ray scattering. It showed that the diameters of the bacteriophage bundles increase while the interbundle distance decreases during the binding process. This resulted in red-shifting of the dominant coherently scattered wavelengths of light. When the bacteriophage film was exposed to certain chemical vapors, the color matrix changed in color, and the color shift was analyzed with pattern analysis algorithms such as the principal component analysis (PCA).

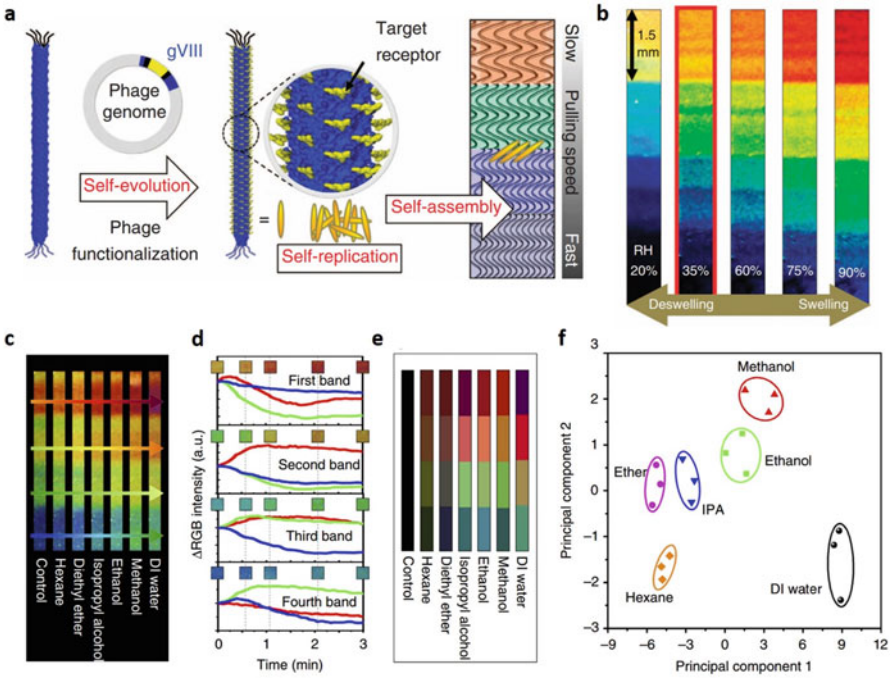


Fig. 1.6 Bacteriophage-based colorimetric sensor. (**a**) Genetically engineering, replication and self-assembly of bacteriophages. (**b**) Bacteriophage-based colorimetric sensor. Changes in relative humidity (RH) result in changes in colors. (**c**) Sensor after exposure to hexane, diethyl ether, isopropyl alcohol, ethanol, methanol, and DI water, respectively. (**d**) Real-time RGB color change of bacteriophage color sensor after exposure to DI water. (**e**) VOC color fingerprints used to selectively distinguish various chemicals. (**f**) Principal component analysis plot of the color changes resulting from different VOCs. (Reprinted with permission from [73])

The PCA showed that the bacteriophage-based colorimetric sensors can distinguish between various volatile organic compounds (VOCs).

Although at an early stage, biomimetic structures based on bacteriophages show a promising future in building well-organized hierarchical structures for the detection of mechanical and chemical stimuli via its piezoelectric and color-changing property. In conjunction with other biological molecules such as collagen, cellulose and other peptides, bacteriophages will be useful as engineering materials for sensor applications.

1.6 Conclusions

After millions of years of evolution, nature has a plethora of models for sensors that will serve as a source of ideas and inspiration for the community of scientists and engineers. Biomimetic materials based on biomimetic materials, soft polymers, inorganic silicon structures, biomineralized crystals, biological materials including bacteriophages and collagen, and so on, will be the center of interest and research for the following years. In particular, sensors based on biomimetic structures and materials are still at an early stage compared to other electrochemical or inorganic material-based gas sensors. Most of the research is focused on the detection of mechanical stimuli such as pressure and vibration. Meanwhile, the sensors related to the electroreception and magnetoreception of aquatic animals and birds remain almost an untrodden land. Even so, biomimetic materials and structures will remain one of the most promising scientific and engineering pathways towards successful sensor development. Nevertheless, there is no doubt that nature will continue to be the most abundant source of inspiration and ideas left for us to discover.

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

References

1. Meyers MA, Chen P-Y, Lin AY-M, Seki Y (2008) Biological materials: structure and mechanical properties. Prog Mater Sci 53:1–206
2. Xia F, Jiang L (2008) Bio-inspired, smart, multiscale interfacial materials. Adv Mater 20:2842–2858
3. Lenau T, Stroble J, Stone R, Watkins S (2009) An overview of biomimetic sensor technology. Sens Rev 29:112–119
4. Sanchez C, Arribart H, Guille MMG (2005) Biomimetism and bioinspiration as tools for the design of innovative materials and systems. Nat Mater 4:277–288
5. Johnson EAC, Bonser RHC, Jeronimidis G (2009) Recent advances in biomimetic sensing technologies. Philos Trans R Soc Lond A 367:1559–1569
6. Mulvaney SP, Sheehan PE (2014) Nature inspires sensors to do more with less. ACS Nano 8:9729–9732
7. Stroble JK, Stone RB, Watkins SE (2009) An overview of biomimetic sensor technology. Sens Rev 29:112–119
8. Casas J, Steinmann T, Krijnen G (2010) Why do insects have such a high density of flow-sensing hairs? Insights from the hydromechanics of biomimetic MEMS sensors. J R Soc Interface 7:1487–1495
9. Junliang T, Xiong Y (2012) Hair flow sensors: from bio-inspiration to bio-mimicking—a review. Smart Mater Struct 21:113001
10. Liu C (2007) Micromachined biomimetic artificial haircell sensors. Bioinspir Biomim 2:S162
11. Krijnen GJM, Dijkstra M, van Baar JJ, Shankar SS, Kuipers WJ, de Boer RJH, Altpeter D, Lammerink TSJ, Wiegerink R (2006) MEMS based hairflow-sensors as model systems for acoustic perception studies. Nanotechnology 17:S84

12. Hein M, Maqableh MM, Delahunt MJ, Tondra M, Flatau AB, Shield CK, Stadler BJ (2013) Fabrication of bioinspired inorganic nanocilia sensors. IEEE Trans Magn 49:191–196
13. Schroeder P, Schotter J, Shoshi A, Eggeling M, Bethge O, Hütten A, Brückl H (2011) Artificial cilia of magnetically tagged polymer nanowires for biomimetic mechanosensing. Bioinspir Biomim 6:046007
14. Nguyen N, Jones DL, Yang Y, Liu C (2011) Flow vision for autonomous underwater vehicles via an artificial lateral line. EURASIP J Adv Signal Process 2011:9
15. Droogendijk H, de Boer MJ, Sanders RGP, Krijnen GJM (2015) Advantages of electrostatic spring hardening in biomimetic hair flow sensors. J Microelectromech Syst 24:1415–1425
16. Tee BCK, Chortos A, Berndt A, Nguyen AK, Tom A, McGuire A, Lin ZC, Tien K, Bae W-G, Wang H, Mei P, Chou H-H, Cui B, Deisseroth K, Ng TN, Bao Z (2015) A skin-inspired organic digital mechanoreceptor. Science 350:313–316
17. Sheng L, Liang Y, Jiang L, Wang Q, Wei T, Qu L, Fan Z (2015) Bubble-decorated honeycomb-like graphene film as ultrahigh sensitivity pressure sensors. Adv Funct Mater 25:6545–6551
18. Su B, Gong S, Ma Z, Yap LW, Cheng W (2015) Mimosa-inspired design of a flexible pressure sensor with touch sensitivity. Small 11:1886–1891
19. Pang C, Lee G-Y, Kim T-I, Kim SM, Kim HN, Ahn S-H, Suh K-Y (2012) A flexible and highly sensitive strain-gauge sensor using reversible interlocking of nanofibres. Nat Mater 11:795–801
20. Ha M, Lim S, Park J, Um DS, Lee Y, Ko H (2015) Bioinspired interlocked and hierarchical design of ZnO nanowire arrays for static and dynamic pressure-sensitive electronic skins. Adv Funct Mater 25:2841–2849
21. Fratzl P, Barth FG (2009) Biomaterial systems for mechanosensing and actuation. Nature 462:442–448
22. Kang D, Pikhitsa PV, Choi YW, Lee C, Shin SS, Piao L, Park B, Suh K-Y, Kim T-I, Choi M (2014) Ultrasensitive mechanical crack-based sensor inspired by the spider sensory system. Nature 516:222–226
23. Valle MD (2011) Bioinspired sensor systems. Sensors 11:10180–10186
24. Hammock ML, Chortos A, Tee BCK, Tok JBH, Bao Z (2013) 25th anniversary article: the evolution of electronic skin (e-skin): a brief history, design considerations, and recent progress. Adv Mater 25:5997–6038
25. Park J, Kim M, Lee Y, Lee HS, Ko H (2015) Fingertip skin–inspired microstructured ferroelectric skins discriminate static/dynamic pressure and temperature stimuli. Sci Adv 1: e1500661
26. Tee BC, Wang C, Allen R, Bao Z (2012) An electrically and mechanically self-healing composite with pressure-and flexion-sensitive properties for electronic skin applications. Nat Nanotechnol 7:825–832
27. Hou C, Huang T, Wang H, Yu H, Zhang Q, Li Y (2013) A strong and stretchable self-healing film with self-activated pressure sensitivity for potential artificial skin applications. Sci Rep 3:3188
28. Kim J, Lee M, Shim HJ, Ghaffari R, Cho HR, Son D, Jung YH, Soh M, Choi C, Jung S, Chu K, Jeon D, Lee S-T, Kim JH, Choi SH, Hyeon T, Kim D-H (2014) Stretchable silicon nanoribbon electronics for skin prosthesis. Nat Commun 5:5747
29. Scheibert J, Leurent S, Prevost A, Debrégeas G (2009) The role of fingerprints in the coding of tactile information probed with a biomimetic sensor. Science 323:1503–1506
30. Zhang Y (2010) Sensitivity enhancement of a micro-scale biomimetic tactile sensor with epidermal ridges. J Micromech Microeng 20:085012
31. Lee LP, Szema R (2005) Inspirations from biological optics for advanced photonic systems. Science 310:1148–1150
32. Song YM, Xie Y, Malyarchuk V, Xiao J, Jung I, Choi K-J, Liu Z, Park H, Lu C, Kim R-H (2013) Digital cameras with designs inspired by the arthropod eye. Nature 497:95–99

33. Floreano D, Pericet-Camara R, Viollet S, Ruffier F, Brückner A, Leitel R, Buss W, Menouni M, Expert F, Juston R, Dobrzynski MK, L'Eplattenier G, Recktenwald F, Mallot HA, Franceschini N (2013) Miniature curved artificial compound eyes. Proc Natl Acad Sci 110:9267–9272
34. Wilson AD, Baietto M (2009) Applications and advances in electronic-nose technologies. Sensors 9:5099–5148
35. Spitzer D, Cottineau T, Piazzon N, Josset S, Schnell F, Pronkin SN, Savinova ER, Keller V (2012) Bio-inspired nanostructured sensor for the detection of ultralow concentrations of explosives. Angew Chem Int Ed 51:5334–5338
36. Zhang W, Tian J, Wang YA, Fang X, Huang Y, Chen W, Liu Q, Zhang D (2014) Single porous SnO2 microtubes templated from *Papilio maacki* bristles: new structure towards superior gas sensing. J Mater Chem A 2:4543–4550
37. Wang J, Zhao X, Li J, Kuang X, Fan Y, Wei G, Su Z (2014) Electrostatic assembly of peptide nanofiber-biomimetic silver nanowires onto graphene for electrochemical sensors. ACS Macro Lett 3:529–533
38. Zhang D, Lu Y, Zhang Q, Yao Y, Li S, Li H, Zhuang S, Jiang J, Liu GL, Liu Q (2015) Nanoplasmonic monitoring of odorants binding to olfactory proteins from honeybee as biosensor for chemical detection. Sens Actuators B 221:341–349
39. Lee SH, Kwon OS, Song HS, Park SJ, Sung JH, Jang J, Park TH (2012) Mimicking the human smell sensing mechanism with an artificial nose platform. Biomaterials 33:1722–1729
40. Goldsmith BR, Mitala JJ Jr, Josue J, Castro A, Lerner MB, Bayburt TH, Khamis SM, Jones RA, Brand JG, Sligar SG (2011) Biomimetic chemical sensors using nanoelectronic readout of olfactory receptor proteins. ACS Nano 5:5408–5416
41. Kim TH, Lee BY, Jaworski J, Yokoyama K, Chung W-J, Wang E, Hong S, Majumdar A, Lee S-W (2011) Selective and sensitive TNT sensors using biomimetic polydiacetylene-coated CNT-FETs. ACS Nano 5:2824–2830
42. Kwon OS, Song HS, Park SJ, Lee SH, An JH, Park JW, Yang H, Yoon H, Bae J, Park TH, Jang J (2015) An ultrasensitive, selective, multiplexed superbioelectronic nose that mimics the human sense of smell. Nano Lett 15:6559–6567
43. Guo CX, Ng SR, Khoo SY, Zheng X, Chen P, Li CM (2012) RGD-peptide functionalized graphene biomimetic live-cell sensor for real-time detection of nitric oxide molecules. ACS Nano 6:6944–6951
44. Song HS, Kwon OS, Lee SH, Park SJ, Kim U-K, Jang J, Park TH (2013) Human taste receptor-functionalized field effect transistor as a human-like nanobioelectronic tongue. Nano Lett 13:172–178
45. Lee M, Jung JW, Kim D, Ahn Y-J, Hong S, Kwon HW (2015) Discrimination of umami tastants using floating electrode-based bioelectronic tongue mimicking insect taste systems. ACS Nano 9:11728–11736
46. Ge J, Yin Y (2011) Responsive photonic crystals. Angew Chem Int Ed 50:1492–1522
47. Vukusic P, Sambles JR (2003) Photonic structures in biology. Nature 424:852–855
48. Doucet SM, Meadows MG (2009) Iridescence: a functional perspective. J R Soc Interface 6: S115–S132
49. Zhao Y, Xie Z, Gu H, Zhu C, Gu Z (2012) Bio-inspired variable structural color materials. Chem Soc Rev 41:3297–3317
50. Kolle M, Salgard-Cunha PM, Scherer MR, Huang F, Vukusic P, Mahajan S, Baumberg JJ, Steiner U (2010) Mimicking the colourful wing scale structure of the *Papilio blumei* butterfly. Nat Nanotechnol 5:511–515
51. Kim JH, Moon JH, Lee S-Y, Park J (2010) Biologically inspired humidity sensor based on three-dimensional photonic crystals. Appl Phys Lett 97:103701
52. Bai L, Xie Z, Wang W, Yuan C, Zhao Y, Mu Z, Zhong Q, Gu Z (2014) Bio-inspired vapor-responsive colloidal photonic crystal patterns by inkjet printing. ACS Nano 8:11094–11100
53. Hou J, Zhang H, Yang Q, Li M, Song Y, Jiang L (2014) Bio-inspired photonic-crystal microchip for fluorescent ultratrace detection. Angew Chem 126:5901–5905

54. Zang X, Tan Y, Lv Z, Gu J, Zhang D (2012) Moth wing scales as optical pH sensors. Sens Actuators B 166–167:824–828
55. Chou H-H, Nguyen A, Chortos A, To JWF, Lu C, Mei J, Kurosawa T, Bae WG, Tok JBH, Bao Z (2015) A chameleon-inspired stretchable electronic skin with interactive colour changing controlled by tactile sensing. Nat Commun 6:8011
56. Yu C, Li Y, Zhang X, Huang X, Malyarchuk V, Wang S, Shi Y, Gao L, Su Y, Zhang Y, Xu H, Hanlon RT, Huang Y, Rogers JA (2014) Adaptive optoelectronic camouflage systems with designs inspired by cephalopod skins. Proc Natl Acad Sci 111:12998–13003
57. Klemm D, Heublein B, Fink H-P, Bohn A (2005) Cellulose: fascinating biopolymer and sustainable raw material. Angew Chem Int Ed 44:3358–3393
58. Kim J-H, Mun S, Ko H-U, Yun G-Y, Kim J (2014) Disposable chemical sensors and biosensors made on cellulose paper. Nanotechnology 25:092001
59. Yun S, Kim J (2010) Multi-walled carbon nanotubes–cellulose paper for a chemical vapor sensor. Sens Actuators B 150:308–313
60. Qi H, Mäder E, Liu J (2013) Unique water sensors based on carbon nanotube–cellulose composites. Sens Actuators B 185:225–230
61. Qi H, Liu J, Pionteck J, Pötschke P, Mäder E (2015) Carbon nanotube–cellulose composite aerogels for vapour sensing. Sens Actuators B 213:20–26
62. Mahadeva SK, Yun S, Kim J (2011) Flexible humidity and temperature sensor based on cellulose–polypyrrole nanocomposite. Sens Actuators, A 165:194–199
63. Hu W, Chen S, Liu L, Ding B, Wang H (2011) Formaldehyde sensors based on nanofibrous polyethyleneimine/bacterial cellulose membranes coated quartz crystal microbalance. Sens Actuators B 157:554–559
64. Maniruzzaman M, Jang S-D, Kim J (2012) Titanium dioxide–cellulose hybrid nanocomposite and its glucose biosensor application. Mater Sci Eng B 177:844–848
65. Dufresne A (2013) Nanocellulose: a new ageless bionanomaterial. Mater Today 16:220–227
66. Matthews JA, Wnek GE, Simpson DG, Bowlin GL (2002) Electrospinning of collagen nanofibers. Biomacromolecules 3:232–238
67. Xu S, Lu X, Yao C, Huang F, Jiang H, Hua W, Na N, Liu H, Ouyang J (2014) A visual sensor array for pattern recognition analysis of proteins using novel blue-emitting fluorescent gold nanoclusters. Anal Chem 86:11634–11639
68. Ju YM, Yu B, West L, Moussy Y, Moussy F (2010) A novel porous collagen scaffold around an implantable biosensor for improving biocompatibility. II. Long-term in vitro/in vivo sensitivity characteristics of sensors with NDGA-or GA-crosslinked collagen scaffolds. J Biomed Mater Res A 92:650–658
69. Wang Z-H, Takada N, Uno H, Ishizuka T, Yawo H, Urisu T (2012) Positioning of the sensor cell on the sensing area using cell trapping pattern in incubation type planar patch clamp biosensor. Colloids Surf B Biointerfaces 96:44–49
70. Farrar D, Ren K, Cheng D, Kim S, Moon W, Wilson WL, West JE, Yu SM (2011) Permanent polarity and piezoelectricity of electrospun α-helical poly(α-amino acid) fibers. Adv Mater 23:3954–3958
71. Chung W-J, Oh J-W, Kwak K, Lee BY, Meyer J, Wang E, Hexemer A, Lee S-W (2011) Biomimetic self-templating supramolecular structures. Nature 478:364–368
72. Lee BY, Zhang J, Zueger C, Chung W-J, Yoo SY, Wang E, Meyer J, Ramesh R, Lee S-W (2012) Virus-based piezoelectric energy generation. Nat Nanotechnol 7:351–356
73. Oh J-W, Chung W-J, Heo K, Jin H-E, Lee BY, Wang E, Zueger C, Wong W, Meyer J, Kim C (2014) Biomimetic virus-based colourimetric sensors. Nat Commun 5

Chapter 2
A Multi-Modal CMOS Sensor Platform Towards Personalized DNA Sequencing

Yu Jiang, Xu Liu, Xiwei Huang, Yang Shang, Mei Yan, and Hao Yu

Abstract Precision medicine requires scalable bio-instrument for a personalized DNA sequencing, which can be label-free, cost-efficient, and high-throughput. This chapter mainly presents three kinds of CMOS-based label-free sensors, including: (1) a high-sensitivity ion-sensitive field-effect transistor (ISFET) sensor with pH-to-time-to-voltage conversion (pH-TVC); (2) a dual-mode sensor with image and chemical modes for high accuracy; and (3) a THz metamaterial sensor with electrical resonance detection. The developed CMOS multi-modal sensor platform can show a scaled solution for future personalized DNA sequencing.

Keywords Personalized DNA sequencing • Label-free sensors • ISFET • Metamaterial sensor • Multi-modal • Genotyping • Dual-mode sensor • Contact imaging • THz-based sensor • High sensitivity

2.1 Introduction

DNA detection, categorized as sequencing and genotyping, plays a significant role for modern human health. Sequencing, detection of the order of nucleotides in DNA strands, enables studies of metagenomics, genetic disorders, diseases, and genomic medicine. Genotyping, targeted sequencing or mutation of specific DNA, is deployed for single nucleotide polymorphism (SNP) detections, most of which are associated with diseases and deficiencies [1]. Sanger sequencing was successfully employed in DNA detection since 1970s, which is expensive and time-consuming for large-scale sequencing [2]. Next generation sequencing (NGS) technologies are later developed for high-throughput sequencing with low cost, including pyrosequencing (454), sequencing by oligo ligation detection (SOLiD), and Illumina sequencing [3]. However, these methods require fluorescent labels and bulky optical instruments and hence are not feasible for personalized diagnosis.

A large-array CMOS-compatible sensor foresees a strong potential in the future personalized DNA sequencing. Same as computer and communication devices, it

Y. Jiang • X. Liu • X. Huang • Y. Shang • M. Yan • H. Yu (✉)
School of EEE, Nanyang Technological University, Singapore, Singapore
e-mail: haoyu@ntu.edu.sg

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_2

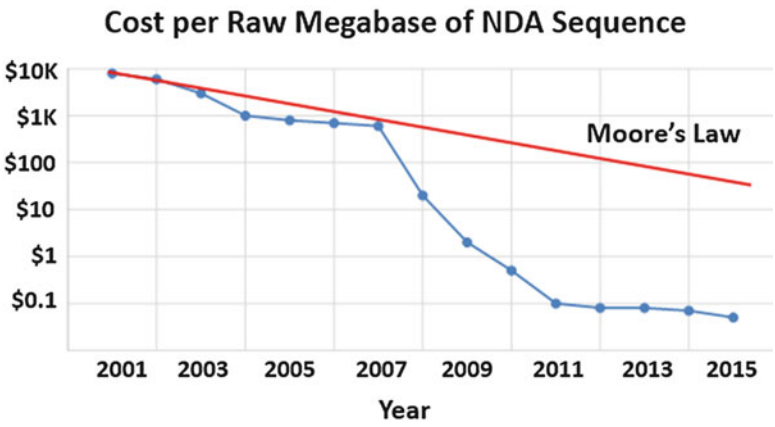


Fig. 2.1 The relationship between DNA sequence cost and Moore's law: since the year 2008, the development of new sequencing technologies and the continuous semiconductor process improving lead to dramatically drop on the cost. Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts

follows the Moore's law that is scalable for millions of DNA strands to be detected simultaneously on a single CMOS chip. As shown in Fig. 2.1, since 2008, the development of new sequencing technologies under the continuous semiconductor process scaling leads to dramatically scale-down on the cost in CMOS process. State-of-the-art CMOS-compatible methods include ion-sensitive field-effect transistor (ISFET) based [4–7] chemical sensing and nanopore based [8] electrical sensing.

This chapter introduces the latest development of CMOS-based multi-modal sensor platform for personalized DNA sequencing, which includes: (1) a high-sensitivity ISFET sensor by pH-to-time-to-voltage conversion (pH-TVC); (2) a dual-mode ion-image sensor with high accuracy; and (3) a proposed THz metamaterial sensor.

2.2 High-Sensitivity CMOS pH-TVC ISFET Sensor

An ISFET sensor builds connection between aqueous solution and solid-state circuits by converting biochemical reaction to electrical signal. Traditional ISFETs are built with expensive specialized processes, since gate oxide is utilized as ion sensing membrane [9]. Improved structure proposed by Bausells et al., in which Si_3N_4 passivation layer is employed as the sensing membrane, makes ISFET sensor compatible and scalable with commercial CMOS technology [10]. Consequently, CMOS-based ISFET sensors have shown great promise in NGS of DNA, due to its low-cost, high-speed, and large-scale advantages [6].

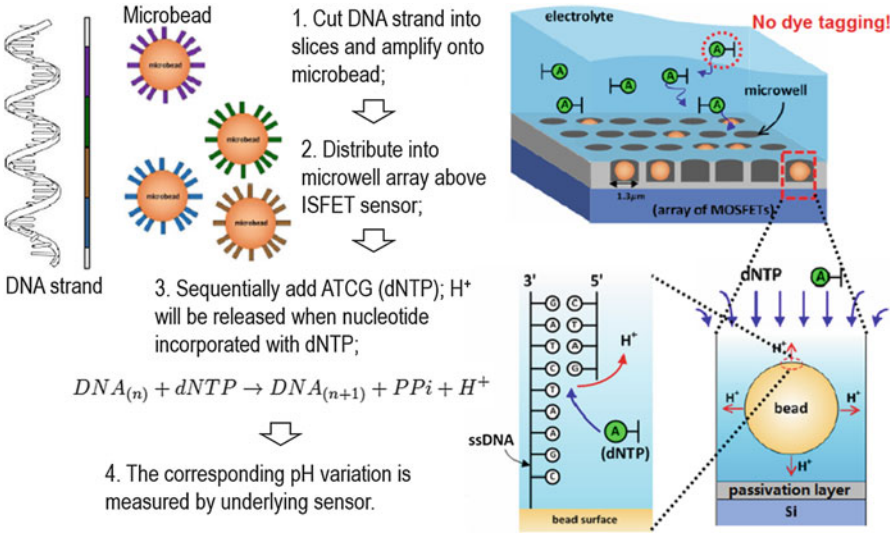


Fig. 2.2 pH-based DNA sequencing process: including DNA chain fragmentation, amplification, and distribution; then H^+ released during sequencing will be detected by ISFET sensor

The CMOS-based ISFET sensor in DNA sequencing is shown in Fig. 2.2. Long DNA chain is first fragmented into short templates, then clonally amplified on microbeads by polymerase chain reaction (PCR). The DNA-templated microbeads are scattered into microwells mounted on chip surface through centrifuge spinning. During detection, four nucleotides (dNTPs) are cyclically delivered, and a hydrogen ion will be released when one nucleotide is incorporated to its complementary template base. Decreased pH value will be observed, which is proportional to the number of incorporations. The ISFET sensor is thereby adopted to detect this pH change to assemble the DNA sequence.

2.2.1 CMOS ISFET Model

For a CMOS ISFET device, an Ag/AgCl electrode in solution works as a remote gate, and a floating gate (FG) structure is formed between the passivation layer and gate oxide, as shown in Fig. 2.3a. The chemical–electrical conversion happening at the solution and ISFET interface can be explained by the site-binding theory, in which chemical groups on the surface of the passivation layer, SiOH and $SiNH_2$ sites, will donate H^+ to or accept from the solution. These sites can be positively charged, negatively charged, or neutral. The equilibrium reactions at these sites can be described as follows [11]:

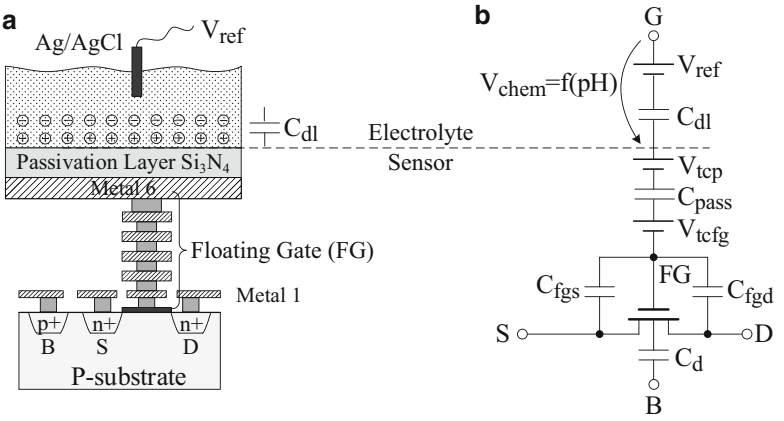


Fig. 2.3 (**a**) Cross section of an ISFET implemented in 1P6M CMOS process. (**b**) A capacitance-dependent CMOS ISFET model. C_{dl} is the double layer capacitance on solid–liquid surface, and V_{chem} represents pH-related potential

$$\mathrm{SiOH} \rightleftharpoons \mathrm{SiO^-} + \mathrm{H_s^+} \tag{2.1}$$

$$\mathrm{SiOH_2^+} \rightleftharpoons \mathrm{SiOH} + \mathrm{H_s^+} \tag{2.2}$$

$$\mathrm{SiNH_3^+} \rightleftharpoons \mathrm{SiNH_2} + \mathrm{H_s^+} \tag{2.3}$$

where H_s^+ represents a hydrogen ion at the surface. The ions SiO^-, $SiOH_2^+$, and $SiNH_3^+$ combined represent the charge density at the passivation surface σ_0, which is equally balanced by an opposite charge density in the solution σ_{dl}. As a result, a double layer capacitance C_{dl} forms at the interface. The potential difference between the bulk solution and passivation surface can be defined as ψ_0 across C_{dl} [12]:

$$\sigma_0 = C_{dl}\psi_0 = -\sigma_{dl} \tag{2.4}$$

Considering the Boltzmann distribution of hydrogen across the double capacitance layer and non-Nernstian response, the sensitivity of ψ_0 to pH change in the bulk solution can be described as [12]:

$$\frac{\Delta\psi_0}{\Delta pH} = -2.303\alpha U_T \tag{2.5}$$

where α is a dimensionless sensitivity factor which varies between 0 and 1, and U_T is the thermal voltage. The value of α depends on the passivation material and the H^+ concentration in the solution. An ideal maximum 59.2 mV/pH Nernstian sensitivity may be reached at $\alpha = 1$.

A widely used, capacitance-dependent ISFET model working in weak-inversion region is illustrated in Fig. 2.3b, in which trapped charge accumulated during fabrication is also depicted. According to [13], the ISFET V_{th} can be expressed as

$$V_{th(ISFET)} = V_{chem} + V_{tc} + V_{th(MOSFET)}/A \tag{2.6}$$

where V_{chem} is the grouped chemical-related potential; V_{tc} is the combined trapped charge potential in the passivation V_{tcp} and the floating gate V_{tcfg}; and A is the capacitive division factor caused by passivation capacitance.

V_{chem}, V_{tc}, and A are further summarized as follows:

$$V_{chem} = \gamma + 2.303\alpha U_T \cdot pH \tag{2.7}$$

$$V_{tc} = V_{tcp} + V_{tcfg}/A \tag{2.8}$$

$$A = \frac{C_{pass}}{\left(C_{pass} + C_{ox}//C_d + C_{fgd} + C_{fgs}\right)} \tag{2.9}$$

where γ represents all non-pH related potential; C_{pass}, C_{ox}, and C_d are the passivation, oxide, and depletion capacitances, respectively; C_{fgd}, and C_{fgs} are the parasitic capacitances associated with the floating gate.

For a weak-inversion ISFET, where $V_{GS} < V_{th(ISFET)}$, the drain current I_D can be expressed as [6]:

$$I_D = I_0 \cdot K \cdot \exp \frac{-A \cdot 2.303\alpha U_T \cdot pH}{nU_T} \tag{2.10}$$

where K represents all the non-pH related terms, and $n = 1 + C_d/C_{ox}$ is the non-ideal slope factor.

According to (2.5), the ISFET drain current is attenuated by A. This division effect can degrade the ISFET sensor sensitivity, especially at advanced technology node.

2.2.2 pH-TVC Readout Scheme

Large-arrayed CMOS ISFET sensors have been commercialized by Ion Torrent™, whose products include 314, 316, 318, Proton I and Proton III with array size from 1.5 million to 660 million pixels. Even though the ISFET array fabricated in CMOS process with low cost has been realized by industry, the poor pH sensitivity is still a major problem remaining to be solved.

In order to reduce fabrication cost, passivation layer in standard CMOS process is utilized as ion-sensitive layer. However, this layer will introduce a small capacitance C_{pass} to the device model, which will capacitively decrease the coupling strength from solution to the ISFET reflected in the capacitive division factor

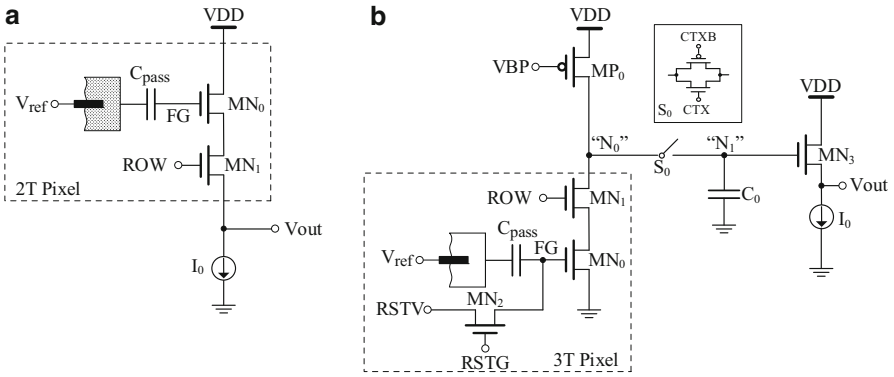


Fig. 2.4 (**a**) A conventional source follower readout method; ISFET works in saturation region. (**b**) pH-TVC readout structure; ISFET works in weak-inversion region

A described in last part. Besides, in conventional testing method shown in Fig. 2.4a, the ISFET works in saturation region as a source follower, and pH change in solution is linearly correlated to pixel output and there is no gain from the source follower. As a result, pH sensitivity is greatly affected by the passivation capacitance, especially in larger sensor array with smaller sensing area. What is more, during the DNA incorporation process, minimum to 0.02 pH shift per base incorporation is reported. Passivation material, Si_3N_4, is reported to have a pH sensitivity of 46~56 mV/pH, which means that as small as 1 mV change on the passivation surface should be detected by the pH sensor. Additionally, the floating gate ISFET structure with the passivation layer will accumulate trapped charge during fabrication, which may lead to large and unpredictable threshold voltage deviation.

There are various technologies proposed to deal with these non-ideal characteristics of CMOS-based ISFET sensor. A second electrical input is capacitively coupled to the ISFET floating gate to realize a high input inferred sensitivity [14]. Correlated double sampling (CDS) method is used to reduce overall long-term drift by cancelling the common drift between two consecutive samples [15]. Ultraviolet (UV) radiation and bulk-substrate bias are employed to remove the trapped charge accumulated by creating an aperture in the top metal right above the ISFET for UV exposure [16]. A robust ISFET front-end readout circuit is proposed to compensate issues of capacitive division, drift, trapped charge by applying a feedback loop to the floating gate [17].

In addition, the high-gain of ISFETs mentioned above is realized mainly by employing a quite large chemical sensing area and a relatively small coupled capacitance. This structure is however not feasible for large denser array, where small sensing area is required. Under this condition, a smaller coupled capacitance will suffer from a larger process variation.

In this chapter, we have developed a large-scale ISFET sensor with a novel pH-TVC readout scheme to compensate capacitive attenuation, and a reset device to alleviate drift and trapped charge influence. The proposed pH-TVC scheme is

shown in Fig. 2.4b. Each pixel contains three transistors: MN_0 is an ISFET device, MN_1 is a row-selected device, and MN_2 is a reset device. At first stage, node N_1 is pre-charged to power supply by turning on the device MP_0 and the transfer switch S_0. During the pH sensing phase, N_1 is discharged to ground, and the discharging time depends on the drain current of weak-inversion ISFET MN_0, which is exponential to pH change. Following this pH-to-time conversion is the time-to-voltage conversion at N_1 by turning off S_0 at a given time. Therefore, the pH change in solution is first converted to discharging time variation then to voltage difference. MN_3 is the source follower (SF) device to transfer pixel output voltage. I_0 is the current bias of MN_3.

For the solution with an initial H^+ concentration of pH_1, the local DNA slice at one microwell above one ISFET pixel will cause a shift in pH value, which results in a new concentration of pH_2. As a result, the pH-related voltage change at node "N_1," ΔV_{pH}, can be defined as

$$\Delta V_{pH1} = V_{DD} - V_{pH1} = I_{pH1} \cdot \frac{\Delta t}{C_0} \tag{2.11}$$

$$\Delta V_{pH2} = V_{DD} - V_{pH2} = I_{pH2} \cdot \frac{\Delta t}{C_0} \tag{2.12}$$

$$\Delta V_{pH} = V_{pH1} - V_{pH2} = \Delta V_{pH1} \cdot \left(\frac{I_{pH2}}{I_{pH1}} - 1 \right) \tag{2.13}$$

where I_{pH1} and I_{pH2} are the ISFET currents related to pH_1 and pH_2, respectively. V_{pH1} and V_{pH2} are the relative output voltages at node "N_1" after a given sensing time Δt. ΔV_{pH} can be improved by increasing the ratio between I_{pH2} and I_{pH1}. Therefore, the small ISFET input signal is amplified at the pixel level instead of using the external amplifier [18], which can greatly increase readout signal to noise ratio.

Based on (2.10), the pH sensitivity to ΔV_{pH} of ISFETs working in subthreshold region is given by

$$\Delta V_{pH} = \Delta V_{pH1} \cdot \left[\exp \frac{A \cdot 2.303 \alpha U_T \cdot (pH_1 - pH_2)}{n U_T} - 1 \right] \tag{2.14}$$

Compared to the linear relationship in conventional source follower method, an exponentially amplified output voltage ΔV_{pH} for a small pH change can be observed after applying pH-TVC readout scheme, which has no more attenuation due to the passivation layer parasitic capacitor. As a result, a denser ISFET array with a smaller pixel pitch can be realized in advanced CMOS process.

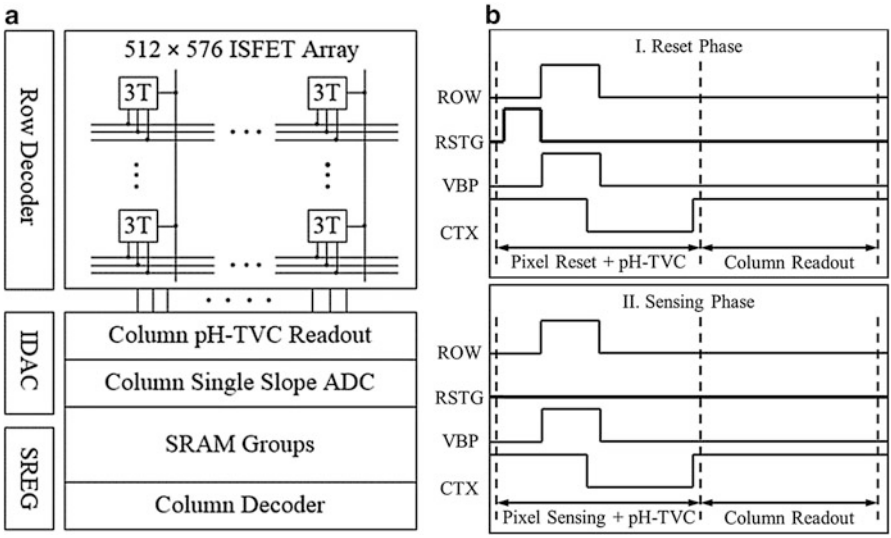


Fig. 2.5 (**a**) Top architecture of the CMOS ISFET sensor. (**b**) Pixel readout timing diagram with pH-TVC

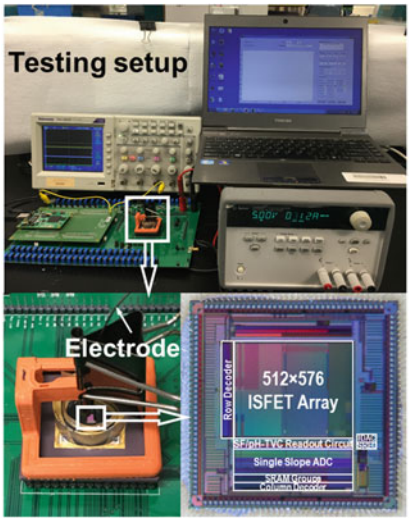
2.2.3 Top Architecture and Operation

Top architecture of the CMOS ISFET sensor is shown in Fig. 2.5a, including a 512×576 pixel array, pH-TVC readout circuits, 10-bit column ADCs, SRAM groups, row decoder, column decoder, etc. For large-scale pixel array, pixel pitch is relatively small, so the column readout circuit should be as simple as possible to be aligned with the pixel size.

Pixel timing consists of two phases: reset phase and sensing phase, as depicted in Fig. 2.5b. The whole sensor will be placed in a solution condition, and a reset phase will be first conducted to alleviate drift and trapped charge effect. Then in sensing phase, the hydrogen reactions on the passivation surface will charge and accumulate in the floating gate. pH-TVC readout scheme is applied to enlarge small pH change, which is then turned into digital data by ADC. One row of pixels are measured in parallel, and digital outputs are first stored in SRAM groups, then read out column by column controlled by clock. Compared to circuit operation time, chemical reaction on the passivation surface is much slower, so that the system should continuously detect the solution to track surface potential trends.

2.2.4 Results and Discussions

The chip implemented in standard 65-nm CMOS process occupies an area of 5 mm × 5 mm. Chip photo, testing setup, and design specifications are summarized



Process	1P6M 65nm CMOS
Die size	5 x 5 mm^2
Pixel size	4.4×4.4 μm^2
Pixel chemical sensing area	15 μm^2
Total pixel array	512(row)×576(column)
pH sensitivity	123.8 mV/pH
ADC resolution	10 bit
Frame rate	375 fps (maximum)
Power supply	2.8 V
Power consumption	33mW @ 15 fps

Fig. 2.6 Chip photo, testing setup, and design specifications

Table 2.1 Summary of 65 nm CMOS ISFET sensor

Pixel pitch	4.4 μm
ISFET size (W/L)	0.4 μm/0.28 μm
Chemical sensing area	3.9 μm × 3.9 μm
C_{pass}	0.39 fF
C_{ox}	0.69 fF
Total pixel array size	512 × 576
ADC	10 bit
Readout clock	200 MHz
Frame rate	375 fps

in Fig. 2.6. The passivation layer is comprised of series of Si_3N_4 and SiO_2 layers, and the overall passivation capacitance per area is 0.026 fF/μm^2. The oxide capacitance per area is 6.16 fF/μm^2. The depletion region capacitance C_d and parasitic capacitances are dependent on ISFET gate voltage. Parameters of ISFET pixel and specifications of the ISFET sensor are summarized in Table 2.1.

Chip surface except the sensing area is covered by a waterproof encapsulation. Pixel sensitivity is calibrated using pH value of NaOH and HCl mixture. Measured results are shown in Fig. 2.7, of which source follower method has 6.3 mV/pH sensitivity, whereas pH-TVC method has 123.8 mV/pH sensitivity, both in 4.4 μm pixel pitch. Nearly 20-time improvement is observed.

By employing a novel pH-TVC scheme, the ISFET output voltage to pH sensitivity can be largely improved, regardless of capacitive division effect introduced by the passivation capacitance in CMOS process. One can observe 123.8 mV/pH sensitivity and 375 fps readout speed, which can be promising towards DNA sequencing.

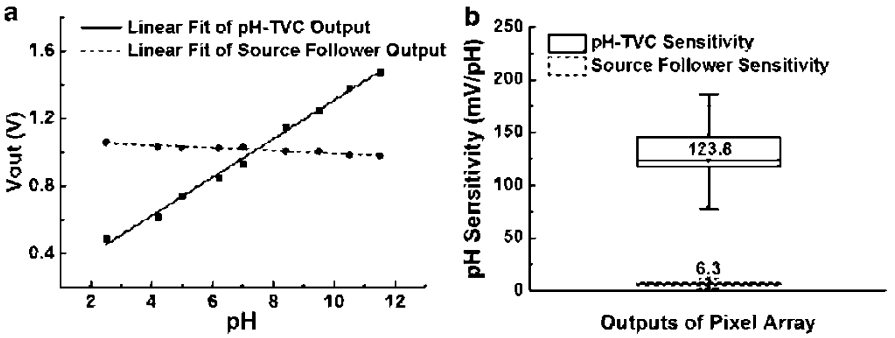


Fig. 2.7 Measured results: (**a**) pH to output voltage curve of source follower and pH-TVC. (**b**) Mean pH sensitivity of pixel arrays read out by the two circuits: 6.3 and 123.8 mV/pH, respectively

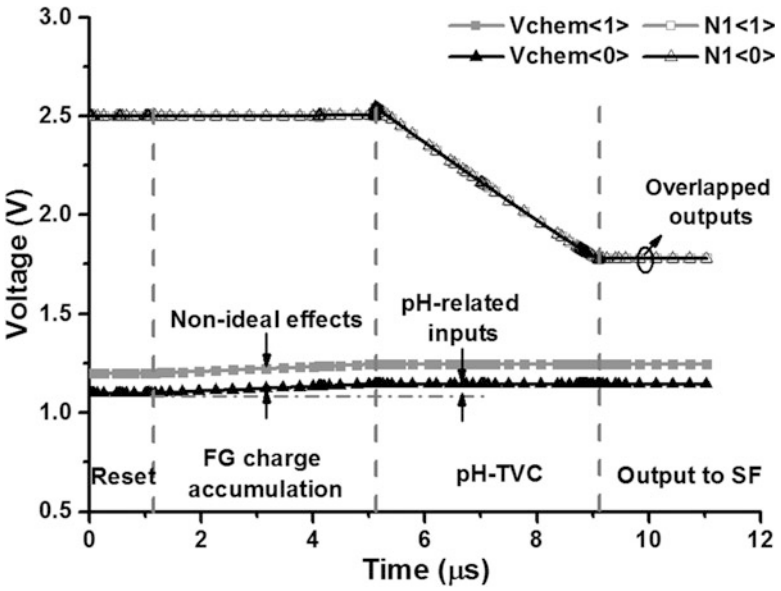


Fig 2.8 System pH-TVC results with reduced drift and trapped charge effects

Additionally, the system is immune to drift and trapped charge effects as shown in Fig. 2.8. Drift and trapped charge may lead to large and unpredictable voltage deviation [17]. Nevertheless the system can still identify the pH change. It is attributed to the reset device in each pixel, since trapped charge can be removed and drift can be eliminated.

Techniques to improve ISFET sensitivity are summarized in Table 2.2. The proposed pH-TVC readout method presents the largest scaling factor (sensitivity/pixel size) that greatly facilitates large-arrayed sensor applications.

Table 2.2 Summarized techniques to improve sensor sensitivity

Ref.	[6]	[14]	[17]	[19]	**This work**
Process	0.18 μm CMOS	0.35 μm CMOS	0.35 μm CMOS	SOI	**65 nm CMOS**
Output signal	Digital	Inverter flip flop	Analog	Analog	**Digital**
Pixel size (μm^2)	10×10	95×200	60×70	20×20	$\mathbf{4.4 \times 4.4}$
Array size	64×64	Single device	Single device	Single device	$\mathbf{512 \times 576}$
pH sensitivity (mV/pH)	103.8	3700	53.1	258	**123.8**
Scaling factor ((mV/pH)/μm^2)	1.038	0.194	0.013	0.645	**6.39**

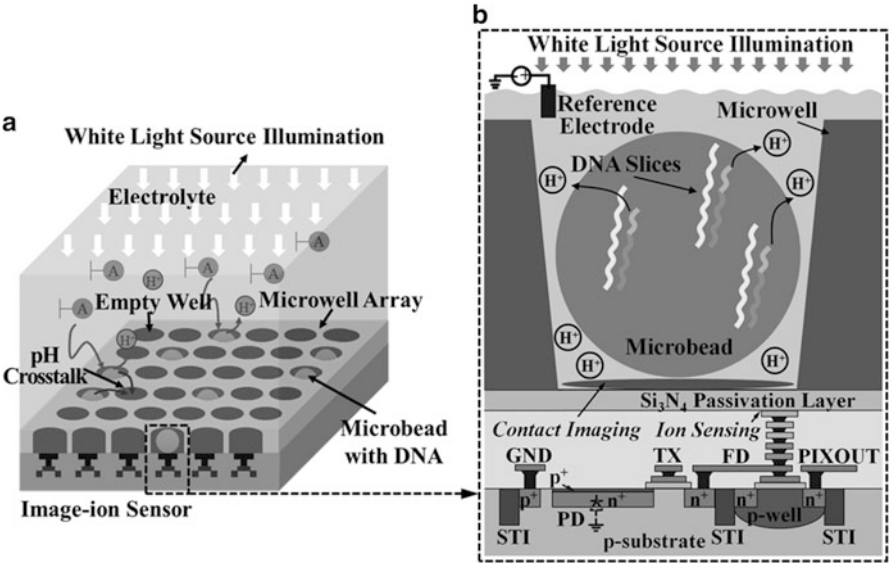


Fig. 2.9 (**a**) Proposed dual-mode sensor to deal with pH crosstalk. (**b**) Cross section of dual-mode pixel with microbead contact imaging and ion sensing

As a conclusion, the developed CMOS ISFET sensor with pH-TVC readout is able to significantly improve the pH sensitivity faced by the traditional CMOS ISFET sensor.

2.3 High Accuracy CMOS Dual-Mode Sensor

For the CMOS ISFET-based DNA sequencing, there is significant inaccuracy existed as illustrated in Fig. 2.9a. As the DNA-templated microbeads are scattered into microwell array by centrifuge spinning, the distribution of microbeads into

microwell array is unknown [12]. Thus, the measured pH responses have no correlation with the physical locations of microwells that contain microbeads. If there is no microbead in the microwell, due to crosstalk from neighboring microbeads in the solution, it will lead to false pH value reported. To tackle this problem, a dual-mode image-ion sensor is introduced by correlating local pH values with the locations of microwells filled with microbeads. The cross-sectional view of the proposed dual-mode pixel is shown in Fig. 2.9b. Note that n^+ and p^+ guard rings are placed around the pixel array to minimize the noise generated from the peripheral circuitry. Since the microbeads are in direct contact with the sensor surface, the imaging of the microbead distribution can be detected based on the contact imaging principle without lens [20]. One can determine the existence of microbead in optical mode and detect the pH value in chemical mode. As such, an accurate pH-image correlation map can be generated to prune the false pH values due to crosstalk for empty microwells.

2.3.1 Contact Imaging and pH-Based Sensing

In this paper, in addition to the pH sensing, we will introduce the optical sensing for the CMOS ISFET such that a dual-mode sensor can be developed with the removal of false pH reporting.

The conventional optical microscope imaging systems require intermediate bulky lens for magnification, which usually constrains the size, weight, and cost with the difficulty of miniaturization. One promising solution is the use of contact imaging, which directly couples the CMOS image sensor array with the sample of interest in small proximity (or contact), as shown in Fig. 2.10. As such, the sample image can be captured by directly projecting light through it with a detected

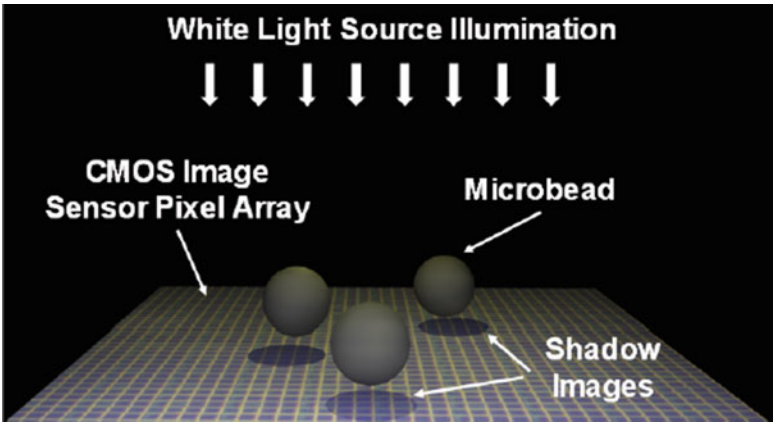


Fig. 2.10 Contact imaging principle: with light source illuminated from above, the contact shadow images of microbeads can be captured by the sensor underneath

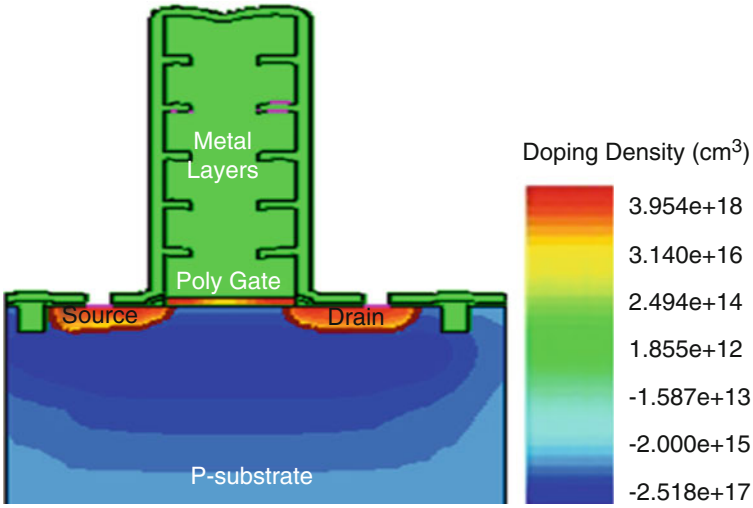


Fig. 2.11 Cross-sectional structure of the ISFET as modeled in Sentaurus TCAD. The electron concentration during ISFET operation is indicated by coloring

shadow. Contact imaging is a kind of near-field sensing without optic lens [21]. As such, contact imaging systems have different geometrical constraints over spatial resolution compared with lens based imaging. In conventional optical imaging systems, the image resolution is determined by the number of pixels in the photo detect array as the scene is entirely projected to the sensor array by optics. By increasing the number of pixels, the spatial resolution for the conventional imaging systems can be increased. Differently in the contact imaging, as the image is directly projected from the object to the image sensor array, the resolution is mainly determined by the pixel dimension as well as proximity distance. Thus, the contact imaging is quite suitable for miniaturized biomedical applications to detect objects such as microbeads used in DNA sequencing. Thereby, if one can leverage a dual-mode ISFET sensor with both pH sensing to detect H^+ at one microbead and also contact imaging to detect the existence of microbead, the false pH reporting problem of the existing ISFET sensor can be resolved during the DNA sequencing.

Moreover, the cross-sectional structure of the ISFET modeled is shown in Fig. 2.11. The electron concentration is indicated by coloring corresponding to the bar graph on the right. As the results shown in Fig. 2.12a, when we increase the concentration of negative charge donor from $10^{16}/cm^{-3}$ to $10^{40}/cm^{-3}$ with acceptors = 0, the threshold voltage V_T has a corresponding linear reduction from 1.11 to 0.06 V. As shown in Fig. 2.12b, when increasing the concentration of positive charge acceptor from $10^{16}/cm^{-3}$ to $10^{36}/cm^{-3}$ with donor = 0, the V_T has a corresponding linear increase from 1.13 to 1.74 V. A natural logarithm scale for charge concentrations is used. As such, although the electrolyte is not directly modeled, the effect of changing the surface charge will cause the changes in ISFET transfer characteristics and linearly modulate the ISFET threshold voltage V_T, which is the basic principle of ISFET-based pH sensing.

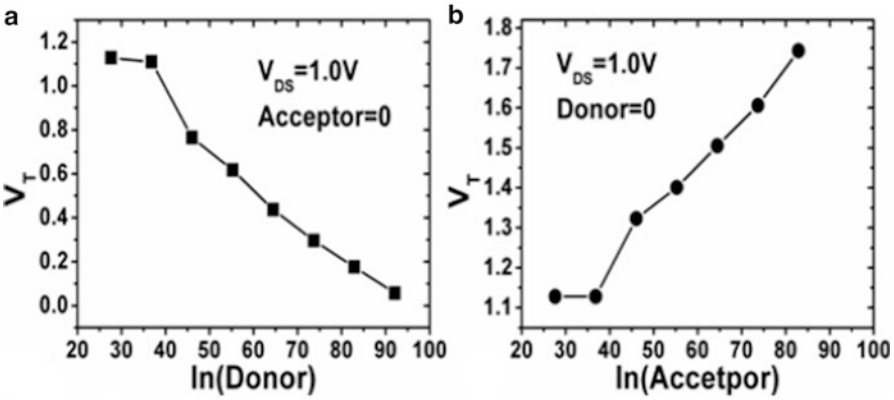


Fig. 2.12 ISFET device simulation results showing the threshold voltage VT change

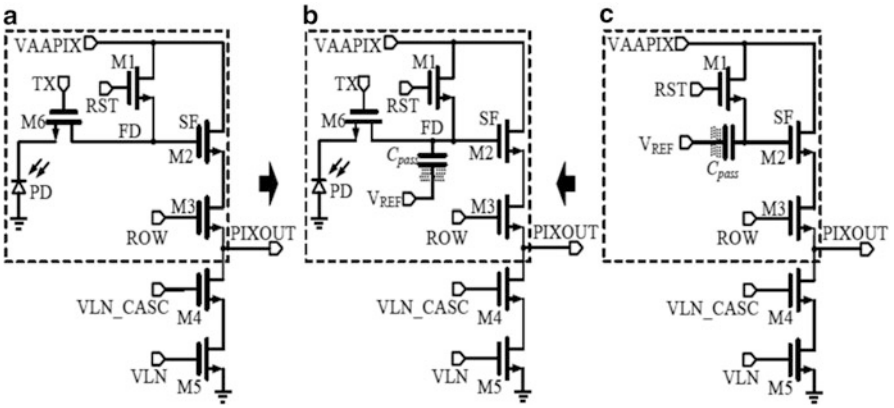


Fig. 2.13 Dual-mode pixel is the combination of 4T-CIS pixel and ISFET pixel. (**a**) 4T-CIS pixel. (**b**) Dual-mode pixel. (**c**) ISFET pixel

In this chapter, we will show a dual-mode sensor based on the CMOS image sensor. The widely used four-transistor CMOS image sensor (4T-CIS) pixel structure is shown in Fig. 2.13a. Photodiode (PD) collects protons and converts them to electrons. The collected charges are closely related to light intensity under a certain integration time. Therefore, microbeads with shadow images can be easily recognized through the detected output voltages. To further have an ISFET chemical sensing, a combined dual-mode pixel structure is given in Fig. 2.13b, where M2 in 4T-CIS pixel (the source follower) also functions as an ISFET device, and the polygate of M2 is connected all the way to the top metal as a sensing plate.

The cross-sectional view of the proposed dual-mode pixel is shown in Fig. 2.9b. Each pixel is in dual-mode to correlate the local pH value to the existence of one microbead detected by the contact imaging. Therefore, the false pH value reporting problem can be pruned in this design.

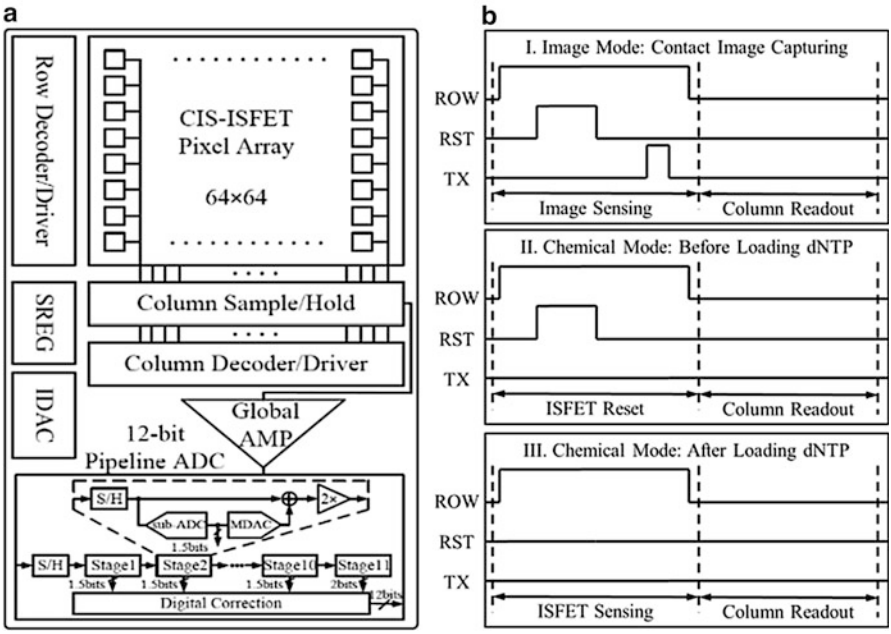


Fig. 2.14 (**a**) Top architecture of the dual-mode sensor. (**b**) Pixel operation diagram of image mode and chemical mode

2.3.2 Top Architecture and Operation

Top architecture of the dual-mode sensor is illustrated in Fig. 2.14a, including a 64×64 dual-mode pixel array, sample-and-hold (S/H) circuit and global switched-capacitor operational amplifier for CDS readout, 12-bit pipelined ADC, row/column decoders, and so on. Basically, there are two sensing modes: image mode and chemical mode.

In the image mode, photodiode (PD) first collects photons and converts them to proportional electrons, whose drifting generates the photocurrent. As the intrinsic junction capacitor of PD can store the generated charges, after a certain integration period, the intensity of incident light carried by the amount of charges is translated to a voltage signal. The charges can be transferred to floating diffusion (FD) by turning on "TX" switch of M6. As such, the voltage signal indicating the shadow image of microbead is detected through contact imaging. Then, the corresponding voltage signal for the optical image is buffered by SF (M2) and read out to PIXOUT node through its source under the control of "ROW" select-signal of M3. Since there are multiple rows of pixels that share the same PIXOUT line, the row-select transistor M3 is used to isolate different pixel outputs, and is enabled only when the row is selected for readout. The cascade current source (M4 and M5) provides biasing current and is shared by the whole column for better current matching.

In the chemical mode, the poly-gate of SF (M2) is all-the-way connected to the top metal and Si_3N_4 passivation layer, acting as ion-sensitive membrane of ISFET. Since the change of ion (H^+) concentration (or pH) can cause the proportional V_T shift of the SF, the corresponding voltage signal is correlated to the pH value that is read out through the source of SF. Note that although ISFET pixel has a switch to the floating gate, the TX leakage has been reduced through process optimization from the CIS aspect. As the cross-sectional pixel layout shown in Fig. 2.9b, a completely depleted pinned photodiode pixel is used, which consists of a pinned diode (p^+-n^+-p) to reduce the surface-defect noise due to dark current. The depletion layer of a pinned photodiode stretches almost to the Si–SiO_2 interface, which is perfectly shielded by the p^+ layer that keeps the interface fully filled with holes, making the leakage extremely low [22].

Pixel timing diagram is given in Fig. 2.14b. Firstly, image mode is conducted to identify microbeads' locations represented by row and column address. After that, chemical mode is utilized for pH detection by turning off TX. Before loading any nucleotides, ISFET gate is reset to high voltage to initiate a uniform sensing condition. During pH sensing phase, dNTPs are added sequentially, and local pH change is converted to output voltage by ISFET. As a result, local pH value is correlated with the microbead address. Therefore, one can achieve an accurate DNA sequencing by removing false pH value through such a CMOS dual-mode sensor.

2.3.3 Results and Discussions

The proposed dual-mode ISFET sensor is fabricated in standard TSMC 0.18 μm CIS process. After fabrication, the chip is packaged in a 100-pin Pin Grid Array (PGA) package with a size of 33.5 mm × 33.5 mm. As the experimental processes need to be conducted in aqueous environments, proper encapsulation of the sensor chip is necessary to protect the circuits. Thus, we use epoxy to encapsulate the whole chip with the sensing pixel array area open only, as shown in Fig. 2.15b, c. Meanwhile, the bonding wires and bonding pads are also covered by epoxy. To retain aqueous samples on the top of sensor chip, in addition, a 3D-printed plastic reservoir that just fits the PGA package is mounted on the package with epoxy to fill the gap at all four sides. The plastic reservoir is also designed to be able to fix the Ag/AgCl reference electrode. The package chip is then mounted on a specially designed printed-circuit-board (PCB) through a 100-pin PGA socket. The PCB, which is further connected with a Xilinx Virtex-6 XC6VLX240T FPGA demo board [23], is designed to provide power supply and digital timing control signals to the sensor chip. We measured the electrochemical characteristics of the chip under the control of a MATLAB-based (MathWorks, Natick, MA) Graphical User Interface (GUI). The chip micrograph with architecture and testing system is shown in Fig. 2.15. The design specifications are summarized in Table 2.3.

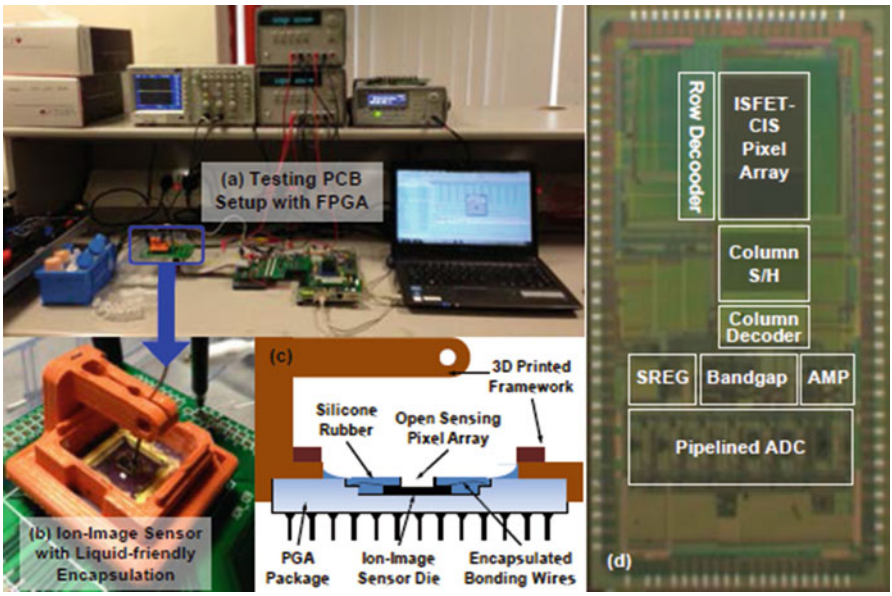


Fig. 2.15 (**a**) FPGA based testing system setup. (**b**) Ion-image sensor with liquid friendly encapsulation with 3D-printed plastic reservoir attached on the chip PGA package. (**c**) Cross-sectional view of the encapsulated packaging strategy. (**d**) Micrograph photo of the dual-mode sensor chip

Table 2.3 Specifications of dual-mode sensor

Parameters	Specifications
Process	Standard TSMC 0.18 μm CIS
Pixel type	Dual-mode (image and chemical)
Pixel size	10 μm × 10 μm
Pixel optical sensing area	20.1 μm^2 (fill factor = 20.1 %)
Pixel chemical sensing area	22.3 μm^2 (fill factor = 22.3 %)
Array size	64 × 64
Die area	2.5 × 5 mm
ADC ENOB	11.4 bits
ADC SNDR	70.35 dB
FPN	0.3 %
Frame rate	1200 fps
Total power consumption	32 mA @ 3.3 V

Firstly, the correlated contact image and pH map of microbeads are shown in Fig. 2.16. The microbeads of 45 μm diameter are used (Product# 07314-5, Polysciences, Warrington, PA). Note that we have not fabricated the microwell array on top of the image sensor die to correspond each microwell with an ISFET pixel. Thus, a relatively larger microbead compared with 10 μm pixel size is

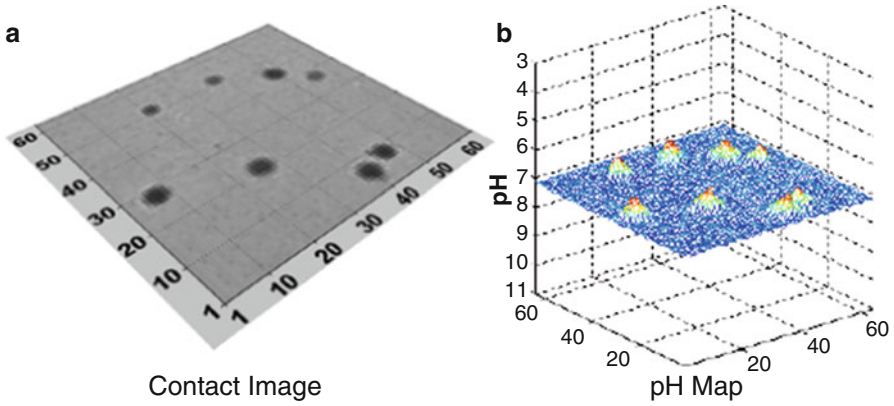


Fig. 2.16 Correlated maps of distributed microbeads: (**a**) contact images and (**b**) pH values

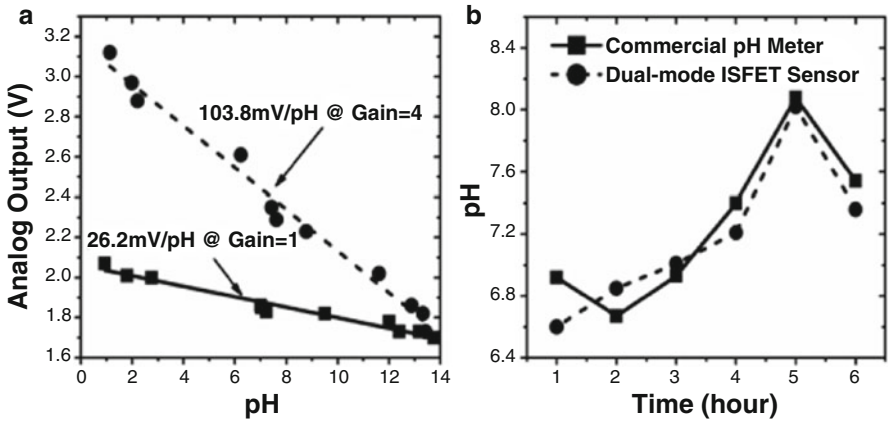


Fig. 2.17 Measurement results: (**a**) pH sensitivity of dual-mode ISFET sensor, and (**b**) the comparison with commercial pH meter for bacteria (*E. coli*) culture solution with glucose at different time intervals

selected such that the contrast of shadow imaging can be better. With the contact shadow imaging, the image size of microbead takes up about a 5×5 pixel array area. Due to the diffraction effect, the center pixels show darker intensity and the pixels near the boundary show lighter intensity. For the proof-of-concept verification, the microbeads are first diluted and prepared in acid solution as they are ideally suited for protein binding using passive adsorption techniques, and then dropped onto the sensor surface to test the local pH changes. The contact image determines the existence of microbeads and provides their addressed distribution. The exposure time of the contact imaging is 160 μs. The pH map is thereby locally associated with microbeads by pruning out those uncorrelated pH data. Due to the diffusion effect, the pH map at microbead locations shows a pattern similar to normal distribution.

To characterize the pH sensing capability of the dual-mode sensor, the pH sensitivity is tested and the measurement results are shown in Fig. 2.17a. The pH

Table 2.4 Comparison of state-of-the-art ISFET sensors

Ref.	[11]	[24]	[25]	[26]	**This work**
Process	5 μm non-CMOS	0.35 μm modified CMOS	0.35 μm standard CMOS	0.35 μm standard CMOS	**0.35 μm Standard CMOS**
Pixel size (μm^2)	200×200	12.8×12.8	10.2×10.2	20×2	$\mathbf{10 \times 10}$
Array size	10×10	16×16	64×64	8×8	$\mathbf{64 \times 64}$
Frame rate	30 fps	–	100 fps	–	**1200 fps**
pH sensitivity (mV/pH)	229	46	20	37	**26.2 (gain = 1) 103.8 (gain = 4)**
Dual-mode	No	No	No	No	**Yes**

of solution is changed by adding HCL and NaOH. The pH readout sensitivity of ISFET by CIS process is measured as 26.2 mV/pH with amplifier gain = 1 and as 103.8 mV/pH with amplifier gain = 4. The device sensitivity at gain = 1 is somewhat lower than the commonly observed response of 45–56 mV/pH for Si_3N_4, this can be due to the low-pressure chemical vapor deposition (LPCVD) technique for Si_3N_4S at low temperature, which generally causes low-density and porous passivation layer. It can be optimized by the LPCVD at a high temperature or do additional depositions, which are still standard CMOS process.

The CMOS ISFET sensor chip is also calibrated by testing the pH change of a bacteria (*Escherichia coli*) culture solution at different time intervals and comparing with commercial tool. By extracting the sample solution of the bacteria culture for testing at 1–6 h time intervals, the measurement results by the dual-mode sensor can correlate well with one commercial pH meter (Checker, Hanna Instruments, RI, USA) in Fig. 2.17b.

Lastly, the comparisons with the state-of-the-art ISFET sensors are summarized in Table 2.4. The proposed dual-mode sensor shows the state-of-the-art results: 10 μm pixel pitch, 64×64 pixel array, fast frame rate of 1200 fps, and readout sensitivity of 103.8 mV/pH in standard CIS process.

As a conclusion, the developed CMOS dual-mode sensor is able to significantly improve the pH detection accuracy faced by the traditional CMOS ISFET sensor.

2.4 CMOS THz Metamaterial Sensor

DNA microarray is widely used in genotyping, where thousands of artificially produced DNA probes are attached to a glass or a plastic plate. When the array is exposed to a solution with DNA samples, the matching probes hybridize with target DNA strands, so that target DNA sequences can be inferred by their given complementary probes. Unlike the detection of nucleotides' order in whole-genome sequencing, the identification of hybridization that relies on labeling and optical

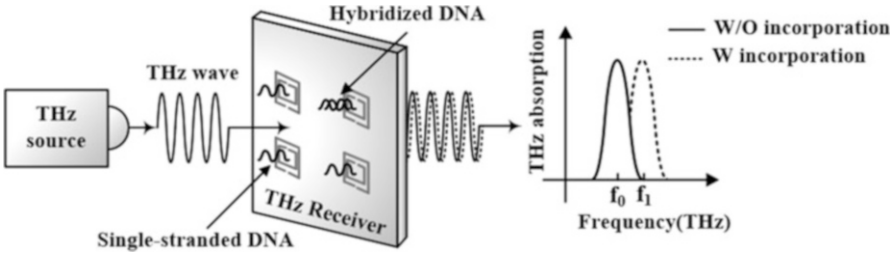


Fig. 2.18 Diagram of THz-based genotyping. Hybridized DNA strand can be identified from resonant frequency shift to single-stranded DNA probes

detection is the main issue to be solved in genotyping. Recently, CMOS THz-based electrical detection sensor has aroused great interest in DNA detection, especially genotyping.

The development of THz-based DNA sequencing method has become one possible new solution to tackle the aforementioned challenges [27, 28]. THz-sensing has attracted a lot of research activities in the past decades as numerous materials can exhibit unique spectrum signatures in THz range. In [29], Brucherseifer et al. first demonstrated that the binding state of DNA can be directly probed through its complex refractive index in THz range. A free-space detection and an integrated detection based on planar waveguides are initially realized [29, 30]. To increase the sensitivity and reduce the amount of sample needed to characterize the sequence, metamaterial THz sensors based on electrical/magnetic resonance are proposed with subwavelength scattering [31]. More recently, one silicon nano-sandwich pump device was proposed that can provide both the excitation of the DNA strands' self-resonant modes and feedback for current–voltage measurements to identify the strands' sequences [32].

Figure 2.18 illustrates the diagram of THz-based genotyping. Several single-stranded DNA probes and a hybridized double-stranded DNA are shown on the surface of detector. The probes are illuminated with THz wave generated by on-chip source and the resonant frequencies are recognized by THz detector. It is reported that a probe at which hybridization takes place exhibits a reduced resonant frequency in THz range [33]. Thereby binding states of DNA strands can be recognized by THz-based sensor.

2.4.1 Metamaterial-Based Source and Detector

The block diagram of one 140-GHz signal source is illustrated in Fig. 2.19. The input is a 35-GHz reference signal, which is doubled to 70-GHz injection signal. The output signals of four 70-GHz zero-phase oscillator unit-cells are first frequency-doubled, and then combined at the center of 70-GHz coupled oscillator network (CON).

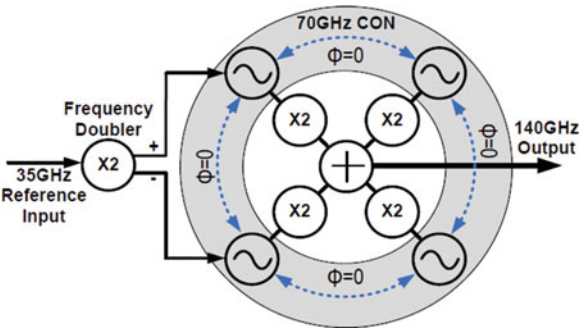


Fig. 2.19 Block diagram of 140-GHz signal source

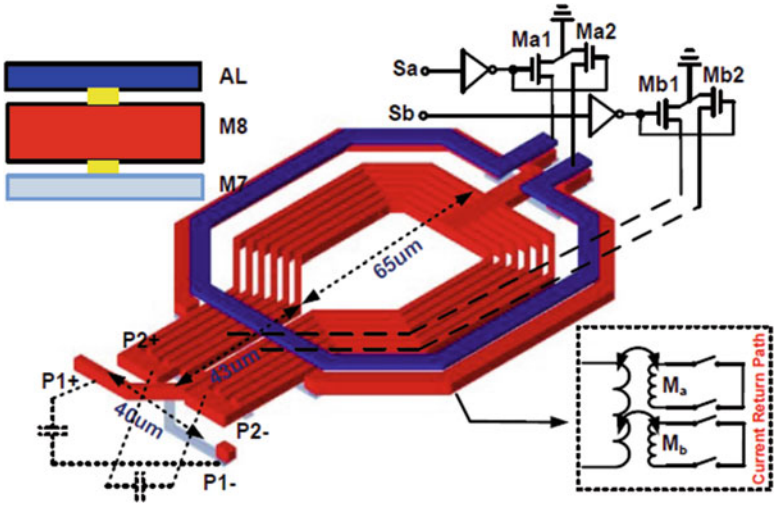


Fig. 2.20 Schematic and layout of on-chip MPW base oscillator unit-cell with inter-digital coupled T-line

Figure 2.20 shows the schematic and layout of on-chip 70-GHz MPW based oscillator unit-cell with inter-digital coupled T-line implemented in the top most copper layer (M8) and parasitic capacitances from transistors in 65-nm CMOS process.

The schematic of the 70-GHz CON is shown in Fig. 2.21. Four 70-GHz MPW based oscillator unit-cells are serially connected in a closed-loop form to generate four in-phase differential output signals at locations A, B, C, and D with the same magnitude and frequency, which is injection-locked to the 70-GHz reference signal. Negative resistance is formed by cross-coupled NMOS pair (e.g., M1 and M2), which can compensate the energy loss in each unit-cell when forming the oscillation signal. A central symmetrical layout is deployed and all active devices are placed as closed as possible to the geometrical center of CON to reduce process variation.

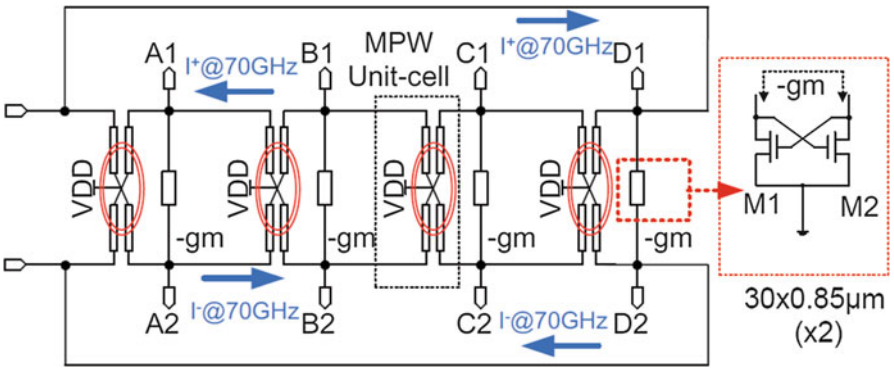


Fig. 2.21 Schematic of injection-locked 70-GHz CON with 4 MPW unit-cells

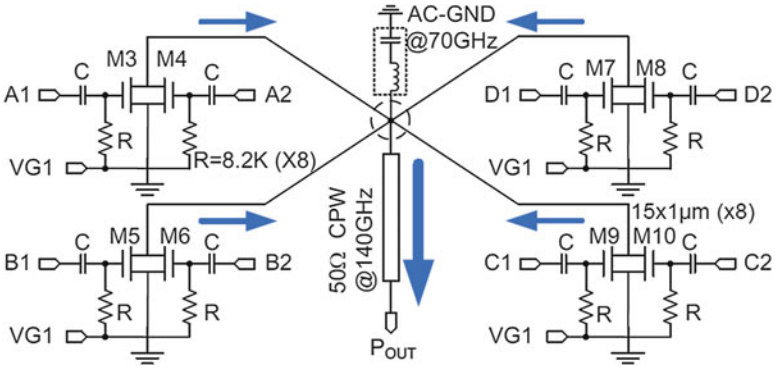


Fig. 2.22 Schematic of 70-GHz to 140-GHz frequency doublers with center combined output

The schematic of 70-GHz to 140-GHz push–push frequency doubler with center combined output is shown in Fig. 2.22. The 70-GHz differential output signals at A, B, C, and D are coupled to the push–push frequency doubler by 28 f. DC-block capacitors. Therefore, the frequency conversion efficiency can be maximized by externally baising all the frequency doublers to the threshold level (VG1). The resulting four in-phase 140-GHz output signals are tied together directly to generate a high power output signal at the center.

Figure 2.23 shows the schematic of 30-GHz to 70-GHz reference frequency doubler. A transformer-based balun is employed to generate a differential 35-GHz reference signal to drive M3 and M4. In addition, another Marchand balun with inter-digital coupling is deployed at 70 GHz to have balanced differential outputs as well as low insertion loss.

In addition to the metamaterial-based 140-GHz source, one can build the metamaterial-based 140-GHz detector such as differential transmission-line loaded split-ring resonator (DTL-SRR). Layout for CMOS on-chip DTL-SRR is shown in Fig. 2.24a, in which stacked SRRs with the same dimensions of $24 \times 24\ \mu m^2$ from

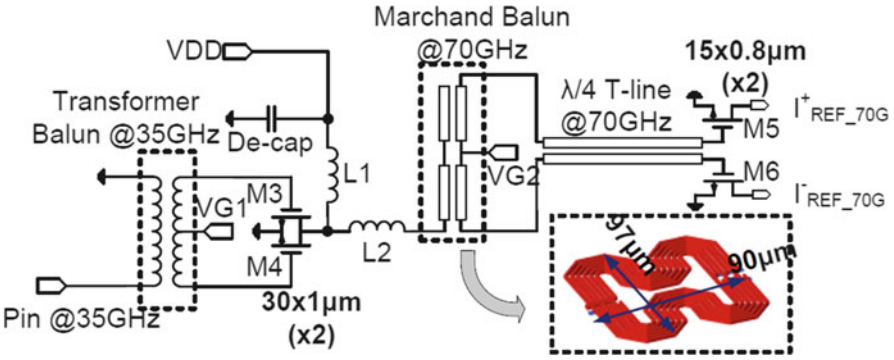


Fig. 2.23 Schematic of injection-locked 70-GHz CON with 2 MPW unit-cells

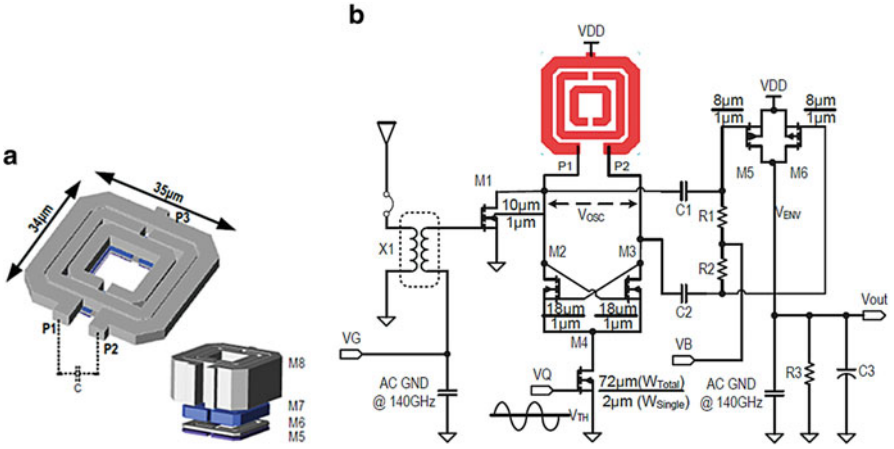


Fig. 2.24 (**a**) Layout for CMOS on-chip implementation of DTL-SRR for 140-GHz SRX. (**b**) Schematic of CMOS 140-GHz SRX with DTL-SRR

M5 to M8 are illustrated. All SRRs are closely coupled to the same host T-line implemented in the top most metal layer (M8). The overall size of the proposed DTL-SRR is $35 \times 34\ \mu m^2$.

The schematic of 140-GHz DTL-SRR-based CMOS receiver is shown in Fig. 2.24b. DTL-SRR is connected to a differential negative resistance formed by cross-coupled NMOS (M2 and M3). The remaining circuit consists of a common source input buffer (M1) for current injection and an envelope detector formed by M5 and M6. The common source stage (M1) is designed for input signal injection and also reverse isolation from the oscillator to the input. The size of M1 is optimized to minimize parasitic capacitance and input mismatch. Transformer-based matching network is used as the electrostatic discharge (ESD) protection when M1 is integrated with the antenna. The detected envelope 140-GHz signal is directly averaged by an on-chip low-pass filter formed by R3 and C3 at the output.

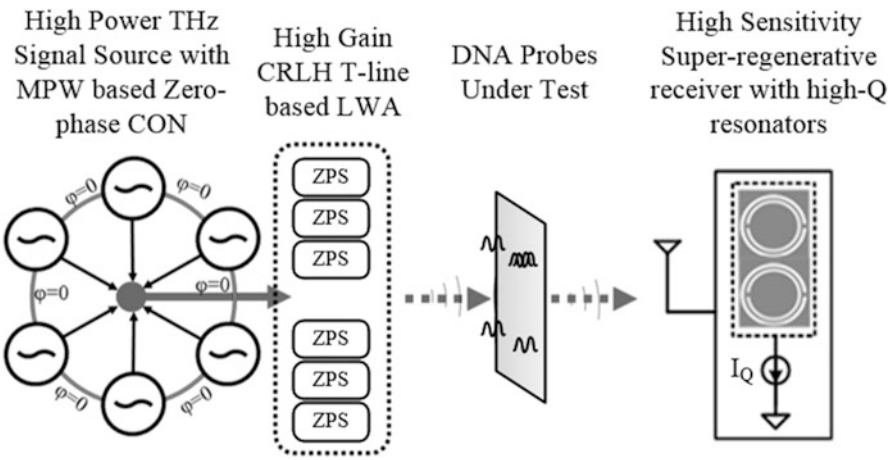


Fig. 2.25 Realization diagram of metamaterial-based THz-sensing system

2.4.2 Top Architecture and Operation

High performance THz-sensing system can be constructed by on-chip metamaterial-based signal source, receiver, and antenna. The block diagram is shown in Fig. 2.25. A high power THz signal source is firstly generated by magnetic plasmon waveguide (MPW) based zero-phase CON and then radiated by the composite right-/left-handed (CRLH) T-line based on-chip leaky wave antenna (LWA). After penetrating through the DNA probes mounted on high-Q resonator array, the resulting THz signal is received by a high-sensitivity super-regenerative differential transmission-line (T-line) loaded with split-ring-resonator (DTL-SRR). In this article, a high power CON based signal source and a DLT-SRR based super-regenerative receiver (SRX) are proposed, which combined together consist a high sensitive and wide band THz-sensing system at 140 GHz. Different from the optics-based THz-sensing systems that are bulky, expensive, lack of portability with low detection resolution by electro-optic sampling techniques, the proposed CMOS THz-sensing system demonstrates both high-sensitivity and wide band THz sensor at 140 GHz.

2.4.3 Results and Discussions

The proposed injection-locked THz signal source is implemented in 65 nm CMOS RF process. The chip photo is shown in Fig. 2.26a, which occupies a total area of $750 \times 550\ \mu m^2$. It was measured on a probe station with a six-pin DC biasing probe, a normal GSG probe for reference signal input and a D-band waveguide to GSG probe for output, which is connected to R&S FSUP signal source analyzer with a

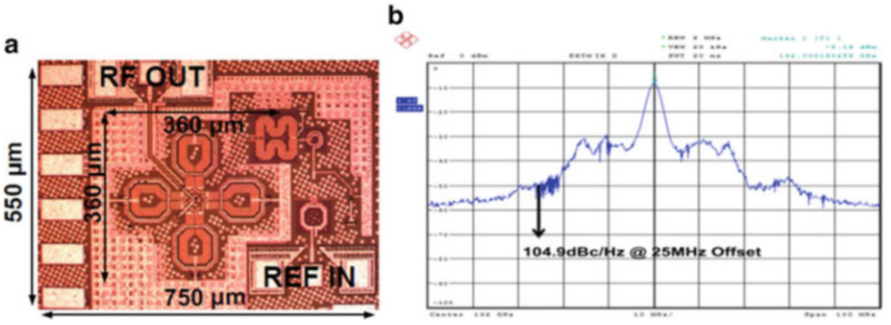


Fig. 2.26 (**a**) Chip photo of the mm-wave source in 65 nm CMOS. (**b**) Measured output spectrum of signal source at 132-GHz

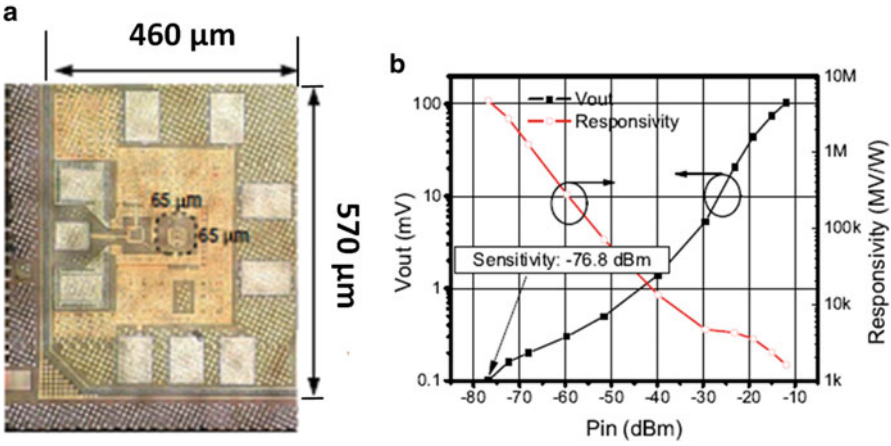


Fig. 2.27 (**a**) Chip photo of DTL-SRR in 65 nm CMOS. (**b**) Output voltage (Vout) and responsivity vs. input power (Pin)

D-band waveguide harmonic mixer, of which the absolute measured power level is calibrated by a VDI PM4 power meter. Operating from a 1.2 V supply, the power consumption of the CON core of signal source is 145 mW, while the input frequency doubler is 3.8 mW.

Figure 2.26b shows the spectrum of 132-GHz output signal when locked to a 5 dBm 33-GHz reference signal. A lack of de-coupling capacitors in the DC probe induces a raised noise floor within ±25 MHz coupled from DC power supplies. In such case, a phase noise of −104.9 dBc/Hz is measured at 25 MHz offset. Moreover, the maximum power density of the proposed signal source is 26.9 mW/mm^2.

Moreover, the proposed DTL-SRR-based mm-wave CMOS receiver is implemented in 65 nm CMOS RF process. The chip photo is shown in Fig. 2.27a, which occupies a total area of 570 × 460 µm^2, and a core area of 0.0085 mm^2. The whole test board is placed on probe station for SRX measurement. The RF input

signal is provided by a VDI D-Band signal generator, of which the output power is calibrated in the range of −85 ~ −10 dBm by R&S FSUP signal source analyzer with a D-band waveguide harmonic mixer. A 12.5 MHz sinusoid quench-control signal is applied from function generator (Agilent 33250a) with voltage sweep range of 0~400 mV. Operating from a 1 V supply, the receiver consumes 6.2 mW.

Figure 2.27b shows normalized Vout against Pin as well as the responsivity, where the receiver sensitivity (S) and the maximum responsivity are observed as −76.8 dBm and 4.82 MV/W, respectively. And the NEP is calculated to be 0.9 fW/pHz. A near linear relationship between Vout and Pin is observed when the input power is below −40 dBm, which can be utilized in post-data processing to generate mm-wave sensing. With the significantly improved receiver sensitivity, the proposed CMOS mm-wave receiver can be used as high-resolution resonance shift detector in DNA sequencing.

As a conclusion, the proposed metamaterial-based CMOS THz sensor has great potential to achieve both high sensitivity and wide spectrum for the future non-contacted electrical/magnetic sensing in DNA sequencing.

2.5 Conclusion

DNA sequencing has a great promise to life sciences, biotechnology, and medicine. The first sequencing method is introduced by Sanger in 1970s. Since then new technologies have been developed from 1st to 3rd generation—Illumina and Ion Torrent sequencing. More recently single-molecule nanopore sequencing that is regarded as the coming 4th generation has aroused industry's interest. The existing DNA detection methods are summarized in Table 2.5, including the contributions

Table. 2.5 Summary of DNA detection methods

Category	Method	Sample labeling (complex)	Optical system (bulky, costly)	CMOS compatible (scalable)	Throughput
Sequencing	Sanger	Yes	Yes	No	Low
	454	No	Yes	No	Moderate
	SMRT	Yes	Yes	No	Moderate
	SOLiD	Yes	Yes	No	Moderate
	Illumina	Yes	Yes	No	High
	Nanopore	No	No	Yes	Potentially high
	pH-based (ISFET[a], Dual-mode[a])	No	No	Yes	Potentially high
Genotyping	Microarrays	Yes	Yes	No	N/A
	THz sensing[a]	No	No	Yes	N/A

[a]Method with sensor presented in this work

from this chapter: the CMOS pH-TVC ISFET sensor; the CMOS dual-mode sensor; and the CMOS THz sensor. Such a lab-on-CMOS integration based approach has shown great potential for a label-free personalized DNA sequencing with low cost in future.

References

1. Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467
3. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11 (1):31–46
4. Toumazou C, Shepherd LM, Reed SC et al (2013) Simultaneous DNA amplification and detection using a pH-sensing semiconductor system. Nat Methods 10(7):641–646
5. Jiang Y, Liu X, Huang X et al (2015) A 201 mV/pH, 375 fps and 512 × 576 CMOS ISFET sensor in 65 nm CMOS technology. In: 2015 IEEE custom integrated circuits conference (CICC), San Jose, CA, September, pp 1–4
6. Huang X, Wang F, Guo J et al (2014) A 64 × 64 1200 fps CMOS ion-image sensor with suppressed fixed-pattern-noise for accurate high-throughput DNA sequencing. In: 2014 symposium on VLSI circuits digest of technical papers, Honolulu, HI, June, pp 1–2
7. Huang X, Yu H, Liu X et al (2015) A dual-mode large-arrayed CMOS ISFET sensor for accurate and high-throughput pH sensing in biomedical diagnosis. IEEE Trans Biomed Eng 62 (9):2224–2233
8. Kumar S, Tao C, Chien M et al (2012) PEG-labeled nucleotides and nanopore detection for single molecule DNA sequencing by synthesis. Sci Rep 2(684):1–8
9. Bergveld P (2003) Thirty years of ISFETOLOGY: what happened in the past 30 years and what may happen in the next 30 years. Sens Actuators B Chem 88(1):1–20
10. Bausells J et al (1999) Ion-sensitive field effect transistors fabricated in a commercial CMOS technology. Sens Actuators B Chem 57:56–62
11. Milgrew M, Cumming D (2008) A proton camera array technology for direct extracellular ion imaging. In: Proceedings of the 2008. IEEE International Symposium on Industrial Electronics, Cambridge, July, pp 2051–2055
12. Rothberg JM et al (2010) Methods and apparatus for measuring analytes. U.S. Patent 20100301398 A1, Dec 2, 2010
13. Liu Y et al (2011) An extended CMOS ISFET model incorporating the physical design geometry and the effects on performance and offset variation. IEEE Trans Electron Devices 58(12):4414–4422
14. Al-Ahdal A, Toumazou C (2012) High gain ISFET based vMOS chemical inverter. Sens Actuators B 171:110–117
15. Premanode B, Silawan N, Toumazou C (2007) Drift reduction in ion-sensitive FETs using correlated double sampling. Electron Lett 43:1–2
16. Milgrew M, Cumming D (2008) Matching the transconductance characteristics of CMOS ISFET arrays by removing trapped charge. IEEE Trans Electron Devices 55(4):1074–1079
17. Hu Y, Georgiou P (2014) A robust ISFET pH-measuring front-end for chemical reaction monitoring. IEEE Trans Biomed Circuits Syst 8(2):177–185
18. Georgiou P, Toumazou C (2009) ISFET characteristics in CMOS and their application to weak inversion operation. Sens Actuators B Chem 143(1):211–217
19. Park JK et al (2014) SOI dual-gate ISFET with variable oxide capacitance and channel thickness. Solid-State Electron 97:2–7

20. Huang X et al (2014) A contact-imaging based microfluidic cytometer with machine-learning for single-frame super-resolution processing. PLoS One 9(8), e104539
21. Ji H et al (2007) Contact imaging: simulation and experiment. IEEE Trans Circuits Syst I Regul Pap 54(8):1698–1710
22. Fossum ER, Hondongwa DB (2014) A review of the pinned photodiode for CCD and CMOS image sensors. IEEE J Electron Devices Soc 2(3):33–43
23. Xilinx Virtex-6 XC6VLX240T FPGA (2014) [On-line] Available: http://www.xilinx.com/products/boards-and-kits/EK-V6-ML605-G.htm
24. Nemeth B et al (2012) High-resolution real-time ion-camera system using a CMOS-based chemical sensor array for proton imaging. Sens Actuators B, Chem 171–172:747–752
25. Chan WP et al (2010) An integrated ISFETs instrumentation system in standard CMOS technology. IEEE J Solid State Circuits 45:1923–1934
26. Manickam A et al (2012) A fully-electronic charge-based DNA sequencing CMOS biochip. In: 2012 Symposium on VLSI Circuits (VLSIC), Honolulu, HI, June, pp 126–127
27. Bolivar PH, Nagel M et al (2004) Label-free THz sensing of genetic sequences: towards 'THz biochips'. Philos Transact A Math Phys Eng Sci 362(1815):323–333
28. Huang X, Jiang Y et al (2015) A CMOS THz-sensing system towards label-free DNA sequencing. In: IEEE international conference on ASIC (ASICON), Nov 2015 (Invited Special Session), Sichuan, China, November, pp 1–4
29. Brucherseifer M, Nagel M, Bolivar PH, Kurz H, Bosserhoff A, Buttner R (2000) Label-free probing of the binding state of DNA by time-domain terahertz sensing. Appl Phys Lett 77 (24):4049–4051
30. Nagel M, Bolivar PH, Brucherseifer M, Kurz H (2002) Integrated THz technology for label-free genetic diagnostics. Appl Phys Lett 80(1):154–156
31. Zheng N et al (2013) Metamaterial sensor platforms for Terahertz DNA sensing. In: 2013 13th IEEE International Conference on Nanotechnology, Beijing, China, August, pp 315–320
32. Chernev AL et al (2015) DNA detection by THz pumping. Semiconductors 49(7):944–948
33. Cao C, Zhang J, Wen X et al (2013) Metamaterials-based label-free nanosensor for conformation and affinity biosensing. ACS Nano 7(9):7583–7591

Part II
Circuit Platforms for Smart Sensors

Chapter 3
Circuit Design in mm-Scale Sensor Platform for Future IoT Applications

Inhee Lee and Yoonmyung Lee

Abstract Since the emergence of the first computers in the 1940s, many different classes of computing systems, such as workstations, desktop PCs, and laptops, have been introduced to meet the ever-changing market needs, as predicted by "Bell's Law." Light-weight mobile computing devices were introduced in the early 2000s, and we expect to be surrounded by millions or trillions of small sensing/computing systems in the upcoming internet of things (IoT) era.

Sensor systems in the IoT era are expected to be several orders of magnitude smaller in volume than their predecessors, consistent with the general trend of increasing compactness observed as computing systems evolve. This means that computing systems with a volume on the order of cubic centimeters or even cubic millimeters are likely. Recent research shows that mm-scale systems can be realized with advances in low-power electronics design, packaging, and battery technologies. These miniature systems are expected to be the driving force for unprecedented IoT applications, such as implanted diagnosis sensors and pervasive environment monitoring sensors.

The key challenge for achieving mm-scale volume is to significantly reduce the power used by every component of a system due to the extremely limited amount of energy storage. Therefore, in this chapter, state-of-the-art low-power design strategies for a few core components which enable such mm-scale systems are reviewed. System-level design approach is also presented with a few recently demonstrated mm-scale sensing systems.

Keywords Low-power circuit • mm-Scale sensor • IoT • Real time clock • Microprocessor • Sensor interface • Energy harvester • Wireless transceiver • Voltage reference • Current reference

I. Lee (✉)
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
e-mail: inhee@umich.edu

Y. Lee (✉)
Department of Semiconductor Systems Engineering, Sungkyunkwan University, Seoul, South Korea
e-mail: yoonmyung@skku.edu

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_3

3.1 Introduction: Bell's Law and Size of Computing Systems

Since the emergence of electronic computing systems in 1940, the form-factor and size of computing systems have, with the help of advances in technology, undergone continuous changes to meet the unique requirements of target applications of the time. This trend was predicted by Gordon Bell in his early article [1] in which he claimed that a new class of smaller computers is developed approximately every decade by using fewer components or fractional parts of state-of-the-art computing system; this idea was formulated as "Bell's Law" [2] in 2008. As Bell's Law predicted, a significant downsizing of computing systems as well as transistors has been observed over the last few decades, as shown in Fig. 3.1.

Mainframe computers in the 1950s were often as big as cabinets and performed powerful and reliable bulk data processing such as census data and enterprise transaction data processing. The minicomputers introduced in the 1960s were as their name claims, smaller computers with reduced functionality and volume. In the 1980s, personal computers with further reduced computing capability suitable for individual usage were introduced at affordable prices. As the demand for mobility increased, laptop computers were introduced in the 1990s, and with advances in wireless networks and electronics manufacturing technology, hand-held portable electronics became commonplace around the beginning of the twenty-first century. These changes in computing systems observed over the years clearly confirm that

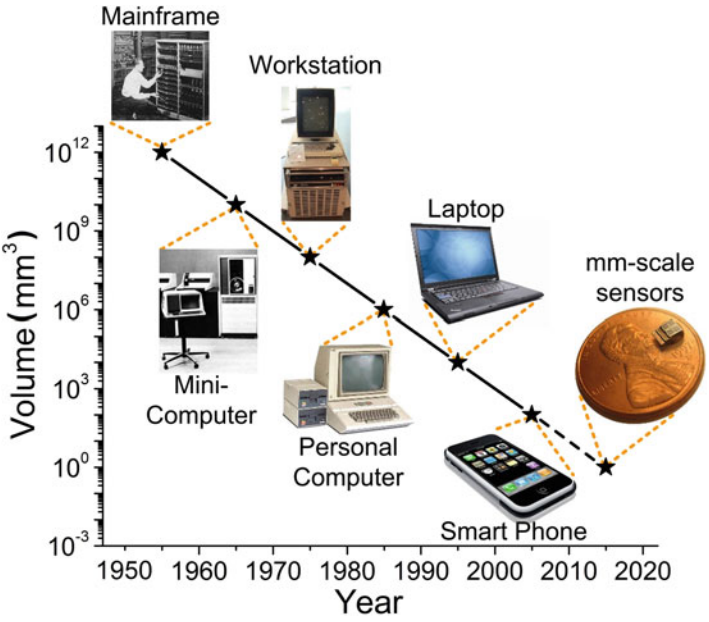


Fig. 3.1 Continuous scaling of computing systems as projected by Bell's Law

smaller and lighter computing systems were repeatedly introduced to meet the requirements of the time—affordability in the early days of computers, and mobility in more recent times.

The aforementioned observation provides some hints as to what the next generation of computing systems might look like. In fact, we are already surrounded by many small electronic devices that are smaller than cell phones and have small computing capabilities suitable for simple tasks. Wearable devices are good examples: they are made light and small so that they can be easily worn or born by users and perform simple health-monitoring or event-logging tasks. There are also many small electronic devices in the form of dongles and tokens that are useful for tagging or tracking objects, especially easily lost objects, such as car keys, remote controllers, and pets.

According to Bell's Law, the next computing platform will be even smaller than these dongles and tokens. Based on the volume reduction trend observed in the past decades, one would project that the next computing platform may be only a few mm^3. Recent research shows that the prototypes of many systems with various sensing modalities can be realized in tiny mm-scale volumes using today's technologies [3–6]. These systems, the size of a grain of rice, would benefit from their extremely high portability and could be deployed almost anywhere due to their tiny size. Although the function of mm-scale systems would be primarily sensing and simple data processing due to their extremely limited volume, encapsulating intelligence and sensing capability in such tiny volumes creates unique opportunities for whole new applications that were never possible earlier. This means that it is now easier to integrate "intelligence" to "things," creating many new opportunities for the upcoming internet of things (IoT) era.

A "smart dental brace" is a good example of a new application made possible with mm-scale systems. It could be created by attaching a small electronic system with tension-measuring capability to a dental brace. The sensors would provide immediate feedback to an orthodontist when he or she adjusts the tension of the brace during a patient's regular visit. It could also track and store data on how the tension has changed between the patient visits so that the orthodontist can make a better informed decision regarding how much tension adjustment to make.

Tiny implantable sensors are another good example of attractive future application for mm-scale systems. Implanted sensors can obtain much more precise diagnostic data compared with wearable or patch-type sensors since they can directly examine target objects such as organ tissue, blood, or other body fluids. However, invasive surgery is required for placing these sensors correctly, which is associated with a risk of infection as well as substantial financial cost. If these implanted sensors can be made with a volume of only a few mm^3, these sensors can be injected with syringes, dramatically reducing the risk and cost of implantation.

There have been several prototypes of mm-scale systems recently demonstrated, as shown in Fig. 3.2, helping us envision many attractive future IoT applications. A pressure-sensing system with a volume of $1.4 \times 2.8 \times 1.6\ mm^3$ was demonstrated in [3], and a temperature-sensing system with a similar volume was demonstrated in [4]. These sensing systems pave the way to realizing implantable diagnosis sensors

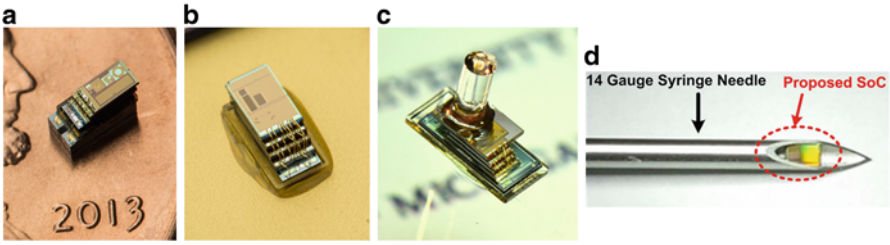


Fig. 3.2 Recently demonstrated mm^3-scale system prototypes: (**a**) pressure-sensing system [3], (**b**) temperature-sensing system [4], (**c**) motion-sensing system [5], and (**d**) ECG-sensing system-on-chip (SoC) [6]

for monitoring blood pressure or cancer tissue growth. They also facilitate wireless pressure/temperature monitoring for areas with limited access, such as the "inside" of sensitive machines, resulting in more reliable manufacturing process control in a factory. A motion detection and imaging sensor system with $2 \times 4 \times 4$ mm^3 volume is also demonstrated in [5] with solar energy harvesting to achieve energy autonomy in a bright environment. This opens up the possibility of implementing versatile motion detection and imaging systems that are not limited by space-constraints. An ultra-low power ECG-sensing system-on-chip (SoC) was also designed to detect arrhythmia with 64-nW power consumption [6]. The ECG-sensing SoC was 1.5×2.3 mm^2, which fits in a 14-gauge needle, as shown in Fig. 3.2d. The SoC can be combined with an mm-scale sensor form-factor that could be injected into patients for implantation without costly surgery.

For the rest of this section, the key challenges for achieving mm-scale volume are addressed in the next sub-chapter, and low-power circuit design strategies for each building block are introduced in the following sub-chapter.

3.2 Challenges for Enabling mm-Scale IoT Systems

Implementing an electronic system on an mm-scale requires a significant volume reduction compared with conventional electronic systems. Such a volume reduction in an electronic system can be achieved by focusing on three distinct parts: electronics, battery, and packaging.

Packaging for electronic systems is required for physical protection and electrical isolation of the system from the outside world. There are a range of packaging techniques suitable for mm-scale systems, such as molding with epoxy or encasing with metal/glass cases. However, since the details of these techniques are beyond the scope of this chapter, we will focus on the volume reduction challenges associated with electronics and batteries.

For electronics in an mm-scale system, the use of standalone components should be avoided since independently packaged electronic components, such as resistors,

capacitors, and discrete transistor devices, are already a few mm^3 in volume even in their smallest form (e.g., SMD components). By utilizing only components that can be integrated on silicon die, mm-scale volume can be achieved. Therefore, a new circuit design approach is required to replace the off-the-shelf components that are widely used for their desirable electronic characteristics, e.g., quartz crystals for clock generation and inductors for DC–DC voltage conversion.

There are a few battery technologies that can achieve mm-scale volume; however, the biggest challenge for battery size reduction is the power consumption of the system rather than the battery technology itself. A system's battery should be able to store enough energy to sustain the system operation for a reasonable lifetime without recharging. Therefore, significant power reduction is the key challenge for keeping battery volume on an mm-scale since the size of batteries used in today's small electronic systems is in cm-scale.

Comparing the amount of energy stored in a battery that fits in an mm-scale system with that of widely used alkaline AA-sized batteries gives us a good hint of how much power reduction is required for implementing mm-scale systems. The thin-film Li battery used in mm-scale systems in [3–6] has an energy density of 1 μAh/mm^2, whereas a typical AA-sized battery contains 1800–2600 mAh. Thus, the AA-sized battery has six orders of magnitude more energy than the thin-film Li battery. Therefore, with rough estimation, the average power consumption must be reduced to the nano-watt order to achieve a comparable battery lifetime. This can be achieved with two strategies: duty-cycling and active power reduction.

Each component in an mm-scale system operates within a different usage scenario as shown in the temperature sensor system example in Fig. 3.3. For monitoring temperatures in a storage facility, sensor measurement is required only once every 10–20 min. The sensor measurement consumes 1 μW of power

	Power	On Time	Frequency	Active Function
Always on	~ 1nW	∞	-	Time Keeper
Short-term	~ 1μW	~ 1s	Every 20 min	Temperature Sensing
Periodic on	~ 100μW	~ 30 ms	Every hour	Radio TX

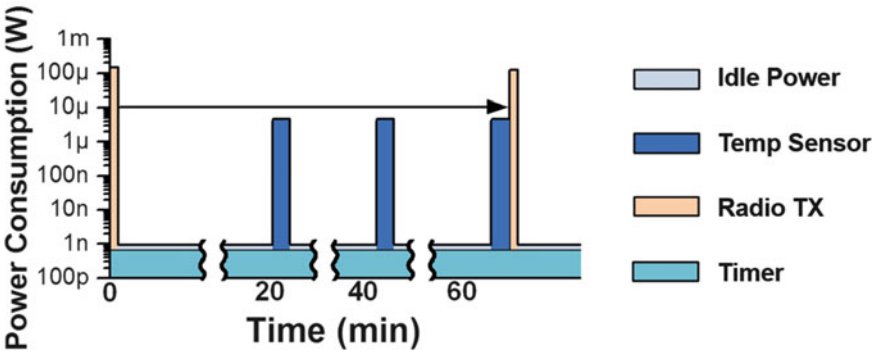


Fig. 3.3 Usage scenarios for components in a temperature monitoring system

and requires a 1 s measurement time. For reporting measured temperature data to a server, radio transmission is required, which consumes 100 s of μW power for 10 s of ms. Therefore, radio transmission should be performed less frequently, i.e., duty-cycled more aggressively, to reduce the total energy consumption to an acceptable range.

Active power reduction for each component is still the primary concern for realizing mm-scale systems. For example, if the radio transmission power can be lowered to 10 s of μW, the activation time can be extended by 10× for a given energy budget. This also means that the lifetime for a given battery can be extended by 10× assuming an identical activity rate. To meet the stringent energy/power budget of mm-scale sensors, various state-of-the-art low-power circuit design methodologies need to be considered when designing each component. However, there is no universal or commonly applicable low-power design strategy for all of the components in an mm-scale system, e.g., processor, memory, radio transmitter/receiver, sensors, and more. Therefore, each component must be carefully revisited, considering its characteristics and required performance. For the rest of this chapter, low-power circuit design techniques recently demonstrated for a variety of components in mm-scale systems will be presented.

3.3 Low-Power Circuit Design Technologies for mm-Scale Systems

An mm-scale system can include various components: sensors, sensor interfaces, a microprocessor, a memory, an energy harvester, a power management unit, a wireless receiver and transmitter, a wake-up timer, voltage/current reference circuits, and so on. To develop ultra-low power mm-scale systems for IoT applications, these building blocks should be implemented with ultra-low power consumption. Recent active research on low-power circuit designs enabled the development of mm-scale sensing systems that operate at unprecedented low power levels. In this sub-chapter, ultra-low power circuit design techniques for key building blocks will be discussed in detail.

3.3.1 Low-Power Sensing Modalities

For a variety of IoT applications, various sensing modalities are required. However, the power budget of the entire system is extremely limited due to the limited storage capacity of mm-scale batteries and the limited power available for harvesting with mm-scale harvesters (e.g., photovoltaic (PV) cells and thermoelectric generators). Therefore, the sensors cannot always be on; they must be duty-cycled. Even with duty-cycling, the active time would need to be extremely short if the active power is

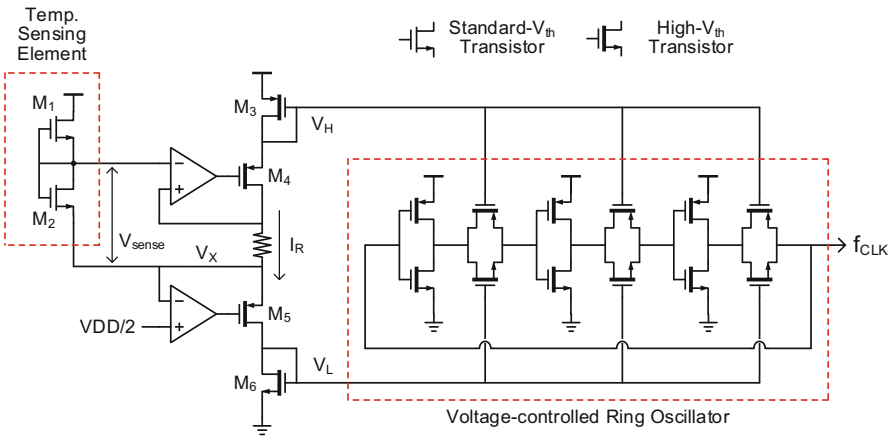


Fig. 3.4 Conceptual circuit diagram of the temperature sensor [4]

too high. Recently, various sensing modalities for miniature IoT system applications were reported with power consumption levels less than 2 μW that could still achieve reasonable accuracy and speed [4, 7, 8].

A temperature sensor is one of the most common sensing modalities for a wide range of applications. The CMOS temperature sensor reported in [4] is implemented to consume 71 nW while achieving an inaccuracy of +1.5/−1.4 °C with a conversion time of 30 ms. A conventional bipolar junction transistor (BJT)-based temperature sensor achieves an inaccuracy of ±0.15 °C but consumes 5.1 μW using a power-hungry analog-to-digital converter (ADC) for digitizing a temperature-dependent voltage [9]. Compared with other CMOS temperature sensors, the proposed design improves inaccuracy, energy per conversion, and power by 2×, 10×, and 2×, respectively, without requiring external clocks with high accuracy.

The CMOS temperature sensor achieves low power consumption by utilizing a low-power MOSFET-based sensing element and modified voltage-to-current converter and current mirror structures. Figure 3.4a shows a circuit diagram of the proposed temperature sensor. In the temperature-sensing unit, the gates of both transistors are connected to the output, and V_{sense} can be expressed as follows:

$$V_{sense} = \frac{m_1 m_2}{m_1 + \gamma_1' m_2} V_T \ln\left(\frac{\mu_1 C_{ox1} W_1 L_2}{\mu_2 C_{ox2} W_2 L_1}\right).$$

Since the same type of transistor is used, the output does not depend on the threshold voltage and mobility, resulting in low process variability. In simulation, it gives a temperature sensitivity of 382 μV/°C and an output voltage of 76.5 mV with 8 pW power. Compared with a conventional 2T structure with different types of transistors [10], the proposed sensing element has 2.2× less variation (0.8 % σ/μ) in temperature sensitivity and 2.8× lower output voltage.

V_{sense} is converted to current across a resistor (I_R). The reduced voltage helps decrease I_R to ~nA. I_R controls bias voltages (V_H and V_L) for the following voltage-controlled ring oscillator by diode-connected transistors (M_3 and M_6). To avoid an nA current mirror, V_H and V_L are created in one rail using two 136-pW amplifiers.

Fully integrated sensors with a standard CMOS technology are popular for their cost-effectiveness and simplicity for implementation. However, off-chip sensors are also widely used for many applications (e.g., chemical-, gas-, and pressure-sensing) for their superior performance and characteristics that cannot be obtained with integrated devices. Sensing through these off-chip sensors is often performed by sensing the change in capacitance or resistance. Thus, to interface with these off-chip sensors, a low-power capacitance-to-digital converter (CDC) or resistance-to-digital converter (RDC) is necessary for mm-scale sensor systems. For these converters, a wide range of input capacitance or resistance is desirable to cover a wide physical input range and to interface with many different types of sensing front-ends with a single converter.

The fully digital CDC in [7] achieves 1.84 μW power consumption for measuring capacitance ranging from 0.7 pF to 10 nF with <0.06 % linearity error. At 11.3 pF, it achieves 0.109 % resolution, 19 μs conversion time, and 35.1 pJ per conversion energy consumption. Compared with the previous CDCs, the energy consumption is reduced by 18×, and the figure-of-merit considering energy, input range, and resolution is improved by 1.3× without an amplifier.

The CDC utilizes an inverter delay chain powered by a precharged input capacitor (C_{SENSE}) as shown in Fig. 3.5. To discharge C_{SENSE} to a fixed voltage (V_{LOW}), the inverter chain must be toggled by a number of times proportional to the capacitance of C_{SENSE}. A counter records the toggling number as a digital output code. The input capacitance range is wide due to the direct charge drawn from C_{SENSE} without an initial capacitance-to-voltage conversion. Also, low conversion energy is achieved by reusing energy for charging C_{SENSE} when switching the inverter chain.

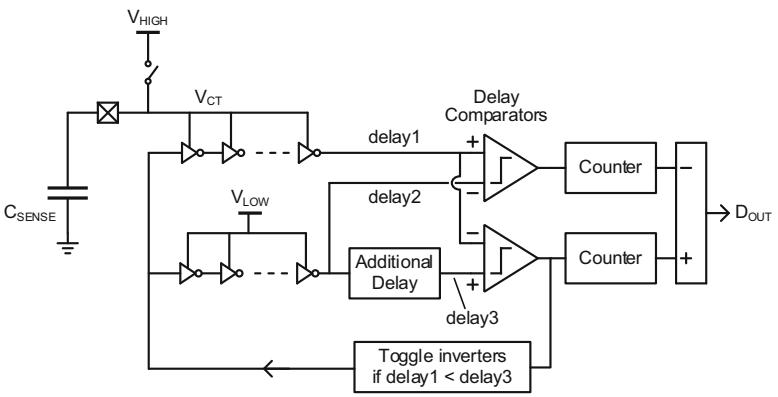


Fig. 3.5 Conceptual diagram of the fully digital capacitance-to-digital converter (CDC) in [7]

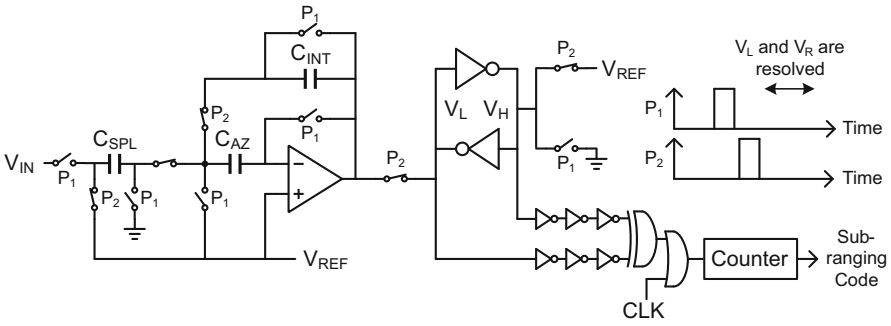


Fig. 3.6 Logarithmic voltage-to-time converter [8]

To increase signal-to-noise ratio (SNR), the sensing capacitor voltage (V_{CT}) is decreased below V_{LOW} by an additional delay. By multiple comparisons around V_{LOW}, false decisions of "$V_{CT} > V_{LOW}$" stochastically compensate for false decisions of "$V_{CT} < V_{LOW}$." Compared with the simple way to use a single comparison decision for a CDC operation, the overall conversion noise is square-rooted at the cost of 3 % energy overhead in simulation.

The 1.7 μW RDC proposed in [8] covers input resistance from 10 kΩ to 10 MΩ at a sampling rate of 1.2 kS/s with 0.21 % measurement error. Compared with the previous RDCs, this work achieves 32× lower energy per conversion and 6.6× less power consumption.

For a wide input resistance range, a logarithmic sub-range detector is used to find one correct sub-range among 14 candidates using a two-step course and find sub-range searching technique. According to the decision, the corresponding digital-to-current converter (I-DAC) is applied to the sensing resistance to generate voltage, which is digitized using switched-capacitor amplifiers and an ADC. The voltage across the resistor is stabilized faster with low resistance and a high I-DAC value. Thus, the active time of the I-DAC is adjusted based on the chosen sub-range to save power, whereas it is set for the slowest settling time in the previous RDC.

Figure 3.6 shows the logarithmic voltage-to-time converter based on comparator metastability resolution time. First, the level-shifted input voltage by V_{REF} (half supply voltage) is assigned to V_L, while V_R is set to a metastable point (V_{REF}).

Next, the switches connected to V_L and V_R are open so as to allow the back-to-back inverters to resolve V_L and V_R. The resolution time is exponentially related to the input voltage difference. Hence, a counter records the time to measure the input resistance value in log fashion, helping to detect the sub-range in one sampling cycle. In the conventional design, eight interactions are required in the worst case, resulting in higher energy consumption and a slower sampling rate.

3.3.2 Low-Power Microprocessor

A microprocessor in a sensor system manages the sensor operation sequence, processes measurement data, and controls data storage/transmission. Therefore, an energy-efficient microprocessor is necessary for an mm-scale system because

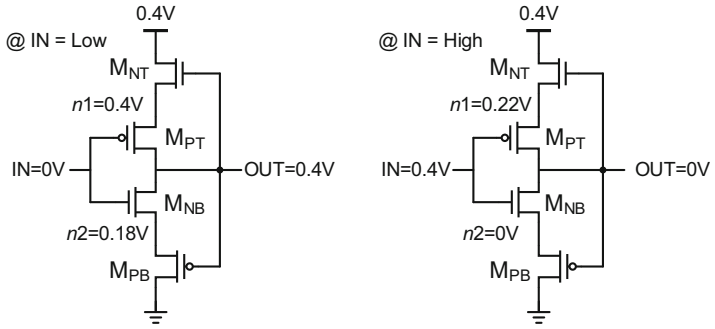


Fig. 3.7 Dynamic leakage-suppression logic (DLSL) inverter in steady state [12]

of the stringent energy budget. The Cortex-M® series from ARM, designed for mobile and sensor applications, is one of the best-known commercial 32-bit microprocessor families (http://www.arm.com/products/processors/cortex-m/). Cortex-M0+ (the lowest power processor in the Cortex-M family) in [11] achieves an energy efficiency of 11.7 pJ per instruction with subthreshold operation and leakage reduction techniques. The standby power is 80 nW with 4 kB retentive memory.

Another recent implementation of Cortex-M0+ is based on a new logic family called dynamic leakage-suppression logic (DLSL) [12]. DLSL minimizes power consumption rather than energy consumption so that it can operate with an energy harvester at a pW power level in a battery-less system. With the help of leakage current suppression with header/footer transistors, a DLSL-based processor reduces its power consumption to 295 pW by slowing the operating frequency to 2 Hz. Such low power can be supported by energy harvested in an ambient environment [13], allowing a battery-less sensing system.

Figure 3.7 shows an inverter design using the DLSL technique. In addition to the standard CMOS inverter (M_{PT} and M_{NB}), an NMOS transistor (M_{NT}) is inserted between the supply voltage and the source of M_{PT}, and a PMOS transistor (M_{PB}) is added between the ground and the source of M_{NB}. The output voltage is connected to the gates of the added transistors (M_{PB} and M_{NT}) such that they act as power gates. The power gates place either a pull-up or pull-down network into the super-cutoff region, reducing the leakage current of the logic to two orders of magnitude lower than that of the conventional CMOS logic. For example, when the input is low, and the output is high, the connecting node ($n2$) between the bottom transistors (M_{NB} and M_{PB}) becomes approximately half the supply voltage (0.18 V) since both M_{NB} and M_{PB} are turned off. Thus, V_{gs} of M_{NB} and V_{sg} of M_{PB} become negative, resulting in super-cutoff. As the supply voltage increases, the leakage current decreases due to a stronger super-cutoff effect. Above 0.55 V, however, p-n junction diode leakage and the drain-induced barrier lowering (DIBL) effect increase the total power consumption.

For switching operations, DLSL charges or discharges the output capacitance with subthreshold leakage current. For instance, when the input is switched from low to high, $n2$ and the output become the same value because M_{NB} is in weak

inversion mode. It moves M_{PB} from the super-cutoff to the normal cut-off region. The decreased output voltage lowers the voltage of $n1$ and places M_{NT} and M_{PT} in the super-cutoff region. Since M_{PB} in the regular cut-off region discharges the output at least an order of magnitude stronger than the pull-up network in super-cutoff, the output slowly switches to low.

The Cortex-M0+ processor implemented with DLSL consumes only sub-nW power at the limited operating speed of 2–15 Hz. It can be a good choice for a battery-less system that needs to sustain its operation under ambient energy conditions.

3.3.3 *Low-Power Energy Harvesters and Power Converter*

Recharging a battery is a critical function for an mm-scale system in order to sustain its operation, achieve energy autonomy, or maximize the system lifetime. In a high energy environment (e.g., bright outdoor light), the quality of mm-scale system operations will not be limited by energy provided by an energy harvester. On the other hand, in low energy conditions (e.g., dim indoor light), if the harvested energy is not sufficient to support the operations, the duty-cycling rate and the standby time of building blocks must be increased, degrading the quality and performance of the system operations. This condition becomes worse for an mm-scale sensor system since the available energy is limited due to the size constraints on energy sources (e.g., PV cells). To operate a sensor system in a low energy condition without performance degradation, developing an efficient energy harvester and power converter is necessary. With higher efficiency, an energy harvester extracts more energy from an energy source and saves it to a battery, and a power converter loses less energy when delivering energy stored in a battery to circuits. Two energy harvesters [13, 14] and one power converter [15] proposed recently enable efficient energy harvesting and power conversion at a low power level.

The fully integrated switched-capacitor (SC) energy harvester in [13] recharges a 4 V battery with $>35\,\%$ efficiency using a 0.84 mm^2 PV cell. At 260 lux, 7 nW input power is converted from 0.25 to 4 V. The harvester itself delivers output power ranging from 5 nW to 5 μW with $>40\,\%$ efficiency. Without an off-chip component (e.g., bulky inductor), previous fully integrated energy harvesters have not been demonstrated for $<\mu W$ input power.

A typical SC energy harvester includes a clock generator, level shifters, and a switched-capacitor network. In low input power conditions, power consumption for clock generation and level conversion to drive power switches becomes relatively significant and degrades the overall harvesting efficiency. In the proposed design, a self-oscillating voltage doubler that does not require a separate clock generator and level converter is developed, and four voltage doublers are cascaded to design a complete energy harvester.

Figure 3.8 shows the self-oscillating voltage doubler. It consists of two ring oscillators connected in series. The output nodes of the inverters are coupled by

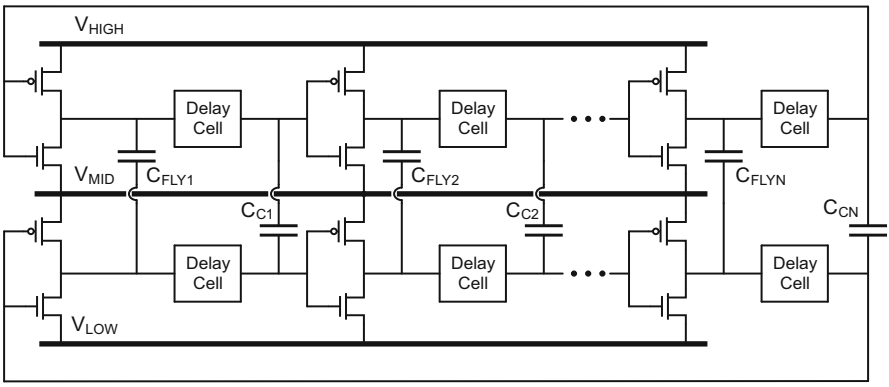


Fig. 3.8 Self-oscillating voltage doubler [13]

flying capacitors (C_{FLY1}−C_{FLYN}). When V_{MID} is connected to an energy source, and V_{LOW} is ground, the charge is transferred from V_{MID} to V_{HIGH} by the flying capacitors since the inverters of the ring oscillators are alternatively connecting the capacitors between V_{MID}-to-V_{LOW} and V_{HIGH}-to-V_{MID} in a similar manner as conventional 1:2 DC–DC converters. As the activity of charging or discharging the flying capacitors propagates through the inverter chains, a clock is internally created. Thus, a dedicated clock generator and level converter are not required. This helps reduce the significant power overhead from clock generation and level conversion.

Furthermore, the voltage doubler self-starts when V_{MID} is higher than 140 mV even if V_{HIGH} is 0 V. Once the bottom ring oscillator begins oscillation due to thermal noise, its inverters switch the bottom nodes of the flying capacitors and inject charge to the top ring oscillator. This enables the charge of a completely depleted battery without connecting an external power supply.

The light harvester in [14] achieves >78 % efficiency from 100 lux to 100 klux light using a 7.8 mm^2 PV-cell network. It can charge a 1.5 V battery even with 7 lux with 26 % efficiency. This harvester stacks the segmented PV cells to charge the battery which has higher voltage than a single PV cell. It only uses DC switches to connect the PV cells in series and avoids loss by switching capacitors or inductors. Hence, this work can achieve a high efficiency across a wide light intensity range, something that previous fully integrated harvesters have failed to do.

However, since light and battery condition change, the number of PV cells connected in series should be flexible so that the individual PV cells can be configured to operate at maximum power point. In this work, a PV cell network is designed with 36 unit cells, and it can be reconfigured to a "3 × 12", "4 × 9", "5 × 7", "6 × 6", or "7 × 5" array mode as shown in Fig. 3.9. The lost area due to the orphan cell is 2.8 % in the "5 × 7" and "7 × 5" modes. Also, the PV cells are sub-grouped to reduce the area loss from wire bonding pads and trench isolation between PV cells. If all 36 PV cells are connected to the harvesting chip, 72 wire

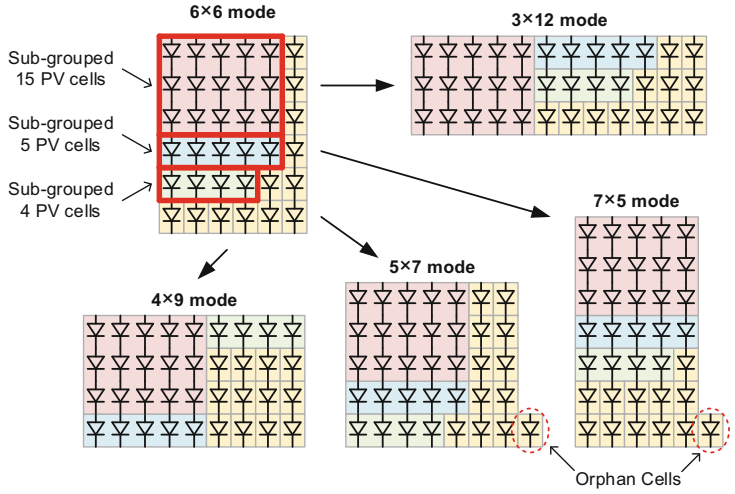


Fig. 3.9 Reconfigurable PV-cell network [14]

bonding pads are required. However, by sub-grouping 15, 5, and 4 PV cells, the loss of area is reduced from 14.1 % to 5.0 %.

In addition to energy harvesting efficiency, power conversion efficiency is a critical factor for an mm-scale system in a low energy environment. One or more power converters are required in an mm-scale system due to the voltage difference between battery voltage (e.g., 4 V) and the supply voltages for analog and digital circuits (e.g., 1–2 V). A linear regulator provides a clean output voltage with smaller ripple and is preferred for analog circuits which are sensitive to supply noise, but it becomes considerably inefficient with the high conversion ratio. An inductor-based switching converter obtains high efficiency even for high conversion ratios, but it requires a bulky off-chip inductor for low input power, which is not acceptable in an mm-scale sensing system. On the other hand, a fully integrated SC converter can be designed using on-chip capacitors for low-power delivery [16–18]. However, its coarse output voltage resolution results in poor efficiency due to significant conduction loss.

The successive-approximation (SAR) SC converter in [15] significantly reduces the resolution by employing a binary searching approach. It achieves 72 % peak efficiency with 31 mV resolution using one 4:1 and five 2:1 SC DC–DC converters. Compared with other conventional SC converters with the same conversion ratio resolution, such as series–parallel and ladder structures, the proposed converter gives a smaller slow-switching limit impedance (R_{SSL}) [19]. A smaller R_{SSL} indicates a higher driving capability of the output power.

Figure 3.10 shows an example of the proposed SAR SC converter. Four 2:1 SC converters are cascaded to control output voltage (V_{OUT}) with fine steps. A stage uses V_{HIGH} and V_{LOW} and generates the middle voltage (V_{MID}). V_{HIGH} and V_{LOW}

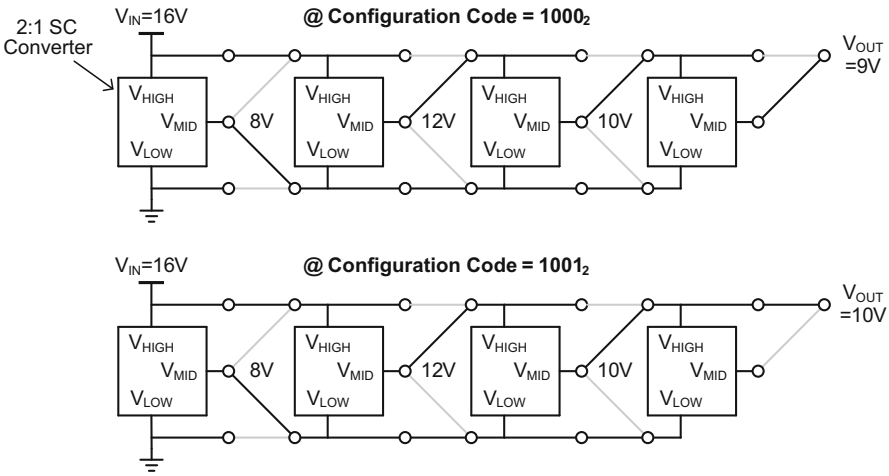


Fig. 3.10 4 b SAR SC converter [15]

of the following stage are connected to "V_{HIGH} and V_{MID}" or "V_{MID} and V_{LOW}", respectively, in this stage, and the next stage generates new V_{MID} and creates smaller conversion ratio resolution. For example, when the input voltage is 16 V, "configuration code $= 1000_2$" sets $V_{OUT} = 9$ V, and "configuration code $= 1001_2$" sets $V_{OUT} = 10$ V. Under no-load conditions, a conversion ratio resolution of 1 V is provided.

3.3.4 Low-Power Wireless Receiver and Transmitter

As discussed earlier, duty-cycling is one of the most effective ways of reducing the average power consumption of an mm-scale system. In standby mode, most of the circuits are turned off, and only essential blocks remain turned on. With the help of the always-active components, the system can be interrupted, synchronized, and reprogramed at any time. A wake-up receiver is one of the always-active circuits for which low-power design is critical due to the low standby power budget of an mm-scale system.

Typical RF wake-up receivers require 10 s of μW (75 μW [20], 52 μW [21]) due to high carrier frequency and throughput. To achieve lower power levels, active research on wake-up receivers has been continued. A dual-mode wake-up receiver consumes 8.5 μW by using two different operation modes [22]. In standby mode, the receiver stays in a low-data rate to save power. Once a wake-up signal is detected, it changes to a high-data rate, dissipating higher power. Also, a wake-up receiver using ultrasound reduces the power consumption to 4.4 μW for a 250 bps data rate at 8 m distance. This is achieved by using a low carrier frequency at the

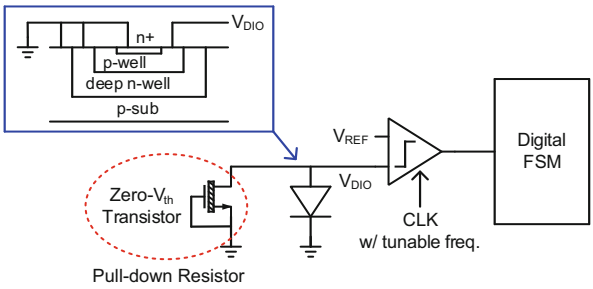


Fig. 3.11 Front-end of the optical wake-up receiver [24]

cost of a slow data rate [23]. However, the power consumption still is not acceptable in an mm-scale sensor system with nW average standby power.

A new approach using visible light as an information carrier achieves 695 pW standby power [24]. Figure 3.11 shows the front-end of the optical wake-up receiver. When light is shined, the voltage across an n+/p-well/n-well diode increases, acting as a PV cell. With lower light intensity, the voltage decreases since the pull-down current by a resistor is stronger than the current generated by the PV cell. The voltage can be lowered in response to weak light without the resistor, but it helps decreasing the discharging time, which increases the data rate. The voltage generation is powered by the forward-biased diode itself, which requires no power. The diode voltage is digitized with a reference voltage and a comparator. Similar to [22], the conversion speed is decreased in standby mode, reducing the power to 695 pW. Once a user-defined pattern is recognized, the conversion speed becomes faster and achieves 91 bps and 140 pJ/b efficiency with 12.7 nW power. With laser light, the maximum distance of 50 m is obtained.

Once a wake-up signal is detected, the system is switched to the active mode. In active mode, signals are received as well as transmitted, and energy efficiency for both directions is important. However, improving the energy efficiency of a wireless transmitter and receiver is challenging. For a transmitter, the communication distance and carrier frequency determine path loss. The transmit power of an mm-scale system is determined by the path loss and the minimum sensitivity level of the base station (e.g., −100 dBm). For a receiver, the base-station transmit power is restricted by the Federal Communications Commission (FCC) limit. Path loss sets the minimum sensitivity of a sensor system receiver, thus determining the gain and noise figure of a low-noise amplifier (LNA).

Despite the challenges, a 10-mm^3 near-field radio system has been recently demonstrated for syringe implantation to minimize the invasiveness of implantation [25]. Integrated with a 1×8 mm^2 magnetic antenna fabricated with 1.5-μm thickness gold on 100-μm thickness glass substrate, the overall system can be injected through a 14-gauge syringe needle. It achieves a link distance up to 50 cm (sensor TX) and 20 cm (sensor RX). The transmitter and receiver consume 43.5 μW at 2 kb/s and 36 μW at 100 kb/s, respectively. Compared with other mm-scale near-field radios (e.g., 3.5 cm [26]), a longer link distance is obtained with improved efficiency using the techniques described below.

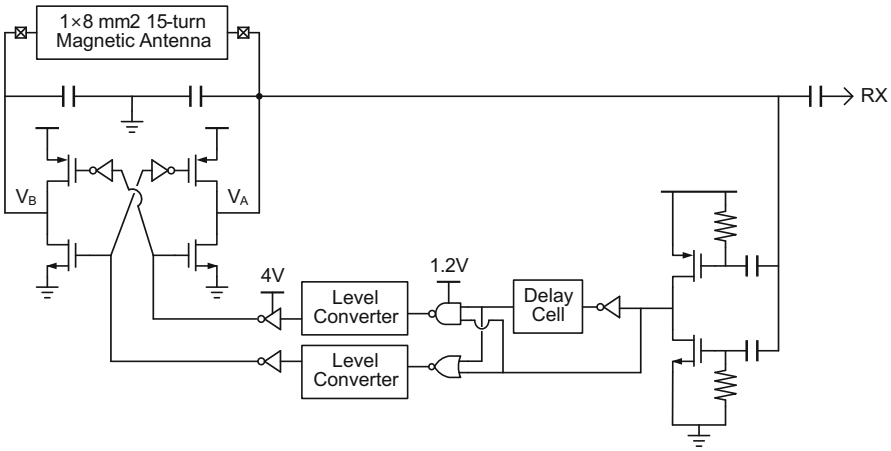


Fig. 3.12 Pulse-inject H-bridge LC-oscillator [25]

In this work, the asymmetric magnetic antenna pair on the implanted sensor system and the base-station device are co-optimized. The antenna size of the implantable sensor system (1×8 mm^2) is more limited than the antenna on the base station (11×11 cm^2). For 1-mm coil width, the coupling strength improves with the longer antenna length, but the lower self-resonant frequency (SRF) and Q-factor result in saturated strength above 8 mm. In addition, the sensor system and base station use asymmetric transmit-signal power levels and carrier frequencies. The sensor system transmits in a signal level under the FCC limit at 3 m at 112 MHz. The low transmit power from the sensor system can be tolerated by the highly sensitive, low-noise figure (~5 dB) receiver in the base station. On the other hand, the base station sends signals with 30 dBm output power at a less-optimal frequency (49.86 MHz) to meet FCC regulations.

The pulse-inject H-bridge LC-oscillator shown in Fig. 3.12 is 1.65× more efficient than a conventional constant-bias cross-coupled LC-oscillator. In the proposed circuit, only when V_A is at the peak, pulsed-current is injected into an H-bridge to sustain the oscillation of the resonant tank formed by the magnetic antenna and on-chip capacitors. To avoid short-circuit current, resistors are used for the inverter that monitors V_A.

Finally, a new sensor-initiating synchronization protocol is used. It does not require an area-demanding off-chip crystal and synchronization baseband processing on the sensor system. The sensor system begins communication by sending multiple pulses at a predefined pseudo-random interval. Based on the received initiation packet, the base station adjusts its local timer frequency and sends a response packet. The response signal is transmitted after a predefined guard time delay set by the timer in the sensor system. Thus, the sensor system returns to sleep mode to save power after sending the initial packet until it receives the response packet from the base station.

3.3.5 *Low-Power Timers*

In a wireless sensor system, an accurate timer should be included to determine when to sense physical quantities of interest and perform wireless data communication. In particular, long-term jitter [27] and temperature sensitivity are important metrics for this timer. For radio communication, a timing error made by a transmitter increases the power consumption of the radio receiver in a receiving-side sensor system by causing a longer waiting time. During this period, the receiver needs to be turned on, and this significantly increases power consumption because a typical radio receiver uses 100 s of μW power consumption.

A quartz crystal provides accurate timing, and its power consumption has recently been reduced to 1 s of nW [28, 29]. However, a quartz crystal cannot be utilized in an mm-scale sensor system due to its bulkiness. A MEMS-based oscillator can be used for volume-limited applications with comparable accuracy, but its 100 s of nW power consumption is not acceptable in a miniature sensing system with an nW standby power budget.

If a sensor system communicates with a base station continuously powered by a wired connection, a power increase due to a timing error is better tolerated, and the timing accuracy requirement is relieved. Thus, low-power timers with acceptable accuracy for mm-scale sensing systems have recently been developed for such conditions [30, 31].

The 5.8 nW CMOS wake-up timer in [30] uses a constant charge subtraction scheme and achieves a temperature stability of 45 ppm/°C and an Allan deviation (long-term stability) of 50 ppm with <11 Hz output frequency. In a conventional relaxation oscillator, the clock period is determined by the time required to charge a capacitor to the threshold voltage. Since a comparator takes time to detect if the capacitor voltage equals the threshold, the comparator delay is included in the output clock period. The comparator delay is highly sensitive to temperature and thus makes the output clock unstable over temperature. For less temperature sensitivity, higher power is required for the comparator to reduce the impact of the delay. For lower power, the proposed work eliminates the comparator delay in the clock delay calculation.

Figure 3.13 shows a conceptual diagram of the proposed constant charge subtraction scheme. Compared with the conventional oscillator that fully discharges the integration capacitor (C_{INT}), it subtracts the constant charge from C_{INT} when the capacitor voltage passes a threshold voltage (V_{SUB}). This is done by a low-power slow comparator, and the comparator delay can be significant. However, the constant charge subtraction technique helps the sawtooth waveform of C_{INT} rejoin the sawtooth waveform using an ideal comparator without time delay. The delay is not added to the output clock delay by using a high-power fast comparator. This fast comparator detects when V_{INT} rises above a different threshold voltage (V_{COMP}), and it toggles the output clock. For a long wake-up period (i.e., 1 min), the output

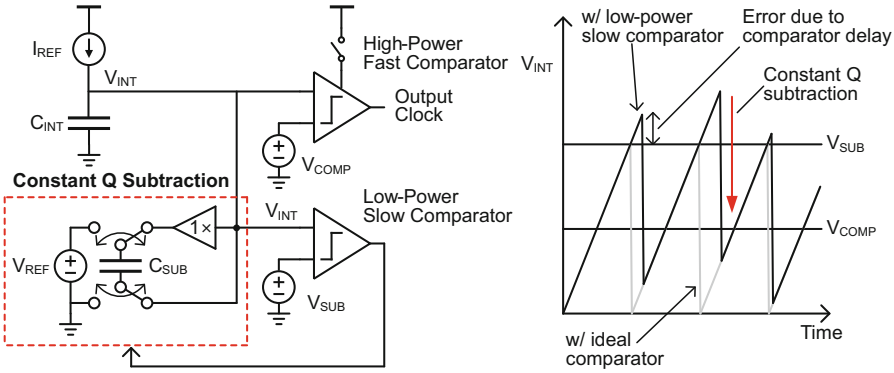


Fig. 3.13 Conceptual diagram and waveform of the constant charge subtraction scheme [30]

clock does not need to be flipped, although the internal period charging or discharging C_{INT} is faster (11 Hz). By activating the fast comparator only for one charging period among multiple charging/discharging periods, the average power is reduced by 7.8× from 45 nW to 5.8 nW. This is enabled by the low-power continuous charging and discharging operation, which is separated from the output clock generation using the high-power fast comparator.

Another low-power timer in [31] reduces power consumption with a frequency-locking technique. Compared with the timer using the constant charge subtraction scheme, it continuously generates a 3-kHz output clock by replacing comparators with a low-power amplifier. The fast output clock can be used for general purposes in the system as well as the wake-up clock. The timer achieves 13.8 ppm/°C at 3 kHz with 4.7 nW power consumption. To save additional power, a duty-cycled resistor and current-reuse schemes are used.

Figure 3.14 shows the timer proposed in [31]. A switched capacitor converts the output frequency of a voltage-controlled oscillator (VCO) to resistance. The resistance changes the pull-down current (I_{SC}) on the control voltage of VCO (V_{CTRL}). The pull-up current (I_{SR}) is generated using a temperature-compensated switched resistor. When I_{SC} equals I_{SR}, the output frequency is stabilized at $1/(MR_{SW}C_{SW})$. Using the duty cycle of M, the effective resistance of 17 MΩ (0.065 mm^2) is increased by M× without additional area. Both terminals of the resistor are connected/disconnected during the switching operation to avoid current flow to parasitic capacitance, thus maintaining the equivalent resistance and reducing temperature sensitivity. In addition, I_{SC} and I_{RC} are placed in series, and current is reused. This approach helps reduce power by 2× compared with the conventional timers that include reference and sensing parts in parallel.

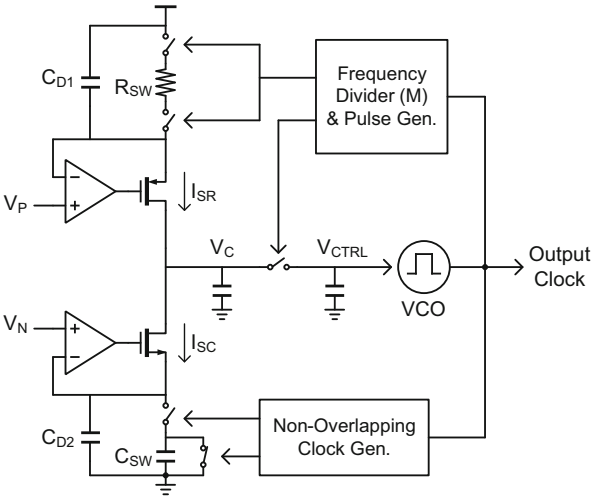


Fig. 3.14 Circuit diagram of the wake-up timer using a switched-resistor scheme [31]

3.3.6 Low-Power Voltage/Current References

Analog circuits require stable voltage and current references over process, voltage, and temperature (PVT) variations to maintain their performance. To sustain the sensor system operation in standby mode, the reference circuits need to support always turned-on blocks such as a power management unit, wake-up receiver, and timer. However, designing the voltage/current references for an mm-scale system is a non-trivial challenge since the power consumption should be below the extremely limited standby-mode power budget (8 nW in [32]) to benefit from low average power by duty-cycling. Here, we discuss two voltage references [10, 33] and one current reference [34], which can be used for miniature sensing systems.

The 2.98 nW bandgap voltage reference in [33] achieves a temperature sensitivity of 25 ppm/°C from −20 to 100 °C. The voltage reference is based on a traditional design with single point trimming. The low-power consumption is achieved by duty-cycled operation (0.003 % at 27 °C), and it is 251× lower than that of the previous bandgap references.

Figure 3.15 shows a circuit diagram of the proposed voltage reference. In active mode (CLK0 = 0), the core bandgap circuit provides the output voltage (VREF) while charging the output capacitor (C_5). In sleep mode (CLK0 = 1), the core circuit is tuned off, but VREF is continuously supported by the output capacitor. Since the difference in power consumption between the active and sleep modes is higher than 10^3×, a low duty cycle ratio is the key to obtaining low average power consumption.

In this voltage reference, a low duty cycle ratio is enabled by three techniques. First, the internal nodes of the core circuits are sampled on C_1–C_4 to speed up the stabilization time of the core circuit. By restoring the stored values when the core circuit is turned on, the active time is reduced by 11.5× (from 55 ms to 4.8 ms in

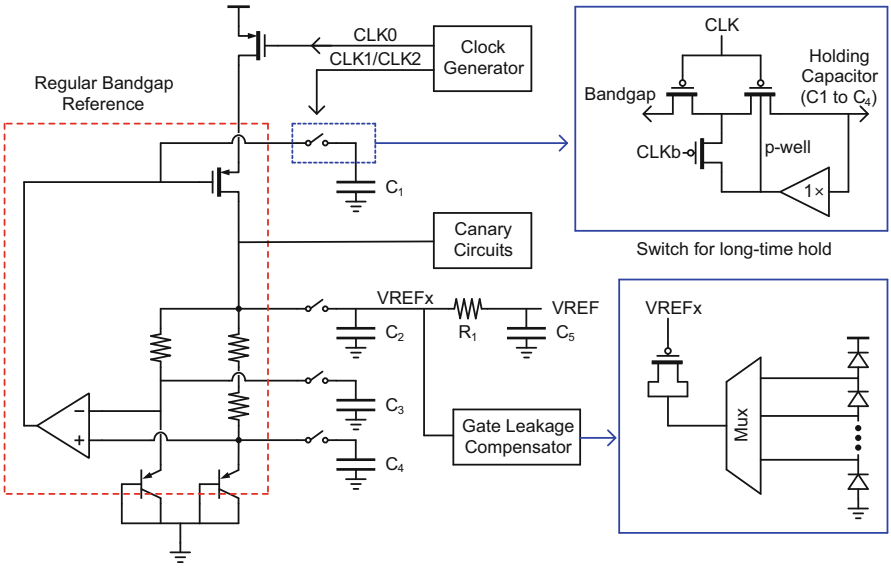


Fig. 3.15 Circuit diagram of the bandgap voltage reference using a sample and hold circuit [33]

simulation). Second, the subthreshold and junction leakage is reduced in the sample and hold circuits for C_1–C_5. The analog unity-gain buffers force V_{sb}, V_{db}, and V_{ds} of the switches directly connected to the capacitors to zero, helping to increase the sleep time by $10^3\times$. Lastly, to compensate for the gate leakage current through the buffer connected to the output capacitor, the leakage compensator injects a small current using a MOS capacitor. The current is controlled by voltage across the capacitor.

To automatically adjust the sleep time for PVT variation, the voltage reference uses a canary sample and hold circuit with 600× lower output capacitance than the original circuit. In sleep mode, it monitors the error between the original and canary circuit and initiates the active mode when the error is larger than a predetermined threshold. Also, this controller is used to set the gate compensation leakage. By optimizing the refresh time and leakage compensation, the power consumption is reduced by 2.75× compared with that of the circuit without the automatic sleep time control.

The nW bandgap voltage reference can be used for wireless sensor systems that have a sleep-mode power budget higher than 10 s of nW. However, it can dominate sleep-mode power in extreme cases with a standby power of less than 10 s of nW [5, 35]. For these systems, the 29.5 pW MOSFET-based voltage reference in [10] can be a good choice, although this type of voltage references typically shows higher process variation than BJT-based bandgap references. Figure 3.16 shows the MOSFET-voltage reference with digital trimming, which uses neither a resistor nor a BJT. It achieves a temperature coefficient of less than 50 ppm/°C, an output voltage spread of 0.72 % σ/μ, a line sensitivity of 0.036 %/V, a power supply

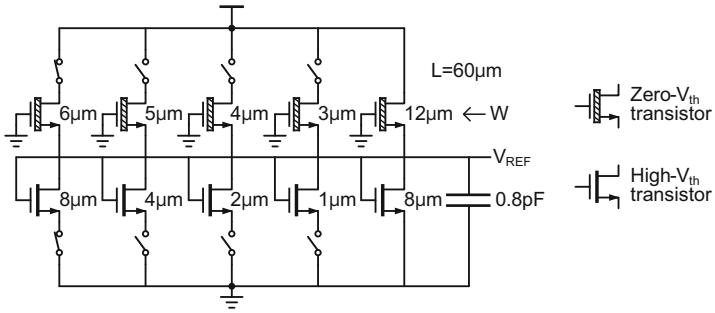


Fig. 3.16 Schematic of the trimmable MOSFET-based voltage reference [10]

rejection ratio of −51 dB, and a minimum supply voltage of 0.5 V. The power consumption is 88× lower than that of the previous low-power voltage references (1181× lower in the non-trimming version).

The voltage reference uses transistors with thick gate oxides to support a high supply voltage, but they have different threshold voltages. The output voltage of the reference is related to the difference of the threshold voltages as follows:

$$V_{REF} = \frac{m_1 m_2}{m_1 + m_2}(V_{th2} - V_{th1}) + \frac{m_1 m_2}{m_1 + m_2} V_T \ln\left(\frac{\mu_1 C_{ox1} W_1 L_2}{\mu_2 C_{ox2} W_2 L_1}\right)$$

where μ is the mobility, C_{ox} is the oxide capacitance, W is the transistor width, L is the transistor length, m is the subthreshold slope factor, V_T is the thermal voltage, and V_{th} is the transistor threshold voltage. The approximate output voltage (V_{REF}) can be expressed as $0.75 \times (V_{th2} - V_{th1})$ if the second log term is negligible and a typical subthreshold swing (90 mV/dec) is applied for both of the transistors. In addition, the output voltage should be larger than ~$5V_T$ to avoid the dependency of V_{ds} on the subthreshold current since the output voltage equals V_{ds} of M_2. Thus, the minimum difference of the threshold voltages needs to be $6.6V_T$ (170 mV at room temperature).

To minimize the spread of the output voltage and temperature coefficient, the following trimming process is performed. First, the output voltages are measured from several dies at two temperature points (20 and 80 °C) with different trimming settings. Using the results, the target output voltage is set to minimize the spread of the output voltages and temperature coefficient. Next, the remaining dies are placed at 80 °C, and the voltage references are trimmed to obtain the output voltage closest to the target value.

To place transistors in a desirable operating region, current biasing is preferred over voltage biasing since important parameters such as transconductance (g_m) and output resistance (r_o) can easily be controlled by current even though the threshold voltage varies due to PVT variations. The 23 pW current reference in [34] can be used to bias analog building blocks in sleep mode with a limited power budget. It obtains a temperature coefficient of 780 ppm/°C, a line sensitivity of 0.58 %/V, and

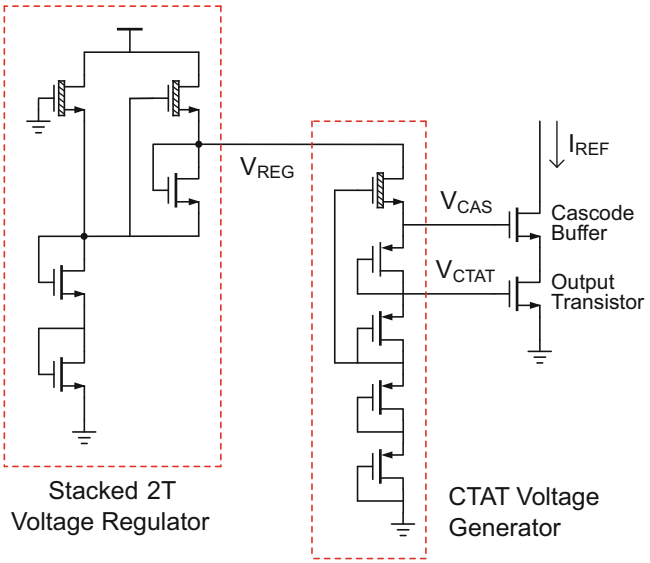


Fig. 3.17 Circuit diagram of the current reference using subthreshold MOSFETs [34]

a load sensitivity of 0.5 %/V. The power consumption is reduced by 50× compared with that of the previous current references.

As shown in Fig. 3.17, the current reference adjusts V_{gs} of the output transistor according to the temperature change to achieve a constant reference current (I_{REF}). To provide a pA reference current, the output transistor is operated in the subthreshold region. For $V_{ds} > 0.1$ V, the relationship between the subthreshold reference current and temperature can be expressed as

$$I_{REF} = a_1 \sqrt{T} e^{\frac{a_2 (V_{gs} - V_{th})}{T}}$$

where a_1 and a_2 are constant, T is the temperature, and V_{th} is the transistor threshold voltage. Note that the temperature dependence on carrier mobility is also considered. In this current reference, the temperature dependency is minimized in two ways. First, it matches a temperature coefficient of V_{gs} to that of V_{th} so they cancel out. Second, the proposed circuit sets V_{gs} to the point where the temperature dependency of $\sqrt{T}$ and $e^{\frac{a_2 (V_{gs} - V_{th})}{T}}$ cancels each other out.

The complementary-to-absolute temperature (CTAT) generator provides V_{gs} to the output transistor by stacking two different types of PMOS transistors. Additional bottom PMOS transistors increase the temperature coefficient from −0.2 mV/°C to −1.26 mV/°C. In the CTAT generator, the top zero-V_{th} NMOS transistor is used to reduce supply sensitivity from 4k %/V to 4 %/V. In addition, by running the CTAT generator under the low-power voltage reference, the supply sensitivity is further improved by 36×.

3.4 Conclusions

The continuous scaling of computing systems predicted the emergence of millimeter-scale systems, which is a promising form-factor for future IoT applications with aggressive volume constraints, such as a smart dental brace. These miniature systems can be realized by significant battery and electronics volume reduction, which can be made possible only by reconsidering every circuit element for ultra-low power operation. Thanks to continuous research on low-power circuit design methodology, many of the circuit components in miniature sensor systems are now re-designed to operate with stringent power and energy budgets. Looking ahead, continued progress in low-power sensors, microprocessors, energy harvesters, power converters, and wireless communication with low power timers represents particularly exciting areas of research in this field. These mm-scale systems will create numerous new extremely volume-constrained IoT applications similar to a smart dental brace and will also replace many of today's bulky and power-hungry electronic systems in other applications. These miniaturized systems will open the door to truly ubiquitous sensing and computing.

Acknowledgments This work is supported by Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project (CISS-2012M3A6A6054193).

References

1. Bell G, Chen R, Rege S (1972) The effect of technology on near term computer structures. Computer 5(2):29–38. doi:10.1109/C-M.1972.216890
2. Bell G (2008) Bell's law for the birth and death of computer classes. Commun ACM 51 (1):86–94. doi:10.1145/1327452.1327453
3. Oh S, Lee Y, Wang J, Foo Z, Kim Y, Jung Y, Blaauw D, Sylvester D (2015) A dual-slope capacitance-to-digital converter integrated in an implantable pressure-sensing system. IEEE J Solid State Circuits 50(7):1581–1591. doi:10.1109/JSSC.2015.2435736
4. Jeong S, Foo Z, Lee Y, Sim JY, Blaauw D, Sylvester D (2014) A fully-integrated 71 nW CMOS temperature sensor for low power wireless sensor nodes. IEEE J Solid State Circuits 49 (8):1581–1591. doi:10.1109/JSSC.2014.2325574
5. Kim G, Lee Y, Foo Z, Pannuto P, Kuo YS, Kempke B, Ghaed M, Bang S, Lee I, Kim Y, Jeong S, Dutta P, Sylvester D, Blaauw D (2014) A millimeter-scale wireless imaging system with continuous motion detection and energy harvesting. In: Symposium on VLSI circuits. doi:10.1109/VLSIC.2014.6858425
6. Chen YP, Jeon D, Lee Y, Kim Y, Foo Z, Lee I, Langhals NB, Kruger G, Oral H, Berenfeld O, Zhang Z, Blaauw D, Sylvester D (2015) An injectable 64 nW ECG mixed-signal SoC in 65 nm for arrhythmia monitoring. IEEE J Solid State Circuits 50(1):375–390. doi:10.1109/JSSC.2014.2364036
7. Jung W, Jeong S, Oh S, Bang S, Sylvester D, Blaauw D (2015) A 0.7pF-to-10nF fully digital capacitance-to-digital converter using iterative delay-chain discharge. In: 2015 I.E. international solid-state circuits conference—(ISSCC) digest of technical papers. doi:10.1109/ISSCC.2015.7063137

8. Choi M, Gu J, Blaauw D, Sylvester D (2015) Wide input range 1.7 μW 1.2 kS/s resistive sensor interface circuit with 1 cycle/sample logarithmic sub-ranging. In: Symposium on VLSI circuits. doi:10.1109/VLSIC.2015.7231311
9. Souri K, Chae Y, Makinwa KAA (2013) A CMOS temperature sensor with a voltage-calibrated inaccuracy of ±0.15°C (3σ) from −55°C to 125°C. IEEE J Solid State Circuits 48(1):292–301. doi:10.1109/JSSC.2012.2214831
10. Seok M, Kim G, Blaauw D, Sylvester D (2012) A portable 2-transistor picowatt temperature-compensated voltage reference operating at 0.5 V. IEEE J Solid State Circuits 47 (10):2534–2545. doi:10.1109/JSSC.2012.2206683
11. Myers J, Savanth A, Gaddh R, Howard D, Prabhat P, Flynn D (2016) A subthreshold ARM cortex-M0+ subsystem in 65 nm CMOS for WSN applications with 14 power domains, 10T SRAM, and integrated voltage regulator. IEEE J Solid State Circuits 51(1):31–44. doi:10.1109/JSSC.2015.2477046
12. Lim W, Lee I, Sylvester D, Blaauw D (2015) Batteryless sub-nW cortex-M0+ processor with dynamic leakage-suppression logic. In: IEEE international solid-state circuits conference. doi:10.1109/ISSCC.2015.706296
13. Jung W, Oh S, Bang S, Lee Y, Foo Z, Kim G, Zhang Y, Sylvester D, Blaauw D (2014) An ultra-low power fully integrated energy harvester based on self-oscillating switched-capacitor voltage doubler. IEEE J Solid State Circuits 49(12):2800–2811. doi:10.1109/JSSC.2014.2346788
14. Lee I, Lim W, Teran A, Phillips J, Sylvester D, Blaauw D (2016) A >78%-efficient light harvester over 100-to-100 klux with reconfigurable PV-cell network and MPPT circuit. In: IEEE international solid-state circuits conference, San Francisco, CA. 370–371. doi:10.1109/ISSCC.2016.7418061
15. Bang S, Wang A, Giridhar B, Blaauw D, Sylvester D (2013) A fully integrated successive-approximation switched-capacitor DC-DC converter with 31 mV output voltage resolution. In: IEEE international solid-state circuits conference. doi:10.1109/ISSCC.2013.6487774
16. Ng V, Sanders S (2012) A 92%-efficiency wide-input-voltage-range switched-capacitor DC-DC converter. In: IEEE international solid-state circuits conference. doi:10.1109/ISSCC.2012.6177016
17. Ramadass YK, Fayed AA, Chandrakasan AP (2010) A fully-integrated switched-capacitor step-down DC-DC converter with digital capacitance modulation in 45 nm CMOS. IEEE J Solid State Circuits 45(12):2557–2565. doi:10.1109/JSSC.2010.2076550
18. Ramadass YK, Chandrakasan AP (2007) Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications. In: IEEE power electronics specialists conference. doi:10.1109/PESC.2007.4342378
19. Seeman MD, Sanders SR (2008) Analysis and optimization of switched-capacitor DC–DC converters. IEEE Trans Power Electron 23(2):841–851. doi:10.1109/TPEL.2007.915182
20. Molnar A, Lu B, Lanzisera S, Cook BW, Pister KSJ (2004) An ultra-low power 900 MHz RF transceiver for wireless sensor networks. In: IEEE custom integrated circuits conference. doi:10.1109/CICC.2004.1358833
21. Cook BW, Berny A, Molnar A, Lanzisera S, Pister KSJ (2006) Low-power 2.4-GHz transceiver with passive RX front-end and 400-mV supply. IEEE J Solid State Circuits 41 (12):2757–2766. doi:10.1109/JSSC.2006.884801
22. Yoon DY, Jeong CJ, Cartwright J, Kang HY, Han SK, Kim NS, Ha DS, Lee SG (2012) A new approach to low-power and low-latency wake-up receiver system for wireless sensor nodes. IEEE J Solid State Circuits 47(10):205–2419. doi:10.1109/JSSC.2012.2209778
23. Yadav K, Kymissis I, Kinget PR (2013) A 4.4-μW wake-up receiver using ultrasound data. IEEE J Solid State Circuits 48(3):649–660. doi:10.1109/JSSC.2012.2235671
24. Kim G, Lee Y, Bang S, Lee I, Kim Y, Sylvester D, Blaauw D (2012) A 695 pW standby power optical wake-up receiver for wireless sensor nodes. In: IEEE custom integrated circuits conference. doi:10.1109/CICC.2012.6330603

25. Shi Y, Choi M, Li Z, Kim G, Foo ZY, Kim HS, Wentzloff D, Blaauw D (2016) A 10 mm^3 syringe-implantable near-field radio system on glass substrate. In: IEEE international solid-state circuits conference, Feb 2016
26. Yakovlev A, Jang J, Pivonka D, Poon A (2013) A 11 μW sub-pJ/bit reconfigurable transceiver for mm-sized wireless implants. In: IEEE custom integrated circuits conference. doi:10.1109/CICC.2013.6658501
27. Allan DW (1966) Statistics of atomic frequency standards. Proc IEEE 54(2):221–230. doi:10.1109/PROC.1966.4634
28. Yoon D, Sylvester D, Blaauw D (2012) A 5.58 nW 32.768 kHz DLL-assisted XO for real-time clocks in wireless sensing applications. In: IEEE international solid-state circuits conference. doi:10.1109/ISSCC.2012.6177043
29. Hsiao KJ (2014) A 1.89 nW/0.15 V self-charged XO for real-time clock generation. In: IEEE international solid-state circuits conference. doi:10.1109/ISSCC.2014.6757442
30. Jeong S, Lee I, Blaauw D, Sylvester D (2015) A 5.8 nW CMOS wake-up timer for ultra-low-power wireless applications. IEEE J Solid State Circuits 50(8):1754–1763. doi:10.1109/JSSC.2015.2413133
31. Jang T, Choi M, Jeong S, Bang S, Sylvester D, Blaauw D (2016) A 4.7 nW 13.8 ppm/°C self-biased wakeup timer using a switched-resistor scheme. In: IEEE international solid-state circuits conference, San Francisco, CA, 31 Jan 2016–4 Feb 2016
32. Kuo YS, Pannuto P, Kim G, Foo ZY, Lee I, Kempke B, Dutta P, Blaauw D, Lee Y (2014) MBus: a 17.5 pJ/bit/chip portable interconnect bus for millimeter-scale sensor systems with 8 nW standby power. In: IEEE custom integrated circuits conference. doi:10.1109/CICC.2014.6946046
33. Chen YP, Fojtik M, Blaauw D, Sylvester D (2012) A 2.98 nW bandgap voltage reference using a self-tuning low leakage sample and hold. In: Symposium on VLSI circuits. doi:10.1109/VLSIC.2012.6243859
34. Choi M, Lee I, Jang TK, Blaauw D, Sylvester D (2014) A 23 pW, 780 ppm/°C resistor-less current reference using subthreshold MOSFETs. In: European solid state circuits conference. doi:10.1109/ESSCIRC.2014.6942036
35. Lee Y, Bang S, Lee I, Kim Y, Kim G, Ghead MH, Pannuto P, Dutta P, Sylvester D, Blaauw D (2013) A modular 1 mm^3 die-stacked sensing platform with low power I^2C inter-die communication and multi-modal energy harvesting. IEEE J Solid State Circuits 48(1):229–243. doi:10.1109/JSSC.2012.2221233

Chapter 4
Smart Sensor Microsystems: Application-Dependent Design and Integration Approaches

Minkyu Je

Abstract With the future filled with a trillion sensors on the way, there is a large variety in the forms of smart sensors for different applications existing or emerging, such as environmental monitoring, smart grid, green transportation, smart home and building, wearables, implants, and so on. The applications of sensors and corresponding use scenarios define desired form factors, operation frequencies and durations, energy sourcing and management strategies, communication distances and data rates, as well as control interfaces and protocols, leading to significantly different microsystem structures as well as design and integration approaches eventually. In this chapter, the application dependence of the microsystem structures and design/integration approaches are investigated, along with several examples of the smart sensor microsystem implementation across different applications introduced. While we find that the optimally crafted system designs and integration strategies can draw the maximum out of currently available technologies on one hand, the study on the other hand reveals the limitations, challenges, and bottlenecks of the technologies to overcome for a leap to the next stage of the sensor world.

Keywords Smart sensors • Microsystems • Wireless communication • Internet of things • Medical devices • Implantable blood flow sensor • Neural recording • Body-channel communication • Wireless capsule endoscopy • Integrated circuits

4.1 Introduction

Internet of things (IoT), or internet of everything (IoE), has numerous applications already existing or emerging. Such applications include environmental monitoring, smart grid, green transportation, smart home and building, wearables, implants, and

M. Je (✉)
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
e-mail: mkje@kaist.ac.kr

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_4

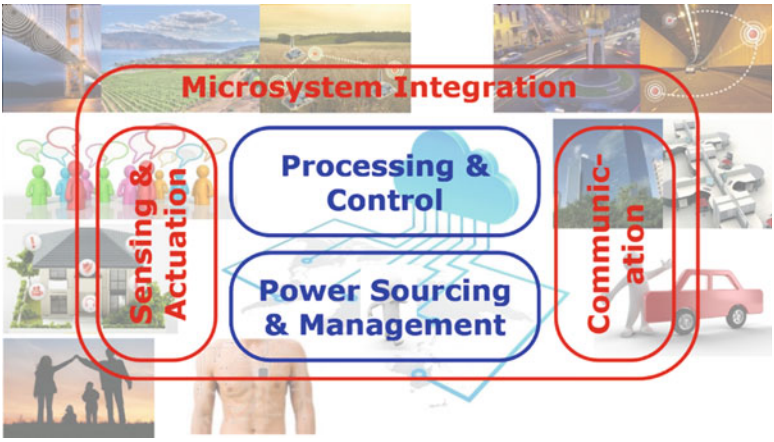


Fig. 4.1 Internet of things (IoT)/IoE (internet of everything) and smart sensor microsystems

so on. From the name, "internet of things," the "internet" means connected cyber world, while the "things" stand for the objects in the physical world. Therefore, implementing IoT means virtualizing physical objects to bring into the networked cyber world. To accomplish this, in the smart sensor microsystem for IoT, we have sensing and actuation function which is necessary for virtualization of physical objects, processing and control function that is also for virtualization, communication function which is for connection, and last but not least, power sourcing and management function which is essential for operating the microsystem. By integrating all these functions, the connected smart sensor microsystem is realized as shown in Fig. 4.1. The aspects of sensing function, communication function, and microsystem integration are mainly investigated in this chapter.

Now, let's look at what are the important factors that are strongly dependent on applications (Fig. 4.2). In the microsystem, its size, lifetime, and physical interface are those factors. In the sensing function, the physical parameters to sense, minimum detection limit, dynamic range, bandwidth, and sensing duty cycles are the application-dependent factors. In the communication function, the communication medium, distance, symmetry, protocol, data rate, and communication duty cycle are such factors. In the following sections, various application-dependent design and integration approaches are investigated and discussed with corresponding examples, mainly in the areas of microsystem integration (Sect. 4.2), sensing (Sect. 4.3), and communication (Sect. 4.4). Then, the conclusion follows in Sect. 4.5.

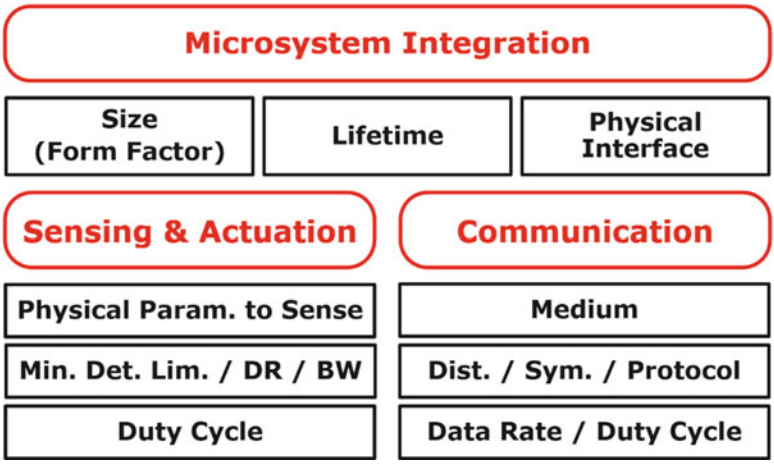


Fig. 4.2 Application-dependent factors in microsystem integration, sensing/actuation, and communication

4.2 Microsystem

4.2.1 Device Size (Form Factor)

The device size or form factor is usually determined by the size of the physical objects where the device is embedded. For example, implants and minimally invasive devices have extreme restriction in their size. In contrast, sensors for smart building and civil structure monitoring don't have much restriction. For wearables, small form factor is important to let users feel comfortable when they wear those devices, but the requirement is not as stringent as the implants and minimally invasive devices.

The size of the device that is constrained by its application in turn determines available capacity of power and energy. Typical primary lithium button cell battery provides 600-mWh energy per cubic centimeter, while rechargeable lithium-polymer cell has the capacity in the range from 250 to 730 mWh per cubic centimeter. Therefore, with current battery technology, if a sensor microsystem of about 1-cm^3 size consumes 1-mW power on average, the microsystem can operate for several hundreds of hours before battery replacement or recharge. If ambient energy sources are tapped to harvest necessary energy for microsystem operation, the power of about 1–100 μW per cubic centimeter is available depending on the kind of sources and environmental conditions (Fig. 4.3 [1]). Consequently, to realize an energy-autonomous microsystem having a form factor less than 1 cm^3, the average power consumption for the overall operation of such a system should be lower than a few microwatts to be on the safe side.

The device form factor also determines microsystem integration scheme. When there is no or little constraint in the device size, the microsystem can be integrated

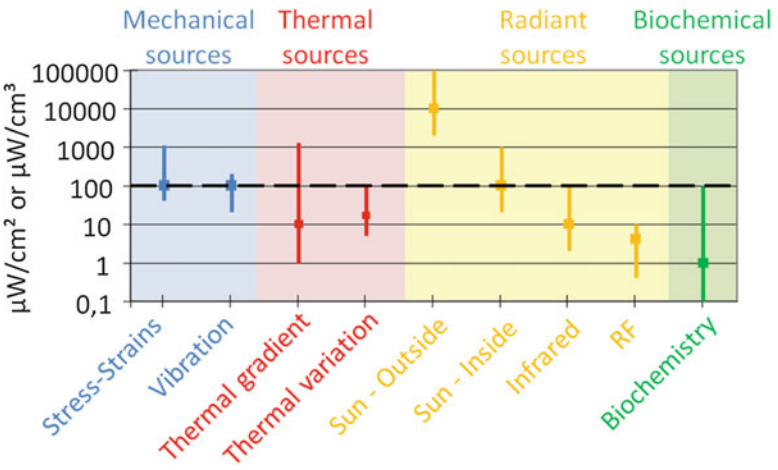


Fig. 4.3 Available power densities from various ambient energy sources [1]

in the form of macromodule. By integrating the microsystem in the form of micromodule, we can make the microsystem form factor much smaller. When there is high restriction in the device form factor, advanced integration technologies have to be used to achieve extreme miniaturization. System-on-chip approach, advanced packaging technology such as wafer-level chip-scale packaging, and 2.5D/3D IC approach based on through-silicon interposer (TSI) and through-silicon-via (TSV) technologies are such integration technologies.

4.2.2 Power Source (or Device) Lifetime

The power source or device lifetime is determined by device usage scenario, accessibility of the installed device, and the number of deployed devices. Since the implant is placed inside a patient's body and surgical procedures are involved to access, its lifetime should be sufficiently long, typically well over 10 years. As the number of sensors in the network grows larger and larger, frequent replacement or recharge of power sources in those sensors will become more and more expensive and eventually impossible. Energy-autonomous sensors will offer nearly infinite lifetime and zero maintenance cost for such large-scale networks. On the other hand, sensors in limited-scale networks that we can find in present smart home and smart building applications, for example, wearable devices, and devices with short-term usage, have less strict requirement on their power source lifetime. Even in those applications of course, longer lifetime is preferred, though not forced.

The required power source or device lifetime determines allowable power consumption and power sourcing/management strategy. For the energy-autonomous sensor node, the average power consumption should be smaller than

the harvested power throughput. When powered by indoor photovoltaic harvesting, for example, the sensor node with a size of 1 cm^3 has to consume the average power lower than 10 μW to achieve energy autonomy, assuming the worst-case harvesting throughput of 10 $\mu W/cm^2$. If the size is 1 mm^3, less than 100 nW should be consumed on average. The implantable cardiac pacemaker with a lithium iodine battery of about 1-Wh capacity should operate with less than 10-μW average power consumption so that the lifetime longer than 10 years can be achieved. To operate the smart watch for longer than 20 h without recharge, the average consumption needs to be lower than 15 mA when it is powered by the rechargeable battery having a capacity of 300 mAh.

4.2.3 Physical Interface with Surroundings

The physical interface of the microsystem with surroundings has to provide an effective isolation barrier between the device and the surroundings to protect one against the other. Hermetic encapsulation isolates the implantable device from surrounding fluid or protects the MEMS sensors against dirt, moisture, and ambient pressure when the sensor is designed to operate in vacuum. Biocompatible packaging is an essential requirement for the implants and on-body devices to minimize the foreign body response and harmful effects on surrounding tissue and cells. Sometimes the package provides EMI shielding if the device operation and performance are sensitive to electromagnetic interference. The MEMS microphone is such an example.

While the effective isolation barrier is constructed on one hand, a proper interface with the surroundings for sensing and actuation has to be provided on the other hand. To do so, controlled minimal exposure of sensors or actuators to the surroundings needs to be implemented somehow, as being done for implantable electrodes, probes, and sensors. For gas sensors, biosensors, and MEMS microphones, proper inlet and guiding microstructures for air and fluid need to be embedded.

4.2.4 Example: Implantable Blood Flow Sensor

The example of application-dependent microsystem integration to investigate is the wireless blood flow monitoring microsystem integrated with a prosthetic vascular graft [2–4]. Prosthetic vascular grafts are widely used for a large number of patients. 20–30 % of the existing renal hemodialysis population has a prosthetic vascular graft in situ, and at least 20 % of bypasses for upper/lower limb ischemia require the use of prosthetic grafts. However, prosthetic grafts are prone to developing progressive stenosis, thrombosis, and ultimately graft abandonment. By monitoring the blood flow rate inside the graft frequently, early stage problem

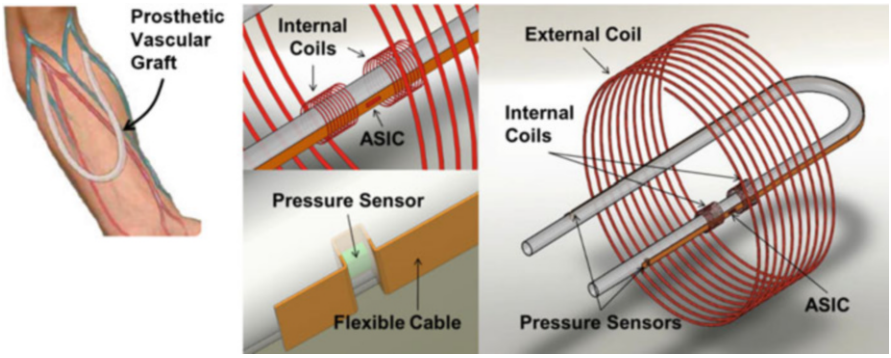


Fig. 4.4 Design and integration approaches of the implantable blood flow monitoring microsystem embedded in the prosthetic vascular graft [2]

detection and intervention can be exercised to prevent graft failure. There exist a few intravascular flow sensing methods such as ultrasound, CT scan, and angiogram. However, they entail morbidity, excessive time and cost, and/or require nephrotoxic contrast, making them ill-suited for frequent monitoring of intravascular blood flow rate.

An implantable wireless blood flow monitoring microsystem embedded in the prosthetic vascular graft provides the most ideal way to frequently monitor the blood flow rate in the graft, when it is paired with an external hand-held monitoring device. Since the microsystem should be integrated within a prosthetic graft, extremely small form factor is required without any room for a battery. The lifetime should be very long as the power source or device replacement incurs a surgery. Since the device is placed inside the human body, biocompatibility is an important concern, while some exposure to the surroundings is also required to sense internal blood flow rate.

Figure 4.4 shows the microsystem design and integration approaches [2]. Since there is no room for battery, we employ passive sensing scheme, where powering the microsystem is accomplished through coil coupling between the implant and the external readout module. The microcoils inside the implant can be placed between the inner and outer layers of the graft. One ASIC, two pressure sensors, and two microcoils are connected using long flexible circuits which can be mounted on the surface of the inner layer. The blood flow rate can be monitored by calculating the difference of the pressure readouts from two pressure sensors located near two ends of the graft. There are minimal incisions at the inner layer to make the pressure sensors exposed to the internal blood stream. The pressure sensors are membrane type rather than cantilever type to avoid endothelialization when exposed to the blood stream. The sensed data are transmitted to the external module through coupled coils by backscattering.

A block diagram of the blood flow monitoring system comprising an implantable microsystem and an external monitoring device is shown in Fig. 4.5 [2]. The microsystem is composed of three key functional subsystems: power link, command/data link, and sensor interface.

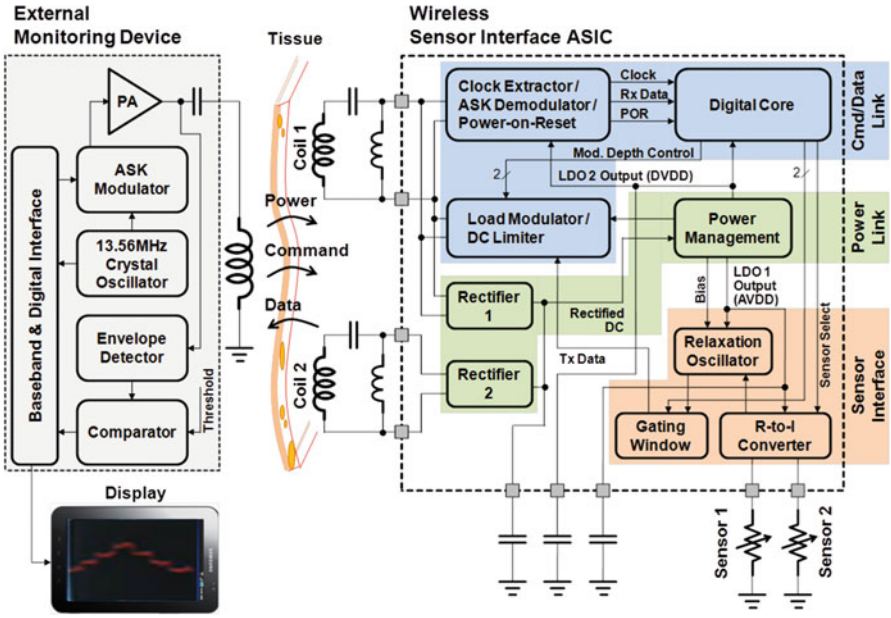


Fig. 4.5 System block diagram of the blood flow monitoring microsystem [2]

- *Wireless power link*: It has two coupling coils, two rectifiers, and a power management block employing low-drop-out voltage regulators (LDOs). The AC power delivered from the primary coil to the secondary coil is converted to the DC power by the rectifiers, which is then regulated by the LDOs to be supplied to circuit blocks in the microsystem.
- *Wireless command/data link*: The command is transmitted using amplitude-shift keying (ASK) with a carrier. After receiving this ASK-modulated RF signal, a clock is extracted from the carrier by a clock extractor and, a power-on-reset (POR) signal is generated to reset the system at the beginning of each measurement cycle. The command is then demodulated by an ASK demodulator. After the microsystem is configured in accordance with the demodulated command, the sensing operation starts. In each measurement cycle, the resistance values of the two pressure sensors are measured in a time-multiplexed manner. The measured resistance values are translated into oscillation frequency by the sensor interface circuit and transmitted to the external monitoring device through backscattering using load-shift keying (LSK).
- *Sensor interface*: The blood flow rate is measured from the difference of pressures sensed by the two MEMS sensors placed at opposite ends of the flexible cable. The piezoresistive sensor transduces the change of the pressure applied on its diaphragm into the resistance change. A sensor interface based on a current-controlled relaxation oscillator is employed to directly translate the resistance value to the oscillation frequency (or the number of counts during a

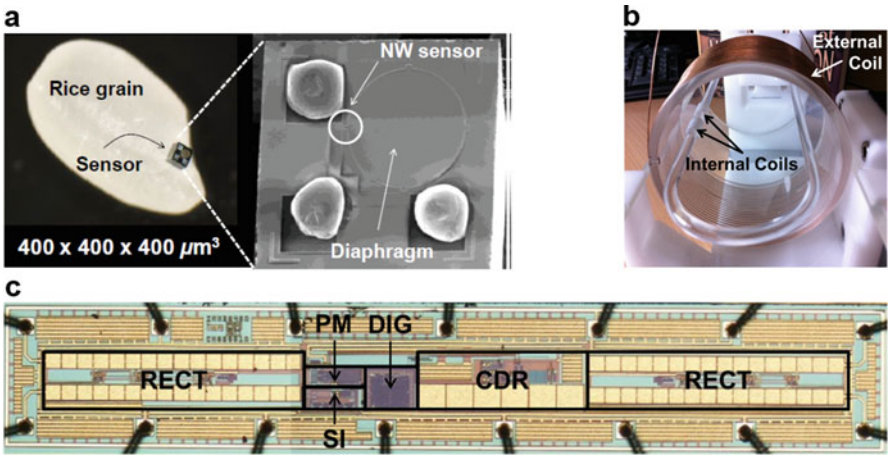


Fig. 4.6 Implementation results of key components [2]: (**a**) MEMS pressure sensor, (**b**) power coupling coils, and (**c**) integrated circuits

predetermined time window). The sensor interface circuit consists of a resistance-to-current converter, a relaxation oscillator, and a gating window control circuit. This approach can greatly simplify the sensor interface design and minimize the power consumption.

The implementation results of key components are shown in Fig. 4.6 [2]. The pressure sensor is made of a silicon nanowire piezoelectric sensor embedded in diaphragm structure. It has an extremely small form factor of $400 \times 400 \times 400\ \mu m^3$, while providing a good sensing performance. For coupling coils, the external coil connected with the readout module is designed to surround the limb. The implanted miniature coils have a diameter of 6 mm. When they operate through 5-cm-thick tissue, the power coupling efficiency is 4.3 %. The integrated circuits support all the necessary functions for microsystem operation, consuming only 12.6 μW.

The results of microsystem integration and wireless sensing test are shown in Fig. 4.7 [2]. To integrate the microsystem within the prosthetic graft, 20-cm-long flexible circuits are mounted on the graft surface and they are coated with the outer layer. The set of materials used for fabricating this new smart graft is the same as that used in the conventional grafts, which is proven to be biocompatible. When the microsystem is tested through 5-cm-thick pork meat as transmission medium, the overall sensing resolution of 0.17 psi is obtained. Importantly, only 600-μW transmission power is sufficient to operate the sensor microsystem, which is possible because of low power consumption and high-efficiency power transfer achieved by the microsystem. It is the very low-level transmission posing no risk of harmful effects on the patient's body.

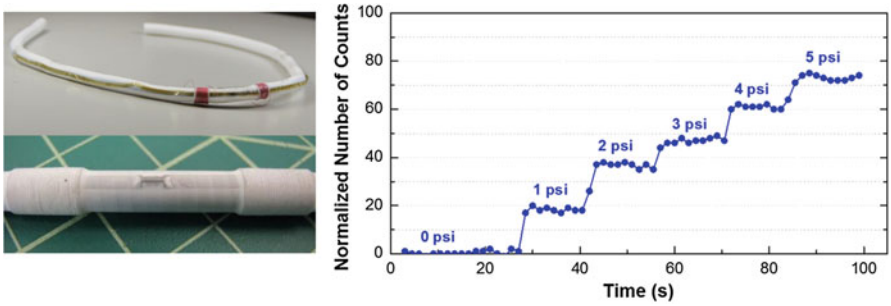


Fig. 4.7 Results of microsystem integration and wireless sensing test

4.3 Sensing

4.3.1 Physical Parameters to Sense

For different applications, different physical parameters need to be sensed for virtualization of "things." The parameters usually sensed are physical forces (e.g., pressure, vibration, acceleration, angular velocity, flow, and tactile force), particles, ions, chemicals, molecules, temperature, humidity, light, potential, and current. To sense different physical parameters, different sensors and corresponding interface circuits are used.

The sensors transduce the physical parameters of interest into electrical quantities. The type of the sensor used determines in which electrical quantity the sensor output is. For example, capacitive accelerometers and pressure sensors are widely used and generate the output in the form of capacitance. Piezoresistive force sensors transduce the physical force into the resistance output. The output of biopotential electrodes and piezoelectric force sensors is in the form of electrical voltage. Ion-sensitive field-effect transistors (ISFET) and optical sensors generate current output.

4.3.2 Minimum Detection Limit, Bandwidth, and Dynamic Range

The minimum detection limit required for the sensing function is determined by the intrinsic strength of the signal that needs to be sensed, the distance from the signal origin, and the signal level of interest. Recording of various biopotential signals illustrates these points very well (Fig. 4.8). Local field potential (LFP), electrocorticogram (ECoG), and electroencephalogram (EEG) have the same signal origin. But the signal amplitudes are not the same as the distance between the location we record and the origin of signal is different. The closer to the origin, the larger the

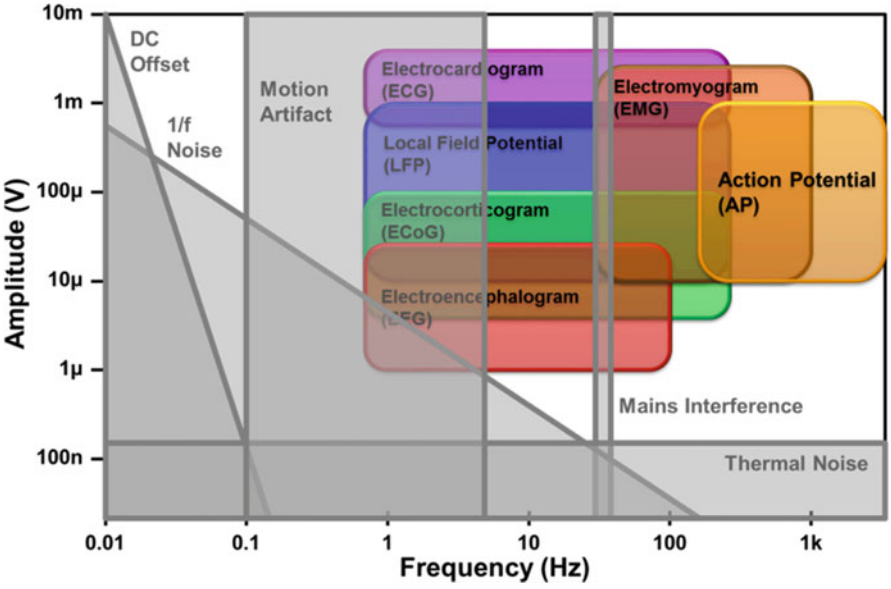


Fig. 4.8 Amplitude and frequency characteristics of various biopotential signals shown with the effect from interferences and disturbances

Table 4.1 Indoor air quality guidelines for VOCs

VOCs	Guideline values
Formaldehyde HCHO	30–120 µg/m^3 (25–100 ppb)
Acetaldehyde CH_3CHO	48 µg/m^3 (27 ppb)
Benzene C_6H_6	16–110 µg/m^3 (5–34 ppb)
Toluene $C_6H_5CH_3$	260–1092 µg/m^3 (69–290 ppb)
Ethylbenzene $C_6H_5CH_2CH_3$	1447–3880 µg/m^3 (264–892 ppb)
Xylene $C_6H_4(CH_3)_2$	870–1447 µg/m^3 (200–333 ppb)
Styrene $C_6H_5CH{=}CH_2$	30–300 µg/m^3 (7–70 ppb)
TVOC	200–3000 µg/m^3

amplitude we obtain, resulting in more relaxed requirement on the minimum detection limit. Compared to LFP and action potential (AP), the electrocardiogram (ECG) signal usually has larger amplitude because of the difference in their intrinsic signal strengths. The LFP and AP are originated from the activities of neurons in the brain, while the ECG is from the activities of electrogenic cells in cardiac muscles. Table 4.1 shows the indoor air quality guidelines especially for volatile organic compounds (VOCs) harming human health significantly. The range of guideline values covers the values used in different countries. Since the harmful effect is caused even with a very little amount of such gases, the required detection limit reaches as low as several ppb levels.

Required sensing bandwidth is mainly determined by the intrinsic signal characteristics and the distance between the signal origin and the sensing site. Except some applications sensing images and videos, the bandwidth requirement is not usually so high, but tends to be higher when the sensor is placed closer to the signal origin. For example, although neural signals such as AP, LFP, ECoG, and EEG share the same signal origin, the signal bandwidth for invasive neural recording is larger than that for non-invasive EEG recording (Fig. 4.8).

The dynamic range requirement is determined by intrinsic signal or sensor characteristics, effect from various interferences and disturbances, and the signal range of interest. For example, assume that we record the LFP and AP signals simultaneously. Since the spikes in the AP signal having an amplitude of tens or hundreds of microvolts appear on top of the LFP signal with an amplitude of a few millivolts, if an input-referred noise of 4 μV_{rms} is needed to meet the signal-to-noise ratio requirement of the spike signal recording and the LFP signal has an amplitude of 2 mV, the dynamic range requirement becomes 54 dB, resulting in about 9-bit resolution. In addition, if we consider interferences as large as a few hundreds of millivolts, the required dynamic range may reach nearly 100 dB. To avoid that, a good common-mode rejection performance needs to be guaranteed. Some sensors such as gas sensors and inertial sensors for navigation require very large dynamic range. Metal-oxide thin-film gas sensors require the interface circuit having a large dynamic range over 140 dB to accommodate high sensitivity, process spread, drift over time, and large variation in sensor resistance values for different doping types and levels used to detect different target gases. The navigation-grade accelerometer requires the dynamic range larger than 100 dB.

Aforementioned requirements on minimum detection limit, bandwidth, and dynamic range determine the necessary sensing performances and power consumption. Stringent requirements on minimum detection limit, bandwidth, and dynamic range usually lead to significant power consumption, and hence low-power circuit techniques are needed more strongly. To provide necessary performances, the sensor and its interface circuit often need to be calibrated. The calibration strategy is another important factor to consider. Especially when the application requires low-cost sensing solution, the time and effort spent for calibration should be minimized while providing necessary performance after calibration. In case of nano-scale sensors, they tend to have large variations in their characteristics at the cost of ultra-high sensitivity offered [5]. Therefore, effective and efficient calibration method has to be developed.

4.3.3 Sensing Duty Cycle

Depending on the characteristics of target physical parameters, the sensing operation can be duty-cycled to minimize the overall energy consumption as well as the amount of the data that need to be transmitted. Very low-duty-cycle sensing can be utilized when slow varying signals are monitored as in the cases of environmental

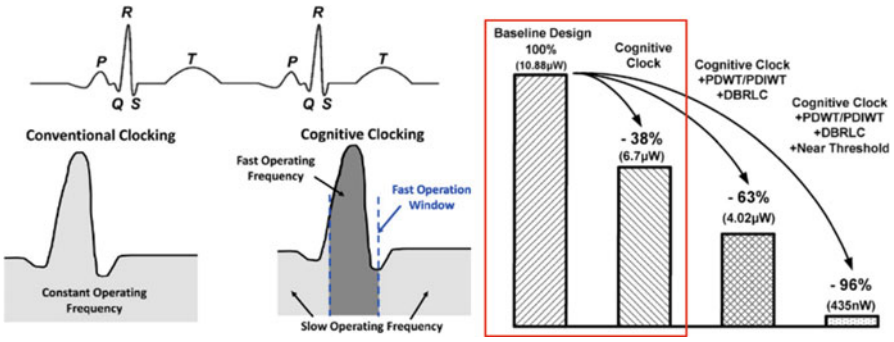


Fig. 4.9 Ultra-low-power electrocardiogram (ECG) processor employing cognitive clocking [6]

monitoring, civil structure monitoring, and progress monitoring of chronic diseases. In environmental monitoring, the sensing interval can be hours, and for intraocular pressure monitoring, the pressure may be recorded every 15 min, for example. When we sense scarce and bursty signals such as ECG signal and neural spikes, fine-grain duty cycling or adaptive operation can be employed. For example, in neural spike recording, by implementing a spike detection function, we may execute data conversion process only when spikes are detected.

Such duty-cycled sensing operation leads to reduction of average power consumption as well as sensor data. Moreover, duty cycling in sensing function does not entail any synchronization issues unlike the case of duty cycling in communication function where the synchronization between sensor nodes or between a sensor node and a host device poses a significant challenge. Figure 4.9 shows the example of adaptive or cognitive clocking scheme used for ultra-low-power ECG signal monitoring and processing [6, 7]. By using this technique, the average power consumption is reduced by 40 % without causing any timing or synchronization issues.

4.3.4 Example: Neural Recording

The example to investigate for application-dependent sensing approach is implantable neural recording and the target application is implantable neuroprobe microsystem for motor prosthesis (Fig. 4.10) [8–11]. This microsystem consists of a probe array inserted into motor cortex of human brain, a neural recording IC (IC1) which is directly integrated on the top plate of the probe array, a wireless power and data link IC (IC2) which is placed between the skull and scalp, a flexible cable connecting the neural recording IC and the wireless power and data link IC. In the microsystem, the neural recording IC working together with the probe array senses multichannel neural signals by providing functions of amplification, filtering, and digitization. This application doesn't require extremely small minimum

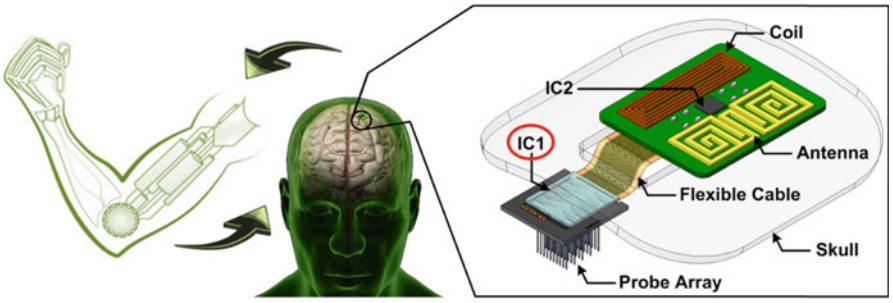


Fig. 4.10 Implantable neuroprobe microsystem for motor prosthesis

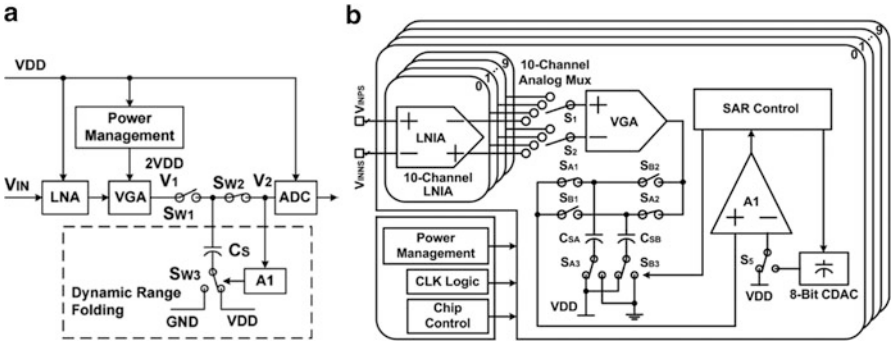


Fig. 4.11 Dual-supply ultra-low-power neural recording IC with dynamic range folding [9]: (**a**) block diagram showing only one recording channel, (**b**) block diagram of 100-channel recording IC

detection limit, as the neural spikes are recorded invasively and therefore have a relatively large amplitude. Compared to other types of neural signals, the spike signal requires a larger bandwidth of over 5 kHz and a moderate dynamic range of about 60 dB. Since the microsystem is implanted inside human body, ultra-low-power consumption should be achieved. Utilizing bursty signal characteristics of the neural spikes, duty-cycled sensing can be implemented.

4.3.4.1 Dual-Supply Ultra-Low-Power Neural Recording IC with Dynamic Range Folding

Figure 4.11a shows the block diagram of the proposed neural recording circuit for one channel [9]. It is dual-supply ultra-low-power neural recording IC with dynamic range folding. The supply voltage of 0.45 V ($=V_{DD}$) is used for the low-noise instrumentation amplifier (LNIA) as it consumes high current for low-noise operation but has low voltage swing. 0.45 V is used for the successive approximation analog-to-digital converter (SAR ADC) as well to reduce the

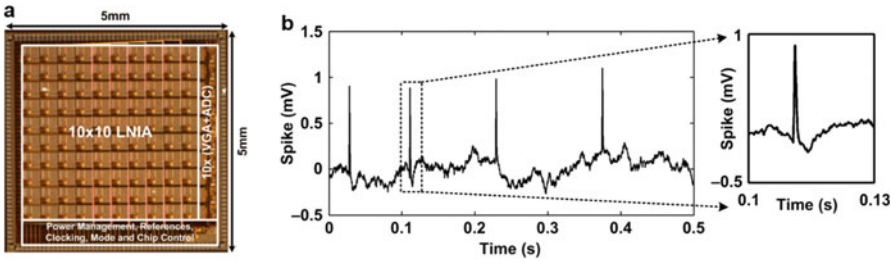


Fig. 4.12 Implementation results of the neural recording IC [9]: (**a**) microphotography of the fabricated chip, (**b**) in vivo neural signal recorded from the rat model

charging and discharging power consumed in its capacitive digital-to-analog converter (DAC) circuit and the digital switching power consumed in its control logic and register circuit. For the variable gain amplifier (VGA), as it consumes very low current but requires high voltage swing, we power this block with $2V_{DD}$ which is 0.9 V. However, we then have to resolve the mismatch in the signal dynamic range between the VGA and ADC which is caused by this dual-supply scheme. To solve this problem, a new dynamic-range-folding technique is applied and the details of this technique can be found in [8, 9]. By doing so, we could save 50 % of the power consumption without degrading noise, bandwidth, and dynamic range performances. The block diagram of the entire 100-channel neural recording IC is shown in Fig. 4.11b [9]. It consists of ten recording blocks, each of which contains ten recording channels. 10-to-1 analog multiplexer is located between ten neural amplifiers and a single VGA not only to minimize the area consumption but also to simplify the implementation of the multiplexer and the pipeline sample-and-hole network.

Figure 4.12a shows the proposed neural recording IC fabricated in 0.18-μm CMOS technology, occupying the die area of 5 × 5 mm [9]. By multiplexing before the VGA, the area consumption is saved by 22 % overall. The dual-supply scheme combined with dynamic range folding technique enabled extremely low power consumption of 730 nW/channel while providing 9-bit sensing resolution. Figure 4.12b presents the recorded in vivo neural signal, which is reconstructed from the ADC output code and scaled for input-referring [9]. The characteristic biphasic neuronal spike is indicative of a recording from the cell body of a neuron in the anterior cingulate gyrus, a region in the frontal cortex of the brain.

4.3.4.2 Ultra-Low-Power Neural Recording IC with Delta-Modulation-Based Spike Detection

Since the spikes are rare events in neural signal (10–100 fires/s typically), it is desirable to record only the spikes in order to minimize power dissipation as well as the amount of recording data to transmit while preserving the fundamental information of neuronal activities. A spike detection function can therefore be

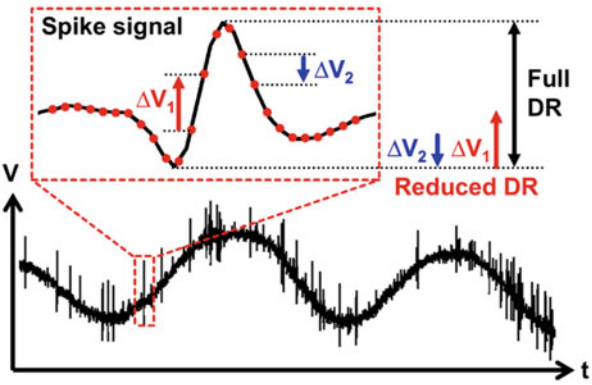


Fig. 4.13 Typical waveform of neural signal waveform (LFP and spikes), with the concept of dynamic range compression and spike detection illustrated in the inset [10]

implemented for duty cycling. Figure 4.13 presents typical neural signal waveform consisting of LFP and spikes together with the concept of delta-modulation-based spike detection [10]. For spike detection, consecutive delta values of the input signal are monitored to extract the frequency and amplitude information. By checking the polarity of the delta values continuously, too high and too low frequency components can be rejected on one hand. On the other hand, too low amplitude and too low frequency components can be ruled out by checking the size of the delta values. Using this delta-modulation-based spike detection technique, we can compress effective dynamic range and bandwidth to cover and enable fine-grain duty cycling, while preserving the neural spike waveform information.

The block diagram of the proposed 16-channel neural recording IC is depicted in Fig. 4.14 [10]. The delta modulator following a low-noise amplifier (LNA) subtracts two consecutive neural signal values so that the signal dynamic range can be reduced by half the normal range. Up to six values resulting from the subtraction in series are stored in analog first-in-first-out (FIFO) memory for a spike detector in the next stage. The detector recognizes the frequency as well as the amplitude of the signal from the stored set of several consecutive delta samples, extracting the spikes from the entire neural signal accurately. The spikes are detected during the sampling phase before the 8-bit SAR ADC operates. The ADC and transmitter (not included in this design) therefore consume power only when they process the spikes, realizing the neural recording system that runs with optimal energy efficiency.

A prototype neural recording IC with 16 channels has been fabricated using 0.18-μm CMOS technology and fully characterized. Figure 4.15a shows the micrograph of the fabricated recording chip [10]. The area of the whole chip and one recording channel is 2.35×2.5 mm and 200×800 μm, respectively. Because the delta modulator compresses the dynamic range, we can reduce 1 bit of the ADC resolution, implementing 8-bit SAR ADC. Moreover, the MSB is equivalent to the sign of the signal, and hence 7-bit capacitive DAC is sufficient to cover the entire signal range. In order to verify the spike detection functionality, in vitro test of arbitrary neural signal acquisition has been performed. We used the pre-recorded

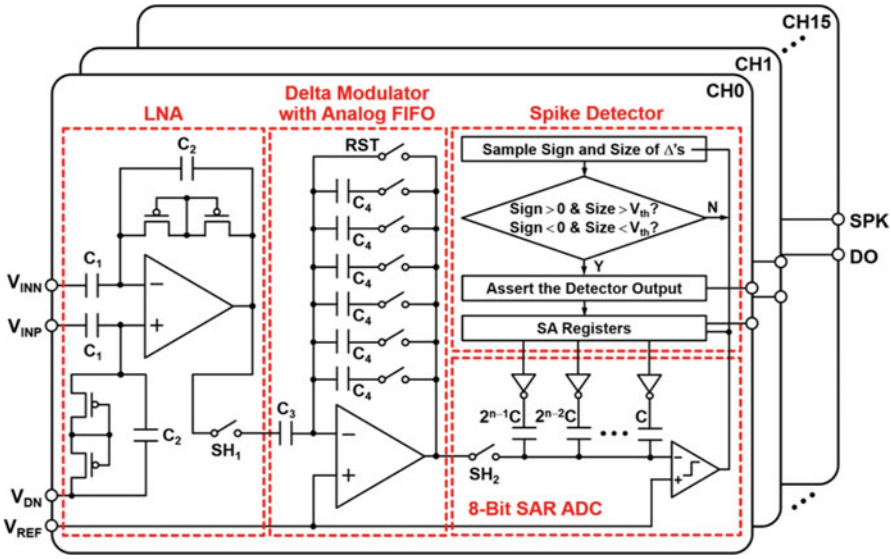


Fig. 4.14 Block diagram of the 16-channel neural recording IC with delta-modulation-based spike detection [10]

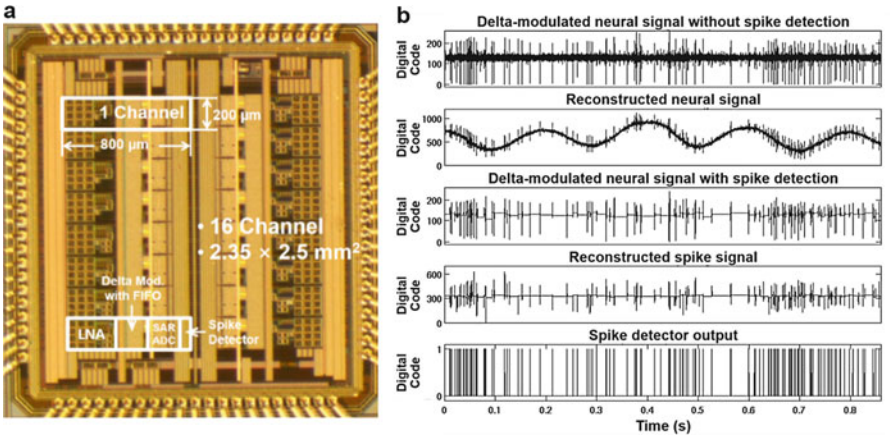


Fig. 4.15 Implementation results of the neural recording IC [10]: (**a**) microphotography of the fabricated chip, (**b**) recorded neural signal waveforms with and without activating the spike detector

neural signal provided by Plexon for the assessment of the spike detector. Figure 4.15b shows the neural signal waveforms recorded for 0.9 s with and without activating the spike detector [10]. The upper two plots are the output of the delta modulator and the reconstructed neural signal obtained by integrating the delta signal. The lower three plots show the extracted spike signal in delta values, the reconstructed spike signal, and the output of the spike detector.

4.4 Communication

4.4.1 Communication Medium

The communication medium is an application factor in the communication function. The medium used for communication is determined by the location of the physical objects where the microsystem is embedded. In most cases, communication is performed wirelessly thorough free air. However, some devices communicate through other communication media. Wearables and implants may use human body as a communication channel. Smart electricity meters may use power lines and sensors for oil exploration may use drilling pipes.

Then, in turn, the communication medium determines communication frequency, bandwidth, and achievable data rate. When we communicate through human body, the frequency band from 40 to 70 MHz as human body has bandpass characteristics. Safety is another important concern in this case. In logging-while-drilling (LWD) tools for oil well exploration, if the telemetry is implemented using electromagnetic propagation, it cannot reach more than a few kilometers and the data rate is limited to a few bits per second (b/s). If mud pulse is used for telemetry, it can reach longer distance, but the data rate is still limited to about 10 b/s. People are thus developing an ultrasound-based telemetry through the drilling pipes, which can achieve much higher data rate (well over 10 kb/s) than other approaches.

4.4.2 Example: Body-Channel Communication

By using the human body as a communication medium, the body-channel communication can achieve high energy efficiency compared to other wireless communication approaches as shown in Fig. 4.16b [13]. However, the data rate tends to be rather low, being insufficient for multimedia communication between wearable devices, such as between smart watch and smart glasses (Fig. 4.16a). A high-data-rate wideband body-channel communication transceiver therefore needs to be developed [12, 14, 15]. As mentioned in Sect. 4.4.1, the body channel has bandpass characteristics and two plots in Fig. 4.17 show such characteristics. The top figure shows the channel gain for the path A which is within one arm, and the bottom one is for the path T which is from one arm to the other across the torso.
The block diagram in Fig. 4.18a shows the structure of the wideband receiver [12]. We use tri-level direct digital Walsh-coded signaling for high data rate and high energy efficiency. To optimize the channel characteristics and achieve high data rate with limited channel bandwidth, we use high input impedance for matching and employ an equalizer in the receiver. Since the channel doesn't pass DC signal information, a transient-detection architecture is proposed using a differentiator followed by an integrator. An injection-locking-based clock recovery

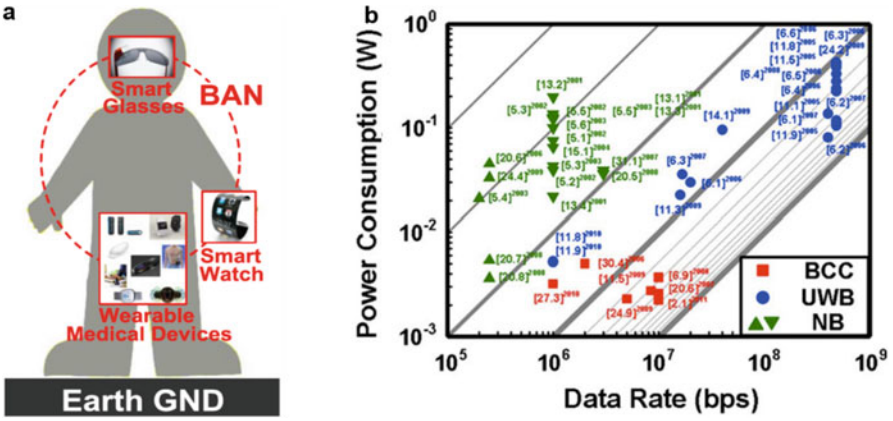


Fig. 4.16 Body-channel communication (BCC): (**a**) energy-efficient communication for wearable devices [12], (**b**) power consumption and data rate of BCC receivers in comparison with ultra-wideband (UWB) and narrow-band (NB) receivers [13]

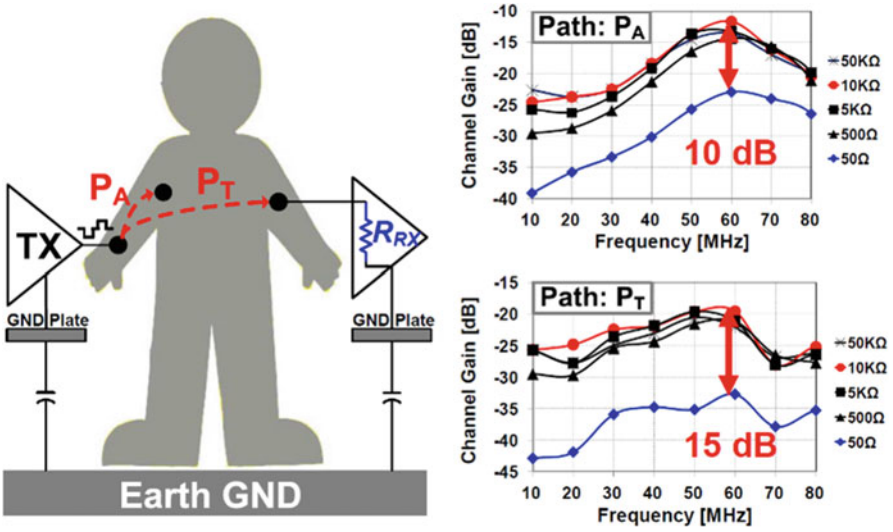


Fig. 4.17 Body-channel characteristics [12]

is also employed to provide the timing signal to the integrator and level detector, eliminating the need of a power-consuming PLL and a bulky crystal oscillator. Figure 4.18b presents the block diagram and measured output of the transmitter [12]. The transmitter consists of tri-level Walsh code modulator followed by tri-level driver. It can be seen that the dominant signal energy of the transmitter output is confined within the passband.

Figure 4.19 presents the implementation results [12]. The microphotography of the fabricated receiver and transmitter chips is shown in Fig. 4.19a. The transceiver

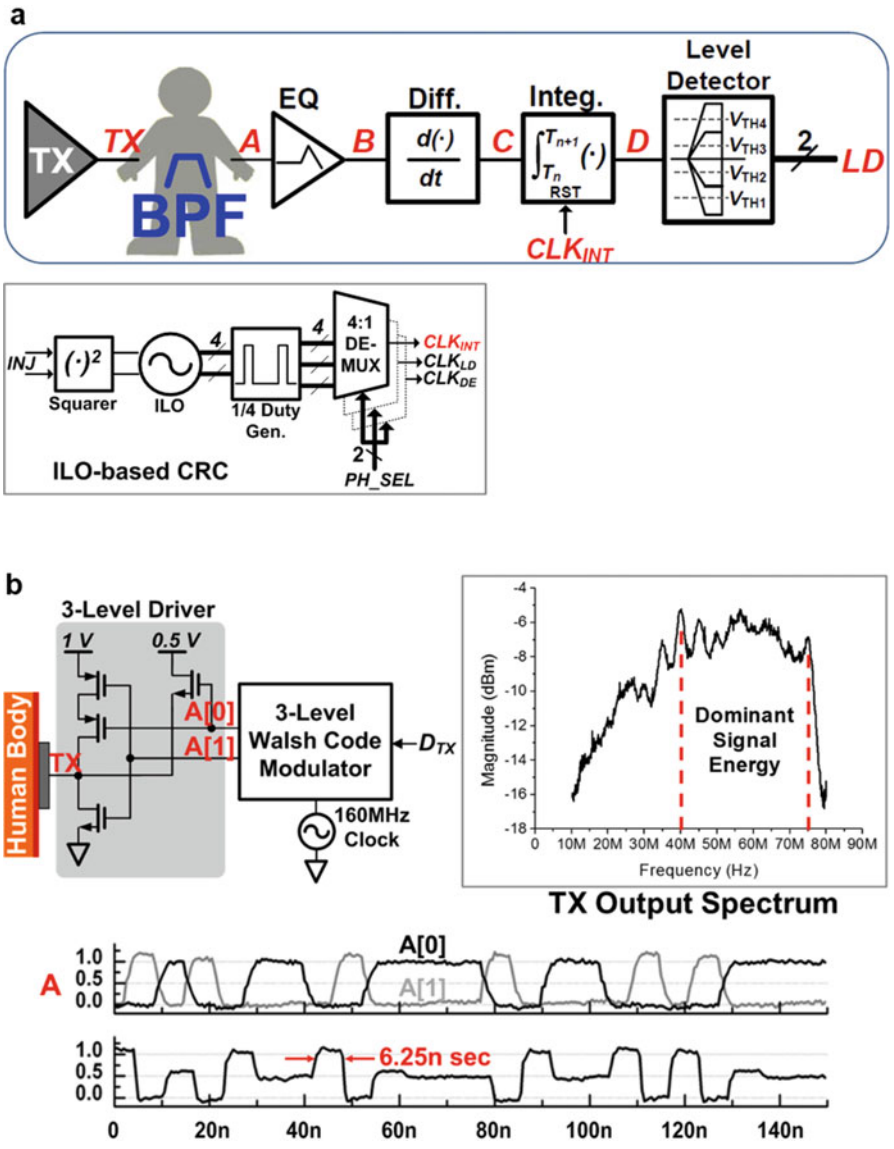


Fig. 4.18 High-energy-efficiency wideband BCC transceiver [12]: (**a**) block diagram of the receiver and clock-recovery circuit, (**b**) block diagram and measured output of the transmitter

is implemented in 65-nm CMOS process and consumes 1.85 mW in the transmitter and 9 mW in the receiver. Figure 4.19b shows the test setup and results of an image data transfer experiment. The transmitter is placed on the left arm and receiver on the right arm. The Lena image was successfully transferred in the testing.

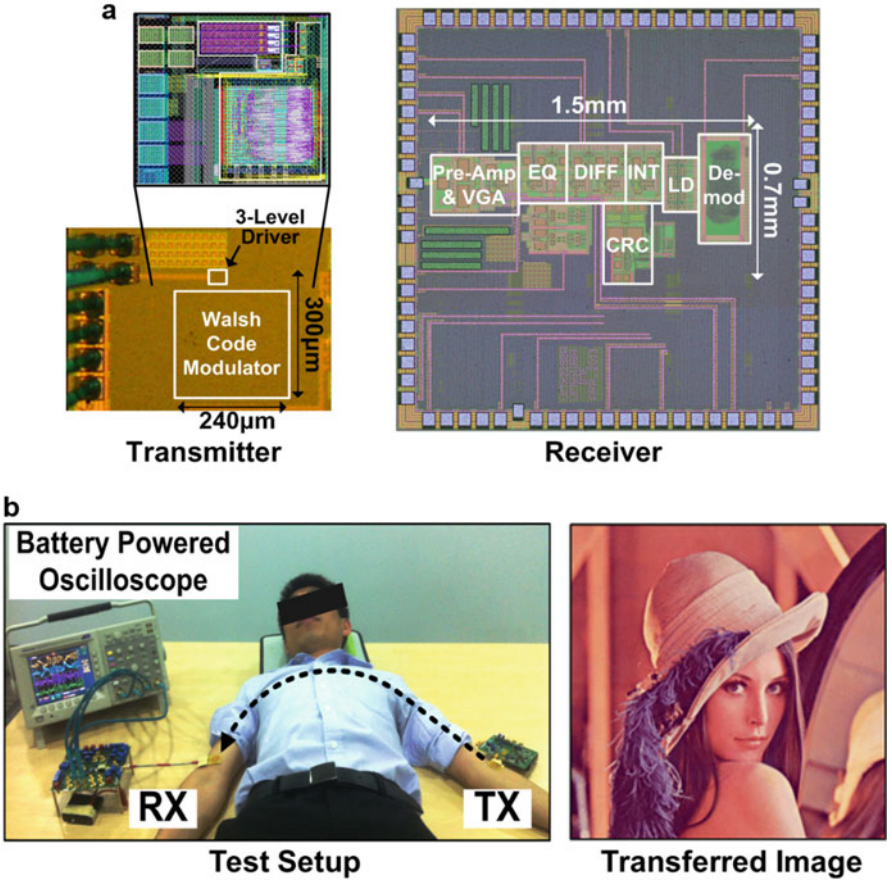


Fig. 4.19 Implementation results of the BCC transceiver ICs [12]: (**a**) microphotography of the fabricated chips, (**b**) test setup and results of an image data transfer experiment

4.4.3 Distance, Symmetry, and Protocol

Other application-dependent factors are communication distance, symmetry, and network topology. Wide-coverage sensor networks with a large number of sensor nodes require long communication distance, symmetric transceiver for multi-hop and ad-hoc communication. Such examples are smart grid and oil pipeline leak detection sensor network. On the other hand, narrow-coverage sensor networks with a small number of sensor nodes operate with short communication distance and asymmetric transceiver for master–slave communication. By doing so, the energy consumption of slave node can be minimized, while burning more power in the master side operating with sufficient energy source typically. Examples are the smart home network and body area network.

4.4.4 *Example: Wireless Capsule Endoscopy*

For wireless capsule endoscopy application (Fig. 4.20a), an asymmetric short-range transceiver is developed, together with a JPEG encoder IC [16–19]. It is for enabling the next-generation wireless capsule endoscopy, which offers high image resolution, high frame rate, and bidirectional wireless link for image transmission (uplink) as well actuator control (downlink). Considering this specific application, the transceiver has to provide high data rate to transmit high-resolution high-frame-rate images. It should consume low power so that the microsystem can operate for several hours with 2 or 3 very small coin cell batteries. To communicate through the human body with a small antenna, the communication frequency should be carefully chosen. Considering large signal attenuation at high frequencies and low antenna gain at low frequencies, 900-MHz frequency band is used in this work. Figure 4.20b shows the system block diagram of the wireless endoscopy capsule. The system has a quadrature phase-shift keying (QPSK) uplink for image transmission with 3–20-Mb/s data rate and 5–6-mW power consumption. The downlink for actuator control is based on OOK. The data rate is 500 kb/s and the power consumption is 4.5 mW. The JPEG encoder is also developed to further boost the effective data rate.

For energy-efficient implementation of the QPSK transmitter, a new injection-locking-based direct phase modulation technique is proposed [18, 19] as shown in Fig. 4.21 [16]. It changes the self-resonance frequency of a sub-harmonic injection-locked oscillator through capacitor bank switching. Since the output phase of the injection-locked oscillator is determined by the difference between the injection frequency and the self-resonance frequency, we can modulate the output phase by controlling the self-resonance frequency of the oscillator.

Figure 4.22a presents a prototype multifunctional endoscopy capsule which is built using the proposed chipset including the RF transceiver IC and JPEG encoder IC [16]. Both ICs have been fabricated in a 0.18-μm CMOS process. Their die micrographs are shown in Fig. 4.22a. The transceiver IC and JPEG encoder IC occupy the die area of 2.6×2.8 mm^2 and 2.6×2.6 mm^2, respectively. The overall

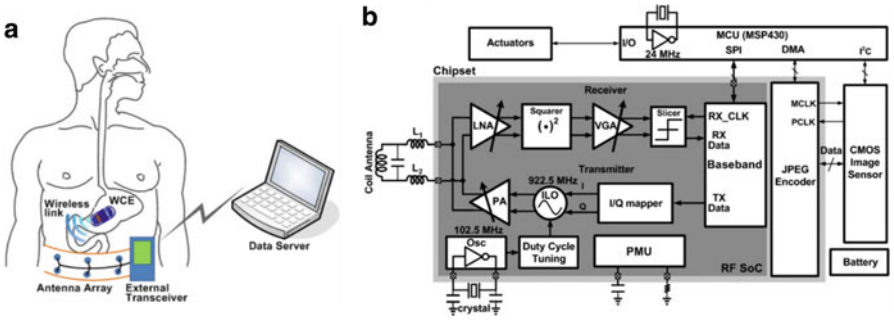


Fig. 4.20 (**a**) Wireless capsule endoscopy [16], (**b**) system block diagram of the wireless endoscopy capsule including the asymmetric short-range transceiver and JPEG encoder ICs [16]

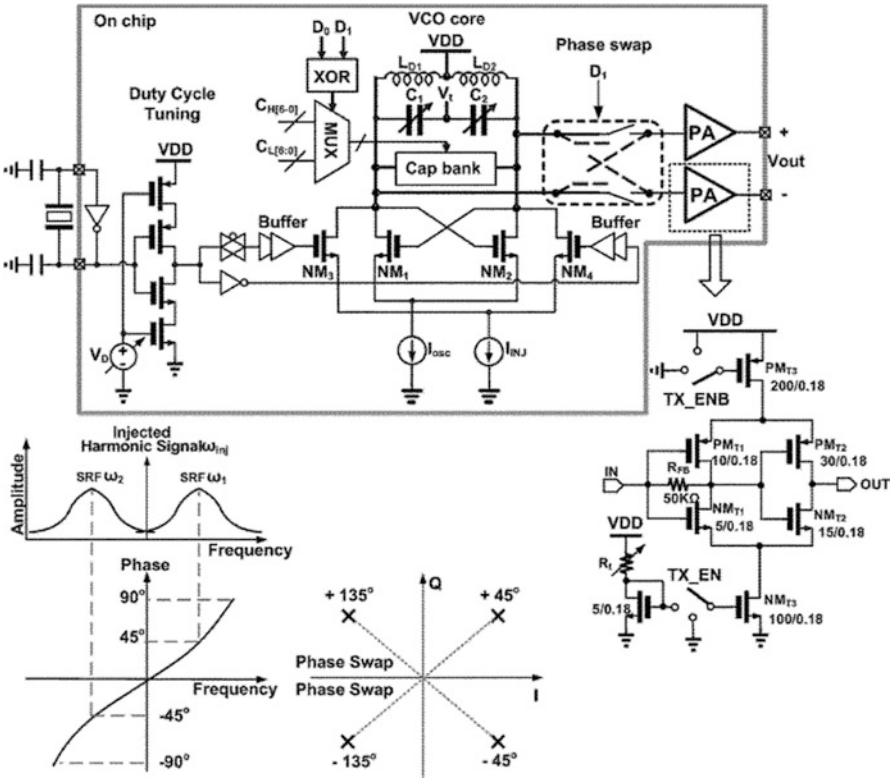


Fig. 4.21 Quadrature phase-shift keying (QPSK) transmitter design using injection-locking-based direct phase modulation [16]

capsule dimensions are 37 mm in length by 11.5 mm in diameter. About 16 % of the capsule space is reserved for a tagging actuator to cater for lesion localization. The RF transceiver, JPEC encoder, and image sensor chips are directly assembled on a rigid-flex PCB and protected by epoxy. The front and back sides of the assembled PCB system are shown in Fig. 4.22a. The full system performance has been verified by an in vivo animal experiment on porcine model as shown in Fig. 4.22b [16]. An external transceiver is implemented with discrete components. The internal images of the stomach muscle were successfully transmitted to the external receiver and displayed on the PC screen.

4.4.5 Data Rate

Being another application-dependent factor, the communication data rate is determined by the amount of sensor data to communicate. The data rate requirement is not usually so high for most of smart sensor microsystems. However, some devices

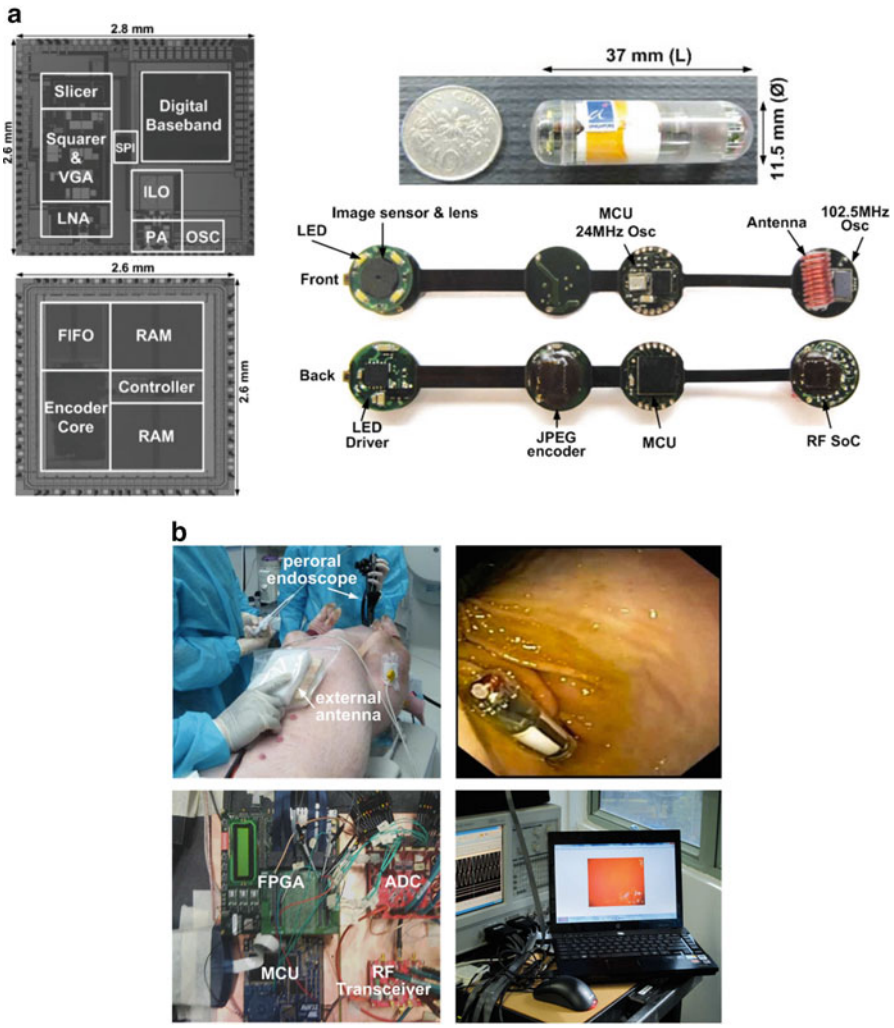


Fig. 4.22 Implementation results [16]: (**a**) chipset fabrication and microsystem integration, (**b**) test setup and results of in vivo animal experiment on porcine model

dealing with image and video data (e.g., video surveillance and capsule endoscopy) and information from large-scale sensor array (e.g., multichannel neural recording) require high-data-rate communication.

4.4.6 Communication Duty Cycle

The communication duty cycle is determined by characteristics of sensed physical parameters and usage scenario. Since the communication transceiver is usually the most power hungry block in wireless sensor microsystems, a significant amount of

energy can be saved by duty cycling the communication function. However, it entails synchronization issues, which can be mitigated by using a low-power high-accuracy clock generator, a low-power wake-up receiver, or a combination of both.

4.5 Conclusion

Unlike PC and mobile phone platforms, the smart sensor platform is used to virtualize physical objects and bring them into the networked cyber world. It therefore has so many different applications, which lead to different approaches for designing and implementing smart sensor microsystems. For different application, different strategies for sensing, processing, communication, powering, and microsystem integration are employed to provide optimum solutions.

For IC designers, it's an important prerequisite to have a good understanding of the applications where the designed circuits and systems are deployed. While improving circuit-level techniques on one hand, optimally crafting system designs on the other hand can draw the maximum out of currently available technologies. As a result, IC designers become more and more like solution providers who develop the best solutions for the problems posed by specific applications, leveraging state-of-the-art IC design techniques.

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

References

1. Boisseau S, Despesse G, Ahmed Seddik B (2012) Electrostatic conversion for vibration energy havesting. In: Lallart M (ed) Small-scale energy harvesting. InTech, Chapter 5, p 92
2. Cheong JH, Ho CK, Ng SSY, Xue R-F, Cha H-K, Khannur PB, Liu X, Lee AA, Endru FN, Park W-T, Lim LS, He C, Je M (2012) A wirelessly powered and interrogated blood flow monitoring microsystem fully integrated with a prosthetic vascular graft for early failure detection. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2012, p 177–180
3. Cheong JH, Ng SSY, Liu X, Xue R-F, Lim HJ, Khannur PB, Chan KL, Lee AA, Kang K, Lim LS, He C, Singh P, Park W-T, Je M (2012) An inductively powered implantable blood flow sensor microsystem for vascular grafts. IEEE Trans Biomed Eng 59(9):2466–2475
4. Khannur PB, Chan KL, Cheong JH, Kang K, Lee AA, Liu X, Lim HJ, Ramakrishna K, Je M (2010) A 21.6μW inductively powered implantable IC for blood flow measurement. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2010, p 9–5
5. Chai KTC, Choe K, Bernal OD, Gopalakrishnan PK, Zhang G-J, Kang TG, Je M (2010) A 64-channel readout ASIC for nanowire biosensor array with electrical calibration scheme. Proceedings of annual international conference of the IEEE engineering in medicine and biology society, Sept 2010, p 3491–3494

6. Liu X, Zhou J, Yang Y, Wang B, Lan J, Wang C, Luo J, Goh WL, Kim TT-H, Je M (2014) A 457 nW near-threshold cognitive multi-functional ECG processor for long-term cardiac monitoring. IEEE J Solid State Circuits 49(11):2422–2434
7. Liu X, Zhou J, Yang Y, Wang B, Lan J, Wang C, Luo J, Goh WL, Kim TT-H, Je M (2013) A 457-nW cognitive multi-functional ECG processor. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2013, p 141–144
8. Han D, Zheng Y, Rajkumar R, Dawe GS, Je M (2013) A 0.45 V 100-channel neural-recording IC with sub-μW/channel consumption in 0.18 μm CMOS. IEEE Trans Biomed Circ Syst 7 (6):735–746
9. Han D, Zheng Y, Rajkumar R, Dawe G, Je M (2013) A 0.45V 100-channel neural recording IC with sub-μW/channel consumption in 0.18μm CMOS. IEEE international solid-state circuits conference digest of technical papers, Feb 2013, p 290–291
10. Kim S-J, Liu L, Yao L, Goh WL, Gao Y, Je M (2014) A 0.5-V sub-μW/channel neural recording IC with delta-modulation-based spike detection. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2014, p 189–192
11. Cheng K-W, Zou X, Cheong JH, Xue R-F, Chen Z, Yao L, Cha H-K, Cheng SJ, Li P, Liu L, Andia L, Ho CK, Cheng M-Y, Duan Z, Rajkumar R, Zheng Y, Goh WL, Guo Y, Dawe G, Park W-T, Je M (2012) 100-channel wireless neural recording system with 54-Mb/s data link and 40%-efficiency power link. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2012, p 185–188
12. Lee J, Kulkarni VV, Ho CK, Cheong JH, Li P, Zhou J, Toh WD, Zhang X, Gao Y, Cheng KW, Liu X, Je M (2014) A 60Mb/s wideband BCC transceiver with 150pJ/b RX and 31pJ/b TX for emerging wearable applications. IEEE international solid-state circuits conference digest of technical papers, Feb 2014, p 498–499
13. Bae J, Song K, Lee H, Cho H, Yoo H-J (2012) A 0.24-nJ/b wireless body-area-network transceiver with scalable double-FSK modulation. IEEE J Solid State Circuits 47(1):310–322
14. Ho CK, Cheong JH, Lee J, Kulkarni V, Li P, Liu X, Je M (2014) High bandwidth efficiency and low power consumption Walsh code implementation methods for body channel communication. IEEE Trans Microwave Theory Tech 62(9):1867–1878
15. Kulkarni VV, Lee J, Zhou J, Ho CK, Cheong JH, Toh W-D, Li P, Liu X, Je M (2014) A reference-less injection-locked clock-recovery scheme for multilevel-signaling-based wideband BCC receivers. IEEE Trans Microwave Theory Tech 62(9):1856–1866
16. Gao Y, Cheng S-J, Toh W-D, Kwok Y-S, Tan K-CB, Chen X, Mok W-M, Win H-H, Zhao B, Diao S, Cabuk A, Zheng Y, Sun S, Je M, Heng C-H (2013) An asymmetrical QPSK/OOK transceiver SoC and 15:1 JPEG encoder IC for multifunction wireless capsule endoscopy. IEEE J Solid State Circuits 48(11):2717–2733
17. Gao Y, Cheng S-J, Toh W-D, Kwok Y-S, Tan K-CB, Chen X, Mok W-M, Win H-H, Zhao B, Diao S, Cabuk A, Zheng Y, Sun S, Je M, Heng C-H (2012) An asymmetrical QPSK/OOK transceiver SoC and 15:1 JPEG encoder IC for multifunction wireless capsule endoscopy. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2012, p 341–344
18. Diao S, Zheng Y, Gao Y, Cheng S-J, Yuan X, Je M, Heng C-H (2012) A 50-Mbps CMOS QPSK/O-QPSK transmitter used in endoscopy by employing injection locking for direct modulation. IEEE Trans Microwave Theory Tech 60(1):120–130
19. Diao S, Zheng Y, Gao Y, Yuan X, Je M, Heng C-H (2010) A 5.9mW 50Mbps CMOS QPSK/O-QPSK transmitter employing injection locking for direct modulation. IEEE Asian solid-state circuits conference digest of technical papers, Nov 2010, p 1–2

Chapter 5
Energy Efficient RRAM Crossbar-Based Approximate Computing for Smart Cameras

Yu Wang, Boxun Li, Lixue Xia, Tianqi Tang, and Huazhong Yang

Abstract Smart cameras have been applied successfully in many fields. The limited battery capacity and power efficiency restrict the local processing capacity of smart cameras. In order to shift vision processing closer to the sensors, we propose a power efficient framework for analog approximate computing with the emerging metal-oxide resistive switching random-access memory (RRAM) devices. A programmable RRAM-based approximate computing unit (RRAM-ACU) is introduced first to accelerate approximated computation, and a scalable approximate computing framework is then proposed on top of the RRAM-ACU. In order to program the RRAM-ACU efficiently, we also present a detailed configuration flow, which includes a customized approximator training scheme, an approximator-parameter-to-RRAM-state mapping algorithm, and an RRAM state tuning scheme. Simulation results on a set of diverse benchmarks demonstrate that, compared with an x86-64 CPU at 2 GHz, the RRAM-ACU is able to achieve 4.06–196.41× speedup and power efficiency of 24.59–567.98 GFLOPS/W with quality loss of 8.72 % on average. The implementation of HMAX application further demonstrates that the proposed RRAM-based approximate computing framework can achieve $> 12.8\times$ power efficiency than the digital implementation counterparts (CPU, GPU, and FPGA).

Keywords Approximate computing • Neural network • Power efficiency • Resistive random-access memory (RRAM) • Smart camera

5.1 Introduction

Smart cameras have been applied successfully in many fields including autonomous vehicles, area surveillance, search and rescue by providing application-specific information out of a raw image or video stream without human operators [1]. Power efficiency has become a major concern for smart camera and modern

Y. Wang (✉) • B. Li • L. Xia • T. Tang • H. Yang
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
e-mail: yu-wang@tsinghua.edu.cn

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_5

computing system design [2]. The limited battery capacity restricts the local processing capacity of smart cameras. Current smart cameras have to rely on the cloud computing by uploading the data to servers and then downloading the results, which leads to a privacy concern and disability of supporting real-time applications like automatic driving cars. In order to shift vision processing closer to the sensors and realize the local processing of smart cameras, power efficiency of hundreds of giga floating point operation per second per watt (GFLOPS/W) is expected to achieve the desirable portability and performance [3]. However, the highest power efficiency of contemporary CPU and GPU systems is only $\sim$ 10 GFLOPS/W, which is expected not to substantially improve in the predictable scaled technology node [4, 5]. As a result, researchers are looking for alternative architectures and technologies to achieve further performance and efficiency gains [6].

In recent years, the device technology innovations have enabled new computational paradigms beyond Von-Neumann architectures, which not only provide a promising hardware solution to neuromorphic system but also help drastically improve the power efficiency of computing systems. The metal-oxide resistive switching random-access memory (RRAM) device (or the memristor) is one of the promising innovations that can advance Moore's law beyond the present silicon roadmap horizons [7]. RRAM devices are able to support a large number of signal connections within a small footprint by taking advantage of the ultra-integration density. And more importantly, RRAM devices can be used to build resistive cross-point structure [8], also known as the RRAM crossbar array, which can naturally transfer the weighted combination of input signals to output voltages and realize the matrix-vector multiplication with incredible power efficiency [9, 10].

Our objective is to use the emerging RRAM devices to design a reconfigurable approximate computing framework with both power efficiency and computation generality. Approximate computing provides a promising solution to close the gap of power efficiency between present-day capabilities and future requirements [11]. Approximate computing takes advantage of the characteristic that many modern applications, ranging from signal processing, pattern recognition to computer vision, are able to produce results with acceptable quality even if many computations are executed imprecisely [12]. This tolerance of imprecise computation can be leveraged for substantial performance and efficiency gains, especially for smart cameras [2, 13].

To realize this goal, the following challenges must be overcome: On the one hand, an architecture, from the basic processing unit to a scalable framework, is required to provide an efficient hardware implementation for RRAM-based analog approximate computing. On the other hand, from the perspective of software, a detailed configuration flow is demanded to program the hardware efficiently for each specific application.

In this work, we explore the potential of RRAM-based analog approximate computing. The main contributions of this work include:

- We propose a power efficient RRAM-based approximate computing framework. The scalable framework is integrated with our programmable RRAM-based

approximate computing units (RRAM-ACUs), which work as universal approximators. Simulation results show that the RRAM-ACU offers less than 1.87 % error for 6 common complex functions.
- A configuration flow is proposed to program RRAM-ACUs. The configuration flow includes three phases: (1) a training scheme customized for RRAM-ACU to train its neural approximator; (2) a parameter mapping scheme to convert the parameters of a trained neural approximator to appropriate RRAM resistance states; and (3) a state tuning scheme to tune RRAM devices to target states.
- A set of diverse benchmarks are used to evaluate the performance of RRAM-based approximate computing. Experiment results demonstrate that, compared with an x86-64 CPU at 2 GHz, our RRAM-ACU provides power efficiency of 249.14 GFLOPS/W and speedup of $67.29\times$ with quality loss of 8.72 % on average. And the implementation of HMAX application demonstrates that the proposed RRAM-based approximate computing framework is able to support large scale applications under different noisy conditions, and can achieve $> 12.8\times$ power efficiency improvements than the CPU, GPU, and FPGA implementation counterparts.

The rest of this paper is organized as follows: Sect. 5.2 provides the basic background knowledge. Section 5.3 introduces the details of the proposed RRAM-based approximate computing framework. The configuration flow is depicted in Sect. 5.4. Experimental results of different benchmarks are presented in Sect. 5.5. Finally, Sect. 5.6 concludes this work.

5.2 Preliminaries

5.2.1 RRAM Characteristics and Device Model

The RRAM device is a passive two-port elements based on TiO_x, WO_x, HfO_x [14], or other materials with variable resistance states. The most attractive feature of RRAM devices is that they can be used to build *resistive cross-point structure*, which is also known as the RRAM crossbar array. Compared with other non-volatile memories like flash, the RRAM crossbar array can naturally transfer the weighted combination of input signals to output voltages and realize the matrix-vector multiplication efficiently by reducing the computation complexity from $O(n^2)$ to $O(1)$. And the continuous variable resistance states of RRAM devices enable a wide range of matrices that can be represented by the crossbar. These unique proprieties make RRAM devices and the RRAM crossbar array promising tools to realize analog computing with great efficiency.

Figure 5.1a demonstrates a model of the HfO_x-based RRAM device[15]. The structure is a resistive switching layer sandwiched between two electrodes. The conductance is exponentially dependent on the tunneling gap distance (d) as:

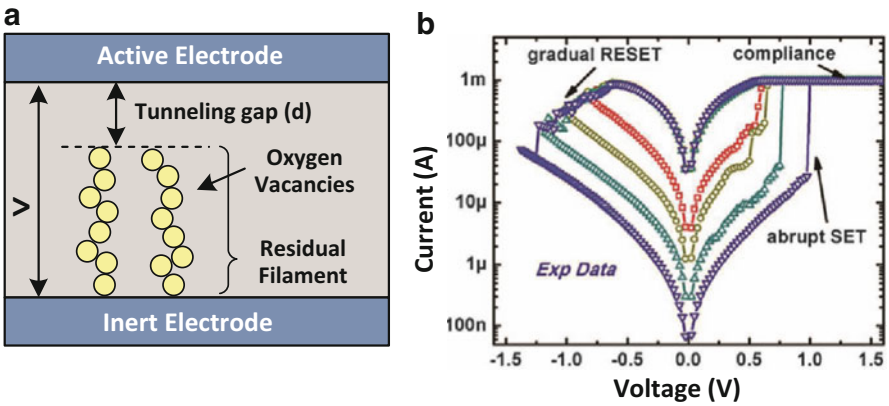


Fig. 5.1 (**a**) Physical model of the HfO_x-based RRAM. The RRAM resistance state is determined by the tunneling gap distance d, and d will evolve due to the field and thermally driven oxygen ion migration. (**b**) Typical DC I–V bipolar switching curves of HfO_x RRAM devices reported in [15]

$$I = I_0 \cdot \exp\left(-\frac{d}{d_0}\right) \cdot \sinh\left(\frac{V}{V_0}\right) \tag{5.1}$$

The ideal resistive crossbar-based analog computing requires both linear I–V relationship and continuous variable resistance states. However, nowadays RRAM devices can't satisfy these requirements perfectly. Therefore, we introduce the practical characteristics of RRAM devices in this section:

- The I–V relationship of RRAM devices is *non-linear*. However, when V is very small, an approximation can be applied as $\sinh(\frac{V}{V_0}) \sim \frac{V}{V_0}$. Therefore, the voltages applied on RRAM devices should be limited to achieve an approximate linear I–V relationship [16].
- As shown in Fig. 5.1b, the SET process (from a high resistance state (HRS) to a low resistance state (LRS)) is abrupt while the RESET process (the opposite switching event from LRS to HRS) is gradual. The RESET process is usually used to achieve multiple resistance states [17].
- Even in the RESET process, the RRAM resistance change is *stochastic and abrupt*. This phenomenon is called '*variability*'. The RRAM variability can be approximated as a lognormal distribution and can make the RRAM device miss the target state in the switching process.

In this paper, we use the HfO_x-based RRAM device for study because it is one of the most mature materials explored [14]. The analytical model is put into the circuit with Verilog-A [15, 18]. We use HSPICE to simulate the circuit performance and study the device and circuit interaction issues for RRAM-based approximate computing.

5.2.2 Neural Approximator

Figure 5.2 illustrates a simple model of a 3-layer feedforward artificial neural network with one hidden layer. The computation between neighbour layers of the network can be expressed as:

$$y_j = f_j\left(\sum_{i=1}^{n} w_{ij} \cdot x_i + b_j\right) \tag{5.2}$$

where x_i is the value of node i in the input (hidden) layer, and y_j represents the result of node j in the hidden (output) layer. w_{ij} is the connection weight between x_i and y_j. b_j is an offset. $f_j(x)$ is an activation function, e.g. sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.3}$$

It has been proven that a universal approximator can be implemented by a 3-layer feedforward network with one hidden layer and sigmoid activation function [19, 20]. Table 5.1 gives the maximum errors of the approximations of six common functions by this method based on the MATLAB simulation. The mean square errors (MSE) of approximations are less than 10^{-6} after the network training algorithm completes. The neural approximator offers less than 1.87 % error for the 6 common complex functions. This precision level is able to satisfy the requirements of many approximate computing applications [2].

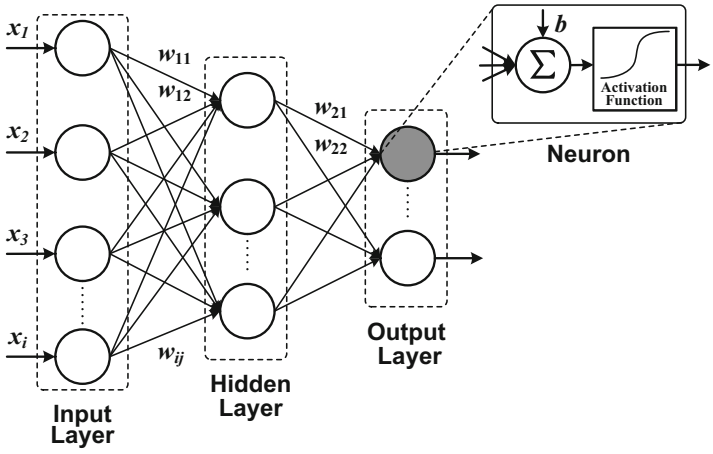


Fig. 5.2 A 3-layer feedforward neural network with one hidden layer

Table 5.1 Maximum errors (%) of neural approximators

Function	Nodes in the hidden layer					
	0	5	10	15	20	25
$x_1 \cdot x_2 \cdot x_3$	22.79	1.10	0.68	0.28	0.34	0.27
x^{-1}	9.53	0.25	0.20	0.14	0.10	0.05
$\sin(x)$	10.9	0.05	0.07	0.05	0.07	0.06
$\log(x)$	7.89	0.21	0.13	0.14	0.12	0.14
$\exp(-x^2)$	20.27	0.04	0.03	0.05	0.03	0.04
$\sqrt{x}$	13.76	1.87	1.19	1.43	0.35	0.49

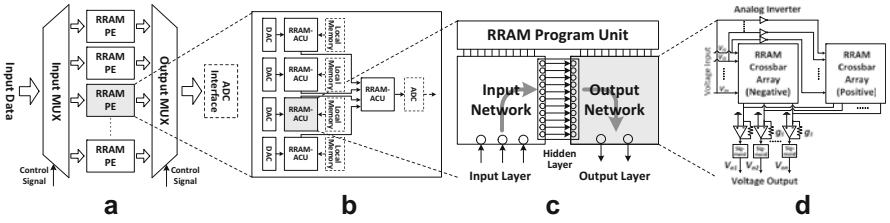


Fig. 5.3 The overview of the hardware architecture of RRAM-based analog approximate computing. (**a**) & (**b**). RRAM approximate computing framework. (**c**) & (**d**). RRAM-based approximate computing unit (RRAM-ACU)

5.3 RRAM-Based Analog Approximate Computing

Figure 5.3 demonstrates an overview of the hardware implementation of RRAM-based analog approximate computing. In this section, we will introduce this framework from the basic RRAM-based approximate computing unit (RRAM-ACU) to the scalable RRAM-based approximate computing framework.

5.3.1 RRAM-Based Approximate Computing Unit

Figure 5.3c,d shows the proposed RRAM-based approximate computing unit (RRAM-ACU). The RRAM-ACU is based on an RRAM hardware implementation of a 3-layer network (with one hidden layer) to work as a universal approximator. The mechanism is as follows.

As described in Eqs. (5.2)–(5.3), the neural approximator can be conceptually expressed as: (1) a matrix-vector multiplication between the network weights and input variations; and (2) a sigmoid activation function.

For the matrix-vector multiplication, this basic operation can be mapped to the RRAM crossbar array illustrated in Fig. 5.4. The output of the crossbar array can be expressed as:

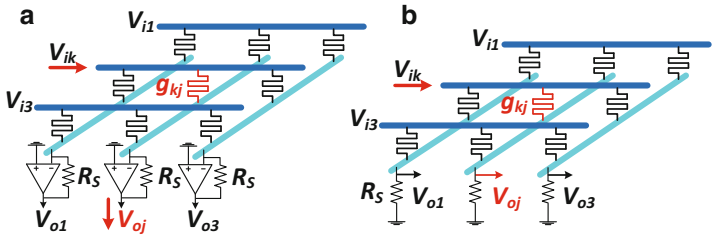


Fig. 5.4 Two implementations of RRAM crossbar arrays: (**a**) with and (**b**) without Op Amps. We use the first implementation in this work

$$V_{oj} = \sum_{k} V_{ik} \cdot c_{kj} \tag{5.4}$$

where, for Fig. 5.4a, c_{kj} can be represented as:

$$c_{kj} = -\frac{g_{kj}}{g_s} \tag{5.5}$$

and for Fig. 5.4b:

$$c_{kj} = \frac{g_{kj}}{g_s + \sum_{l=1}^{N} g_{kl}} \tag{5.6}$$

where g_{kj} is the RRAM conductance state in the crossbar array. And g_s represents the conductivity of the load resistance.

Both two types of crossbar array are efficient to realize matrix-vector multiplication by reducing the computation complexity from $O(n^2)$ to $O(1)$.

The latter one, which does not require Op Amps, consumes less power and can be smaller in size. However, there are some drawbacks with the latter implementation when building multilayer networks:

- First of all, c_{kj} not only depends on the corresponding g_{kj}, but also depends on all the RRAM devices in the same column. It's difficult to realize a linear one-to-one mapping between the network weight w_{ij} and the RRAM conductance g_{ij}. Although previous work proposed some approximate mapping algorithms, the computation accuracy is still a problem [21].
- Secondly, the parameters of neighbour layers will influence each other through R_S. Voltage followers or buffer amplifiers are demanded to isolate different circuit stages and guarantee the driving force [22, 23]. The size and energy savings compared with the first type implementation will be wasted.

The first implementation can overcome these drawbacks. Op Amps can enhance the output accuracy, make c_{kj} linearly depend on the corresponding g_{kj}, and isolate neighbour layers. So we choose the first implementation to build RRAM-ACU.

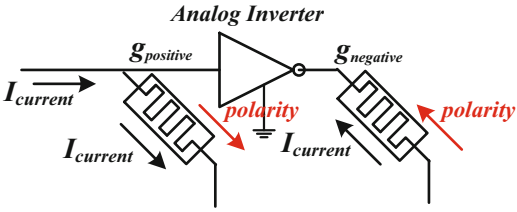


Fig. 5.5 RRAM pairing technique

Since both R (the load resistance) and g (the conductance states of RRAM devices) can only be positive, two crossbar arrays are needed to represent the positive and negative weights of a neural approximator, respectively, with the help of analog inverters [24] as shown in Fig. 5.5.

The practical weights of the network can be expressed as:

$$w_{kj} = R \cdot (g_{kj_{\text{positive}}} - g_{kj_{\text{negative}}}) \tag{5.7}$$

We also note that the polarities of the terminals of the RRAM devices in two crossbar arrays should be set to opposite directions. This technique is aimed to make the resistance state deviations caused by the currents passing through the paired RRAM devices cancel each other [25]. We refer to this technique as RRAM pairing and it's shown in Fig. 5.5.

The sigmoid activation function can be generated by the circuit described in [26] and a complete feedforward network without hidden layer is accomplished.

Finally, by combining two networks together, a three-layer feedforward network unit is realized. As described in Sect. 5.2.2, this network can work as a universal approximator to perform approximated computation. And a basic RRAM approximate computing unit is accomplished.

5.3.2 RRAM-Based Approximate Computing Framework

The overview of the proposed RRAM approximate computing framework is shown in Fig. 5.3a,b. The building blocks of the framework are the RRAM processing elements (RRAM PE). Each RRAM PE consists of several RRAM-ACUs to accomplish algebraic calculus. Each RRAM PE is also equipped with its own digital-to-analog converters (DACs) to generate analog signals for processing. In addition, the RRAM PE may also have several local memories, e.g., analog data stored in form of the resistance states of RRAM devices, or digital data stored in the DRAM or SRAM. Both the use and the type of local memory depend on the application requirement and we will not limit and discuss its implementation in detail in this work. On top of that, all the RRAM PEs are organized by two multiplexers with Round-Robin algorithm.

In the processing stage, the data will be injected into the platform sequentially. The input multiplexers will deliver the data into the relevant RRAM PE to perform approximate computing. The data will be fed into the RRAM PE in digital format and the DACs in each RRAM PE will convert the date into analog signals. Each RRAM PE may work under low frequency but a group of RRAM PEs can work in parallel to achieve high performance. Finally, the output data will be transmitted out from the RRAM PE by output multiplexer for further processing, e.g., be converted back into digital format by a high performance analog-to-digital converter (ADC).

The framework is scalable and the user can configure it according to individual demand. For example, for tasks requiring power efficiency, it's better to choose low power Op Amps to form the RRAM-ACUs and each RRAM PE may work in a low frequency. On the other hand, high speed Op Amps, AD/DAs, and even hierarchical task allocation architecture will be preferred for high performance applications.

5.4 Configuration Flow for RRAM-ACU

The RRAM-based analog approximate computing hardware requires a configuration flow to get programmed for each specific task. In this section, we discuss the detailed configuration flow for the proposed RRAM-based approximate computing units (RRAM-ACUs). The flow is illustrated in Fig. 5.6. It includes three phases to solve the following problems:

1. Training Phase: How to train a neural approximator in an RRAM-ACU to learn the required approximate computing task?
2. Mapping Phase: The parameters of a trained approximator can NOT be directly configured to the RRAM-ACU. We need to map these parameters to appropriate RRAM resistance states in the RRAM crossbar array.
3. Tuning Phase: After we achieve a set of RRAM resistance states for an approximate computing task, how to tune the RRAM devices accurately and efficiently to the target states?

All these phases will be introduced in detail in the following sections.

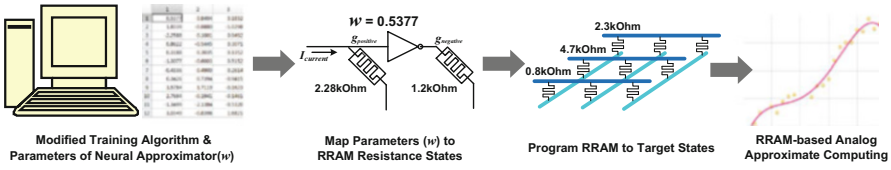


Fig. 5.6 Configuration flow for RRAM-ACU. The flow includes three phases: (1) a training scheme customized for RRAM-ACU to train the neural approximator; (2) a parameter mapping scheme to convert the parameters of a trained neural approximator to appropriate RRAM resistance states; and (3) an RRAM state tuning scheme to tune RRAM devices to target states efficiently

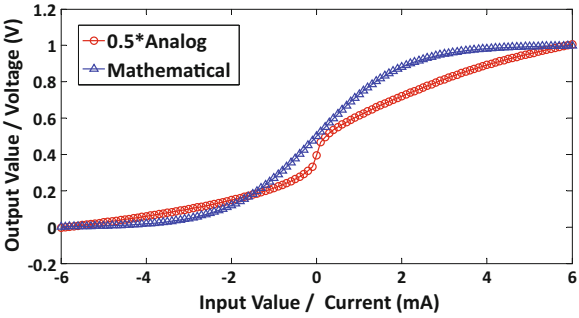


Fig. 5.7 Comparison between the mathematical sigmoid function and its analog implementation reported in [26]. The output of analog implementation is multiplied by 0.5 for normalization. A significant difference can be observed

5.4.1 *Training Phase: Neural Approximator Training Algorithm*

The RRAM approximate computing unit is based on an RRAM implementation of neural approximator. The approximator must be trained efficiently for each specific function. The training process can be realized by adjusting the weights in the network layer by layer [27]. The update of each weight (w_{ji}) can be expressed as:

$$w_{ji} \leftarrow w_{ji} + \eta \cdot \delta_j \cdot x_i \tag{5.8}$$

where x_i is the value of node i. η is the learning rate. δ_j is the error back propagated from node j in the next neighbour layer. δ_j depends on the derivative of the activation function (e.g. sigmoid function) as described in Sect. 5.2.2.

In the RRAM-ACU training phase, both the calculations of sigmoid function and its derivative should be adjusted according to the analog sigmoid circuit. Figure 5.7 illustrates a comparison between the accurate mathematical sigmoid function and its hardware implementation reported in [26]. The I–V relationship is simulated with HSPICE. There is a significant difference between them. Therefore, we replace the mathematical sigmoid activation function by its simulation results in the training scheme of RRAM-ACU.

Finally. it's worth noting that most weights are small (around zero) after a proper training.[1] For example, more than 90 % weights of the trained network[2] are within

[1] A neural network will tend to overfit when many weights of the network are large [28]. Overfitting is a problem that the model learns too much, including the noise, from the training data. The trained model will have poor predictive performance on the unknown testing data which are not covered by the training set.

[2] We use ℓ_2 regularization in the training scheme. Regularization is a technique widely used in the neural network training to limit the amplitude of network weight, avoid overfitting, and improve model generalization [28]. To be specific, for the ℓ_2 regularization, a penalty of the square of the 2-norm of network weights will be proportionally added to the loss function of the network. So the error of the network and the amplitude of weights will be balanced and optimized simultaneously in the training process [28].

the range of $[-1.5, 1.5]$ for all the benchmarks used in this paper. The limitation of weight amplitude can simplify the design of RRAM state tuning scheme and help improve the tuning speed.

5.4.2 Mapping Phase: Mapping Neural Approximator Weights to RRAM Conductance States

Once the weights of a neural approximator are determined, the parameters need to be mapped to the appropriate states of RRAM devices in the crossbar arrays. Improperly converting the network weights to the RRAM conductance states may result in the following problems:

- The converted results are beyond the actual range of the RRAM device.
- The dynamic range of converted results is so small that the RRAM state may easily saturate.
- The converted results are so high that the summation of output voltages will exceed the working range of Op Amps.

In order to prevent the above problems, we propose a parameter mapping algorithm to convert the weights of neural approximator to appropriate conductance states of RRAM devices.

The mapping process can be abstracted as an optimization problem. The feasible range of the weights of neural approximators can be expressed as a function of RRAM parameters:

$$-R_S \cdot (g_{ON} - g_{OFF}) \leq w \leq R_S \cdot (g_{ON} - g_{OFF}) \tag{5.9}$$

where $g_{ON} = R_{ON}^{-1}$ and $g_{OFF} = R_{OFF}^{-1}$. R_{ON} and R_{OFF} are the lowest and highest resistance states of RRAM devices. All the weights should be scaling within this range.

In order to extend the dynamic range and reduce the impact of process variation, we adjust g_{ON} and g_{OFF} to:

$$g_{ON}^{'} = \frac{1}{\eta \cdot \Delta_{ON} + R_{ON}} \tag{5.10}$$

$$g_{OFF}^{'} = \frac{1}{R_{OFF} - \eta \cdot \Delta_{OFF}} \tag{5.11}$$

where Δ_{ON} and Δ_{OFF} represent the maximum deviation of R_{ON} and R_{OFF} induced by process variation of the crossbar array, respectively. η is a scale coefficient which is set to 1.1–1.5 in our design to achieve a safety margin.

The risk of improper conversion can be measured by the following risk function:

$$Risk(g_{pos}, g_{neg}) = |g_{pos} - g'_{mid}| + |g_{neg} - g'_{mid}| \tag{5.12}$$

where:

$$g'_{mid} = \frac{g'_{ON} + g'_{OFF}}{2} \tag{5.13}$$

and g_{pos} and g_{neg} represent the conductance states of each paired RRAM devices in the positive and negative crossbar arrays, respectively, as Eq. (5.7).

Combining the constraints and the risk function, the parameter mapping problem can be described as the optimization problem shown below:

$$(g^*_{pos}, g^*_{neg}) = \arg\min \text{Risk} \tag{5.14}$$

$$s.t. \begin{cases} R_S \cdot (g^*_{pos} - g^*_{neg}) = w \\ g'_{ON} \leq g_{pos} \leq g'_{OFF} \\ g'_{ON} \leq g_{neg} \leq g'_{OFF} \end{cases} \tag{5.15}$$

The optimal solutions of this optimization problem are

$$\begin{cases} g^*_{pos} = g'_{mid} + \dfrac{w}{2R_S} \\ g^*_{neg} = g'_{mid} - \dfrac{w}{2R_S} \end{cases} \tag{5.16}$$

These are the appropriate conductance states of RRAM devices with the minimum risk of improper parameter conversion.

5.4.3 *Tuning Phase: Tuning RRAM Devices to Target States*

After the weights of neural approximator are converted into RRAM conductance states, a state tuning scheme is required to program RRAM devices in an RRAM-ACU to target states.

Due to the stochastic characteristics of RRAM resistance change, program-and-verify (P&V) method is commonly used in multi-level state tuning [29]. As shown in Fig. 5.8, the RRAM device will be first initialized to LRS. Then a sequence of write pulses will be applied to tune RRAM devices gradually. Each write pulse is followed by a read pulse to verify the current conductance state. The amplitude of read pulse should be small enough to not change the RRAM conductance state.

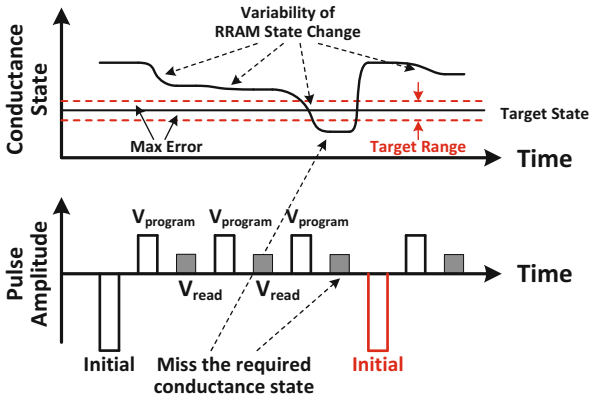


Fig. 5.8 Program-and-verify (P&V) scheme for multi-level RRAM conductance state tuning

The P&V operation will keep on performing until the verify step detects that the RRAM device has reached the target range.

The P&V method choose LRS as the initial state because of the following reasons:

- LRS is much more uniform than HRS. When an RRAM device is switched between HRS and LRS repeatedly, LRS is able to be uniform while HRS usually varies a lot among different cycles [14, 15, 30];
- As shown in Fig. 5.1b, the resistance change process from LRS to HRS is gradual, while the opposite process is abrupt. It is easier to achieve multiple resistance states from LRS than HRS, although HRS may help reduce the power consumption.
- Finally, the target resistance states are closer to LRS according to Eq. (5.16). As HRS is usually $> 100\times$ larger than LRS, initializing RRAM devices to LRS will require much less pulses to reach the target resistance range.

However, tuning RRAM devices to accurate g'_{mid}, g^*_{pos}, or g^*_{neg} as Eq. (5.16) still requires large effort with P&V method. Considering the physical characteristics of RRAM devices and the circuit architecture of RRAM-ACU, we propose a simple but efficient RRAM state tuning scheme as illustrated in Fig. 5.9. The proposed RRAM state tuning scheme includes the following two steps:

Step 1: Initializing all the RRAM devices in the paired crossbar arrays to the same initial state g_i. We hope that only one RRAM device in the pair needs tuning after we initialize all the RRAM devices to g_i. The choice of g_i is a major concern in this state tuning scheme. It should be able to approximate most of the optimal states $\left(g'_{mid} + \frac{|w|}{2R_S}\right)$ in the crossbar array, and should be both uniform and easy to reach for RRAM devices. Therefore, we choose g_i to be close to g'_{mid} because most w_{ij} are close to zero as discussed in Sect. 5.4.1 and the optimal states $(g'_{mid} + \frac{|w|}{2R_S})$ will be close to g'_{mid}. On top of that, we

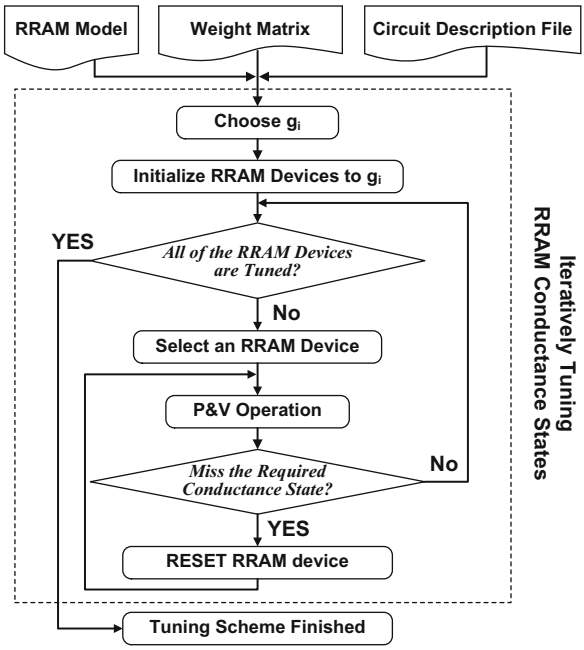


Fig. 5.9 Proposed state tuning scheme for RRAM-ACU

choose g_i, which should be a uniform low resistance state that can be achieved easily according to the physical characteristics of RRAM devices. For example, for the HfO_x RRAM devices used in this paper, the lowest resistance state is $R_{ON} \approx 290\Omega$ [15]. And we set g_i to $\sim 500\Omega^{-1}$ as it's both close to $g_{ON}/2$ and can be easily achieved by limiting the compliance current [15].

Step 2: Tuning the positive and negative crossbar arrays to satisfy $R_S \cdot (g_{pos} - g_{neg}) = w$. After initializing RRAM devices to $g_i \approx g'_{mid}$, only one RRAM device in each paired RRAM devices will need to be tuned according to Eq. (5.16). The state tuning scheme will perform P&V operations on the corresponding RRAM device until Eq. (5.15) is satisfied.

Another problem of the state tuning scheme is that the variability of resistance state change may make RRAM devices miss the target conductance range. Considering that the set back process is abrupt and hard to control, and most target states that are close to g_i (e.g. the requirement of resistance change is usually around tens of ohms), in this work, the proposed state tuning scheme will reset the RRAM device to the initial state g_i. There is no need to prepare a complicated partial setback operation at the cost of increasing the circuit complexity.

The last problem in the state tuning scheme is the sneak path problem. Sneak path usually exits in the memory architecture. As only one cell will be selected in the memory read or write operation, it will be difficult for the architecture to isolate

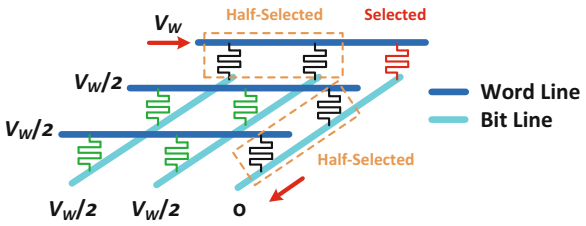


Fig. 5.10 Tuning RRAM devices with half-select method to mitigate sneak path problem

the selected RRAM device from the unselected cells. The unselected cells will form a sneak path, which will disturb the output signals and impact the unselected cells' states [31]. However, when an RRAM crossbar array is used for computation, all the cells will be selected for computation. In other words, no sneak path can be formed in this case. By contrast, each output port can only be used to tune one RRAM device in the corresponding column. We cannot select and tune all the RRAM devices in the crossbar array at the same time. The sneak path still exists in the state tuning scheme.

In order to mitigate the impact of sneak path in the state tuning scheme, the half-select method is adopted [8]. Figure 5.10 illustrates the principle of half-select method. The method is aimed to reduce the voltage drop between the selected and unselected cells to reduce the sneak path current and its impact. A half-select voltage ($V_W/2$), instead of connecting to the ground, will be applied on the unselected word line and bit line. The maximum voltage drop between the selected and unselected cells is $V_W/2$ instead of V_W. Therefore, the sneak path current is reduced and the unselected cells are protected.

The half-select method mitigate the sneak path problem at the cost of extra power consumption. We further reduce the direct component in the original half-select method to alleviate this problem. To be specific, a ($V_W/2$) and ($-V_W/2$) voltage will be applied on the WL and BL of the selected cell, respectively. And other unselected cells will be connect to the ground instead of a half-select voltage ($V_W/2$). This technique can reduce around 75 % of the power consumption compared with the original method.

Finally, we note that only the RRAM devices in different word lines and bit lines can be tuned in parallel. A parallel state tuning scheme can significantly improve the tuning speed of RRAM-ACU but will require extra copies of peripheral circuits and additional control logic. As the energy consumption (the product of tuning time and power consumption) of tuning the entire RRAM crossbar array remains almost the same, there will be a trade-off between the tuning speed and the circuit size in the RRAM state tuning scheme. In order to save more area for AD/DAs and Op Amps, each RRAM-ACU is equipped with only one set of tuning circuit in this work.

5.5 Evaluation

To evaluate the performance and efficiency of the proposed RRAM-based analog approximate computing, we apply our design to several benchmarks, ranging from the signal processing, gaming, and compression to the object recognition. A sensitivity analysis is also performed to evaluate the robustness of the RRAM-based computing system.

5.5.1 Experiment Setup

In the experiment, a Verilog-A RRAM device model reported in [15, 18] is used to build up the SPICE-level crossbar array. We choose the 65 nm technology node to model the interconnection of the crossbar array and reduce the IR drop. The parameters of the interconnection are calculated with the International Technology Roadmap for Semiconductors 2013 [32]. The sigmoid circuit is the same as reported in [26]. The Op Amps, ADCs, and DACs used for simulation are that reported in [33, 34] and [35], respectively. The working frequency of each RRAM-ACU is set to 800 MHz. Detailed parameters of peripheral circuits are summarized in Table 5.2. Moreover, the maximum amplitude of input voltage is set to 0. 5 V to achieve an approximate linear I–V relationship of RRAM devices. All the simulation results of the RRAM crossbar array are achieved with HSPICE.

5.5.2 Benchmark Evaluation

Table 5.3 summarizes the benchmarks used in the evaluation. The benchmarks are the same as that described in [2], which are used to test the performance of an ×86-64 CPU at 2 GHz equipped with a CMOS-based digital neural processing unit. The 'NN Topology' term in the table represents the size of each neural network. For example, '9 × 8 × 1' represents a neural approximator with 9 nodes in the input layer, 8 nodes in the hidden layer, and 1 node in the output layer. The MSE is tested both on CPU and SPICE-based RRAM-ACU after training. The training scheme

Table 5.2 Detailed parameters of peripheral circuits in RRAM-ACU

Technology node	180 nm
RRAM tunneling gap	0.2–1.9 nm
RRAM resistance range	290Ω–500 kΩ
R_S	2 kΩ
Op amp	∼ 4.8 mW
ADC	8bit, ∼ 3.1 mW
DAC	12bit, ∼ 40 mW
Frequency	800 MHz

Table 5.3 Benchmark description

Name	Description	Type	x86-64 Insts	Training set	Testing set	NN topology	NN MSE (CPU)	NN MSE (RRAM)	Error metric	Error (%)
FFT	Radix-2 Cooley–Tukey fast fourier	Signal processing	34 floating point numbers	32,768 random floating point numbers	2,048 random	1 × 8 × 2	0.0046	0.0071	Average relative error	10.72
Inversek2j kinematics for 2-Joint arm	Inverse	Robotics	100	10,000 (x, y) random coordinates	10,000 (x, y) random coordinates	2 × 8 × 2	0.0038	0.0053	Average relative error	9.07
Jmeint	Triangle intersection detection	3D gaming	1,079	10,000 pairs of 3D triangle coordinates	10,000 pairs of 3D triangle coordinates	18 × 48 × 2	0.0117	0.0258	Miss rate	9.50
JPEG	JPEG Encoding	Compression	1,257	Three 512 × 512 color images	One 220 × 220 color images	64 × 16 × 64	0.0081	0.0153	Image diff	11.44
K-Means	K-Means clustering	Machine learning	26	50,000 pairs of (R,G,B) values	One 220 × 220 color image	6 × 20 × 1	0.0052	0.0081	Image diff	7.59
Sobel	Sobel edge Detector	Image processing	88	One 512 × 512 color image	One 220 × 220 color image	9 × 8 × 1	0.0286	0.0026	Image diff	4.00

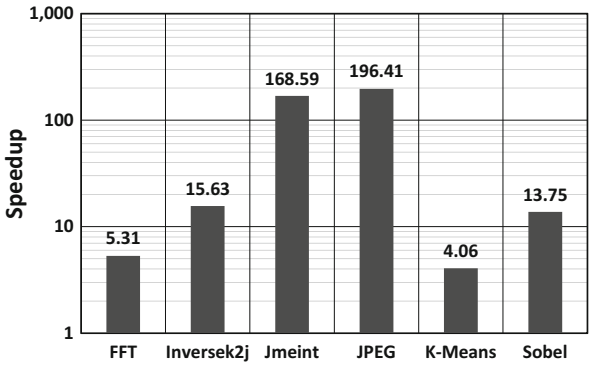


Fig. 5.11 Speedup of the RRAM-ACU under different benchmarks

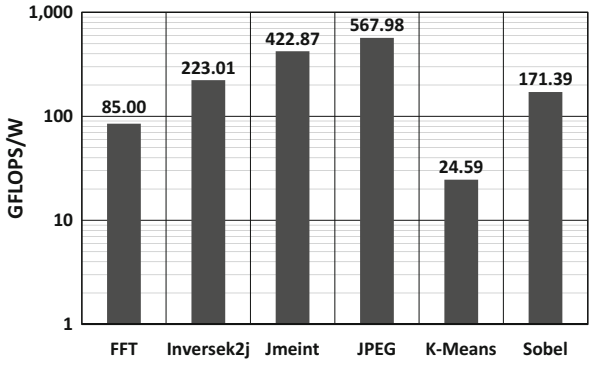


Fig. 5.12 Power efficiency of the RRAM-ACU under different benchmarks

has been described in Sect. 5.4.1, which is modified for RRAM-ACU. The size of the crossbar array in the RRAM-ACU is set to 64 × 64 to satisfy all the benchmarks. The unused RRAM devices in the crossbar array are set to the highest resistance states to minimize the sneak path problem. And the unused input and output ports are connected to the ground.

The simulation results are illustrated in Figs. 5.11 and 5.12. Compared with the x86-64 CPU at 2 GHz, the RRAM-ACU achieves 567.98 GFLOPS/W power efficiency and 196.41× speedup at most. And for the whole set of selected diverse benchmarks, the RRAM-ACU provides 249.14 GFLOPS/W and speedup of 67.29× with quality loss of 8.72 % on average. The improvement of processing speed mainly depends on the capability of the neural approximator. As **the RRAM-ACU is able to transfer a set of instructions into a neural approximator and execute them with only one cycle**, the speedup achieved by an RRAM-ACU increases linearly with the number of instructions the neural approximator represents. For example, the 'Jmeint' and 'JPEG' benchmarks achieve $> 150\times$ speedup as their neural approximators successfully implement the complex tasks that require more than a thousand instructions in traditional x86-64 architectures. In contrast, the 'K-Means' and 'FFT' benchmarks achieve the least speedup ($\sim 10\times$) because of the simplicity of tasks. And for the improvement of power efficiency, although the

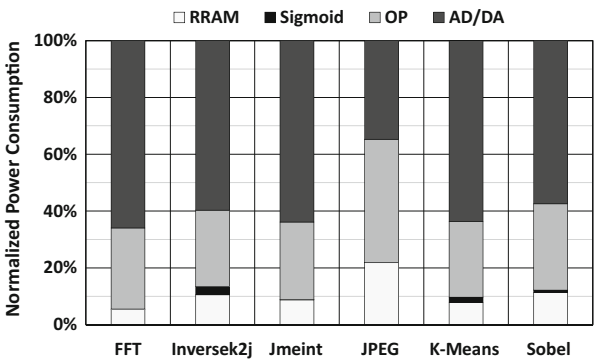


Fig. 5.13 RRAM-ACU power consumption breakdowns

RRAM-ACU for a complex task is able to achieve more speedups, a bigger neural approximator may also be demanded to accomplish more power-consuming tasks. However, as the NN topology increases slower than the instruction number in the experiment, the complex tasks still achieve better power efficiency.

Figure 5.13 illustrates the power consumption breakdowns of RRAM-ACUs. The sigmoid circuit is power efficient as there are only 6 MOSFETs used in the circuit [26]. The power consumption of sigmoid circuit mainly depends on the output voltage. For example, most outputs will be close to zero after the JPEG encoding. And therefore, the sigmoid circuit takes a negligible part of power consumption in the 'JPEG' benchmark. In contrast, the outputs of sigmoid circuits in the 'Inversek2j' and 'K-Means' are much larger and the power consumption increases as a result. Compared with the sigmoid circuit, most of the power is consumed by Op Amps and AD/DAs. RRAM devices only take 10–20 % of the total energy consumption in RRAM-ACU, and the ratio increases with the NN topology. Therefore, how to reduce the energy consumed by peripheral circuits may be a challenge to further improve the efficiency of RRAM-based analog approximate computing.

In conclusion, the simulation results demonstrate the efficiency of RRAM-ACU as well as the feasibility of a dynamic reconfiguration.

5.5.3 System Level Evaluation: HMAX

In order to evaluate the performance of RRAM-ACU at system level, we conduct a case study on HMAX application. HMAX is a famous bio-inspired model for general object recognition in complex environment[36]. Figure 5.14 demonstrates the framework of HMAX. The model consumes more than 95 % amount of computation to perform pattern matching in S2 layer by calculating the distance between the prototypes and units [7, 36]. The amount of computation is too huge to realize real-time video processing on conventional CPUs while the computation accuracy requirement is not strict [37]. In this section evaluation, we apply the

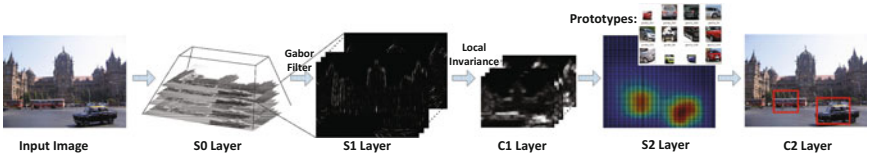


Fig. 5.14 Overview of HMAX framework for object recognition

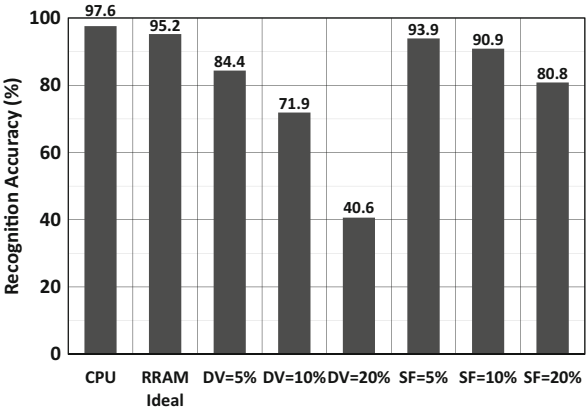


Fig. 5.15 Performance of RRAM-based HMAX under different noise conditions, where 'DV' represents device variation and 'SF' represents input signal fluctuation

proposed RRAM-based approximate computing framework to conduct the distance calculations to promote the data processing efficiency.

We use 1,000 images (350 of cars and 650 of the other categories) from PASCAL Challenge 2011 database [38] to evaluate the performance of the HMAX system on the digital and the RRAM-based approximated computation framework. Each image is of 320 × 240 pixels with complex background. The HMAX model contains 500 patterns of car images which remain the same on each platform. A correct result indicates both a right judgment on the classification of the object and a successful detection on the object location.

The RRAM approximate commuting framework illustrated in Fig. 5.3 is used to support the HMAX approximate. Each RRAM processing element consists of four 6-input RRAM-ACU for Gaussian calculations and one for 4-input multiplication. Therefore, each RRAM PE can realize a 24-input distance calculation per clock cycle [7].

The results of correct rate are shown in Fig. 5.15. The performance of RRAM-based approximate computing under different noise conditions is also considered. The device variation represents the deviation of the RRAM conductance state and the signal fluctuation represents the deviation of the input signals. As we can observe, the correct rate degradation is only 2.4 % on the ideal RRAM-based approximate computing w.r.t. the CPU platform. This degradation can be easily compensated by increasing the amount of patterns [36].

Table 5.4 Power efficiency of the RRAM-based HMAX

AD/DA (mW)	Analog (mW)	Total (mW)	x86-64 Insts	Frequency (MHz)	Efficiency (GFLOPS/W)
963.1	511.96	1475.06	558	800	302.64

Table 5.5 Power efficiency comparison with different platforms (FPGA, GPU, and CPUs in [37])

Parameters	Proposed	FPGA	GPU	CPUs
Size of input image	320 × 240	256 × 256		
HMAX orientations	12			
HMAX scale	12			
HMAX prototypes	500	5120		
Average size of prototypes	8			
Cycles for calculation	32	–		
Calculation amount/frame	5455 × 500	–		
Frequency (MHz)	800	–		
Power (W)	1.475	40	144	116
Unified fps/W	6.214	0.483	0.091	0.023
Speed Up	–	12.86	68.29	270.17

Moreover, when taking the noise into consideration, the device variation will significantly impact the recognition accuracy. As the performance of neural approximator mainly depends on the RRAM conductance states, the device variation will significantly impact the computation quality and make the recognition accuracy decrease a lot. For example, a 10 % device variation can result in a > 50 % decrease of the recognition accuracy. Therefore, the device variation should be suppressed to satisfy the application requiring high accuracy. Compared with the device variation, the impact of signal fluctuation is much less, which demonstrates that we may use DACs with less precision but less power consumption, in the RRAM-ACU to further improve the power efficiency of the whole system.

The power efficiency evaluation of the RRAM-based HMAX accelerator is given in Table 5.4. The detailed comparisons with other platforms are given in Table 5.5. The parameters of the HMAX model as well as the evaluation image dataset are different among different platforms. It's hard to compare the recognition accuracy of different implementations. However, we can still compare the efficiency of different platforms through the unified power consumption per frame. The simulation results show that the power efficiency of RRAM-based approximated computation framework is higher than 300 GFLOPS/W. And compared to other platforms like FPGA, GPU, and CPU [37], RRAM-based HMAX achieves a performance up to 6.214 fps/W, which is 12.8–270.2× higher than its digital counterparts.

5.6 Conclusion and Discussion

In this work, we propose a power efficient approximate computing framework with the emerging RRAM technology to shift vision processing closer to the smart camera sensors. We first introduce an RRAM-based approximate computing framework by integrating our programmable RRAM-ACU. We also introduce a complete configuration flow to program the RRAM-based computing hardware efficiently.

The RRAM-based neuromorphic computing still faces many challenges at different levels.

First, at the application level, which algorithm should be used to support real-world applications with acceptable performance? In this work, we realize an artificial neural network for approximate computing. By contrast, researchers have also developed RRAM-based spiking neural networks (SNN) in time domain [39, 40]. By encoding and processing information with bionic spikes, RRAM SNN avoids high cost AD/DA and enables even lower power consumption. Besides, sparse coding [41], PUF [42], Boltzmann machine [43], and many other applications have also been proposed in recent years. A detailed analysis and comparison of ANN, SNN, and other algorithms are expected to extend the application scenarios of RRAM neuromorphic computing.

Second, at the architecture level, an architecture framework with both scalability and flexibility is demanded to support a wide range of applications. Besides RRAM-based approximate computing [44], researchers also proposed many frameworks like a hybrid memristor crossbar-array/CMOS system [45] and RENO [46]. To evaluate those frameworks, simulators [47] are needed. We also need compilers and software support to map different algorithms and codes to the RRAM hardware [21]. At the same time, as discussed in Sect. 5.2, AD/DAs contribute to a large portion of the area and power consumption of RRAM-based approximate computing, which significantly limits the potential efficiency gains of RRAM crossbar-based computing system. Techniques are required to reduce the interface overhead while maintain good compatibility with von-Neumann systems [48].

Finally, at the circuit level, many important non-ideal factors may significantly influence the performance of RRAM crossbar-based computing systems [49, 50]. For example, the IR-drop caused by the interconnect resistance influences the RRAM computation quality and severely limits the scale of the crossbar system [51]. EDA tools, which are able to consider and mitigate different non-ideal factors (e.g. non-linear devices, IR-drop), are demanded to help the circuit design and fabrication for RRAM neuromorphic computing [52].

Acknowledgements This work was supported by 973 Project 2013CB329000, National Natural Science Foundation of China (No. 61373026), Brain Inspired Computing Research, Tsinghua University (20141080934), Tsinghua University Initiative Scientific Research Program, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

References

1. Graf R, Belbachir A, King R, Mayerhofer M (2013) Quality control of real-time panoramic views from the smart camera 360 scan. In: 2013 I.E. international symposium on circuits and systems (ISCAS), pp.650–653
2. Esmaeilzadeh H, Sampson A, Ceze L, Burger D (2012) Neural acceleration for general-purpose approximate programs. In: International symposium on microarchitecture(MICRO), pp 449–460
3. DARPA (2012) Power efficiency revolution for embedded computing technologies [Online]. Available: https://www.fbo.gov/
4. NVIDIA Tesla K-Series, DATASHEET (2012) Kepler family product overview [Online]. Available: http://www.nvidia.com/content/tesla/pdf/tesla-kseries-overview-lr.pdf
5. Intel. (2016) Intel microprocessor export compliance metrics
6. Esmaeilzadeh H, Blem E, Aman RS, Sankaralingam K, Burger D (2011) Dark silicon and the end of multicore scaling. In: 2011 38th annual international symposium on computer architecture (ISCA). IEEE, pp 365–376
7. Li B, Shan Y, Hu M, Wang Y, Chen Y, Yang H, Memristor-based approximated computation. In: Low power electronics and design (ISLPED), pp 242–247
8. Xu C, Dong X, Jouppi NP, Xie Y (2011) Design implications of memristor-based RRAM cross-point structures. In: Design, automation & test in Europe conference & exhibition (DATE). IEEE, pp 1–6
9. Jo SH, Chang T, Ebong I, Bhadviya BB, Mazumder P, Lu W (2010) Nanoscale memristor device as synapse in neuromorphic systems. Nano Lett 10(4):1297–1301
10. Hu M, Li H, Wu Q, Rose GS (2012) Hardware realization of BSB recall function using memristor crossbar arrays. In: Design automation conference, pp 498–503
11. Chakradhar S, Raghunathan A (2010) Best-effort computing: re-thinking parallel software and hardware. In: 47th ACM/IEEE design automation conference (DAC), pp 865–870
12. Ye R, Wang T, Yuan F, Kumar R, Xu Q (2013) On reconfiguration-oriented approximate adder design and its application. In: Proceedings of the international conference on computer-aided design. IEEE, pp 48–54
13. Venkataramani S, Chippa VK, Chakradhar ST, Roy K, Raghunathan A (2013) Quality programmable vector processors for approximate computing. In: Proceedings of the 46th annual IEEE/ACM international symposium on microarchitecture. ACM, pp 1–12
14. Wong HSP, Lee H-Y, Yu S, Chen Y-S, Wu Y, Chen P-S, Lee B, Chen F, Tsai M-J (2012) Metal-oxide RRAM. Proc IEEE 100(6):1951–1970
15. Yu S, Gao B, Fang Z, Yu H, Kang J, Wong H-SP, (2013) A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation. Adv Mater 25(12):1774–1779
16. Deng Y, Huang P, Chen B, Yang X, Gao B, Wang J, Zeng L, Du G, Kang J, Liu X (2013) RRAM crossbar array with cell selection device: a device and circuit interaction study. IEEE Trans Electron Devices 60(2):719–726
17. Alibart F, Gao L, Hoskins BD, Strukov DB (2012) High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. Nanotechnology 23(7):075201
18. Guan X, Yu S, Wong H-S (2012) A spice compact model of metal oxide resistive switching memory with variations. IEEE Electron Device Lett 33(10):1405–1407
19. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366
20. Ito Y (1994) Approximation capability of layered neural networks with sigmoid units on two layers. Neural Comput 6(6):1233–1243
21. Gu P, Li B, Tang T, Yu S, Cao Y, Wang Y, Yang H (2015) Technological exploration of RRAM crossbar array for matrix-vector multiplication. In: The 20th Asia and south pacific design automation conference (ASPDAC). IEEE, pp 106–111

22. Cannizzaro SO, Grasso AD, Mita R, Palumbo G, Pennisi S (2007) Design procedures for three-stage CMOS OTAs with nested-Miller compensation. IEEE Trans Circuits Syst I Regul Pap 54 (5):933–940
23. Oh W, Bakkaloglu B (2007) A CMOS low-dropout regulator with current-mode feedback buffer amplifier. IEEE Trans Circuits Syst II Express Briefs 54(10):922–926
24. Allen PE, Holberg DR (2002) CMOS analog circuit design. Oxford University Press, Oxford
25. Li B, Wang Y, Chen Y, Li HH, Yang H (2014) Ice: inline calibration for memristor crossbar-based computing engine. In: Proceedings of the conference on design, automation & test in Europe. European Design and Automation Association, p 184
26. Khodabandehloo G, Mirhassani M, Ahmadi M (2012) Analog implementation of a novel resistive-type sigmoidal neuron. IEEE Trans Very Large Scale Integr (VLSI) Syst 20 (4):750–754
27. Fausett L (ed) (1994) Fundamentals of neural networks: architectures, algorithms, and applications. Prentice-Hall, Inc., Upper Saddle River
28. Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. Neural Comput 7(2):219–269
29. Bedeschi F, Fackenthal R, Resta C, Donze E, Jagasivamani M, Buda E, Pellizzer F, Chow D, Cabrini A, Calvi G, Faravelli R, Fantini A, Torelli G, Mills D, Gastaldi R, Casagrande G (2009) A bipolar-selected phase change memory featuring multi-level cell storage. IEEE J Solid-State Circuits 44(1):217–227
30. Lee H, Chen P, Wu T, Chen Y, Wang C, Tzeng P, Lin C, Chen F, Lien C, Tsai M (2008) Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust HFO2 based RRAM. In: IEEE international electron devices meeting (IEDM), pp 1–4
31. Kannan S, Rajendran J, Karri R, Sinanoglu O (2013) Sneak-path testing of crossbar-based nonvolatile random access memories. IEEE Trans Nanotechnol 12(3):413–426
32. ITRS (2013) International technology roadmap for semiconductors
33. Gulati K, Lee H-S (1998) A high-swing CMOS telescopic operational amplifier. IEEE J. Solid-State Circuits 33(12):2010–2019
34. Kull L, Toifl T, Schmatz M, Francese PA, Menolfi C, Braendli M, Kossel M, Morf T, Andersen TM, Leblebici Y (2013) A 3.1 mw 8b 1.2 gs/s single-channel asynchronous SAR ADC with alternate comparators for enhanced speed in 32nm digital SOI CMOS. In: 2013 I.E. international solid-state circuits conference digest of technical papers (ISSCC). IEEE, pp 468–469
35. Lin W-T, Kuo T-H (2013) A 12b 1.6 gs/s 40 mw dac in 40 nm CMOS with $>$ 70db SFDR over entire Nyquist bandwidth. In: 2013 I.E. international solid-state circuits conference digest of technical papers (ISSCC). IEEE, pp 474–475
36. Mutch J, Lowe DG (2008) Object class recognition and localization using sparse features with limited receptive fields. Int J Comput Vision 80(1):45–57
37. Maashri AA, Debole M, Cotter M, Chandramoorthy N, Xiao Y, Narayanan V, Chakrabarti C (2012) Accelerating neuromorphic vision algorithms for recognition. In: Proceedings of the 49th annual design automation conference, pp 579–584
38. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision 88(2):303–338
39. Tang T, Xia L, Li B, Luo R, Chen Y, Wang Y, Yang H (2015) Spiking neural network with RRAM: can we use it for real-world application? In: Design, automation test in Europe conference exhibition (DATE), pp 860–865
40. Liu C, Yan B, Yang C, Song L, Li Z, Liu B, Chen Y, Li H, Wu Q, Jiang H (2015) A spiking neuromorphic design with resistive crossbar. In: Proceedings of the 52nd annual design automation conference. ACM, p 14
41. Seo J-S, Lin B, Kim M, Chen P-Y, Kadetotad D, Xu Z, Mohanty A, Vrudhula S, Yu S, Ye J et al. (2015) On-chip sparse learning acceleration with CMOS and resistive synaptic devices. IEEE Trans Nanotechnol 14(6):969–979
42. Mazady A, Rahman MT, Forte D, Anwar M (2015) Memristor PUF—a security primitive: theory and experiment. IEEE J Emerging Sel Top Circuits Syst 5(2):222–229

43. Bojnordi M, Ipek E (2016) Memristive Boltzmann machine: a hardware accelerator for combinatorial optimization and deep learning. In: International symposium on high performance computer architecture (HPCA)
44. Li B, Gu P, Shan Y, Wang Y, Chen Y, Yang H, Rram-based analog approximate computing. IEEE Trans Comput Aided Des Integr Circuits Syst 34(12):1905–1917
45. Kim K-H, Gaba S, Wheeler D, Cruz-Albrecht JM, Hussain T, Srinivasa N, Lu W (2011) A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. Nano Lett 12(1):389–395
46. Liu X, Mao M, Liu B, Li H, Chen Y, Li B, Wang Y, Jiang H, Barnell M, Wu Q et al. (2015) Reno: a high-efficient reconfigurable neuromorphic computing accelerator design. In: 2015 52nd ACM/EDAC/IEEE design automation conference (DAC). IEEE, pp 1–6
47. Xia L, Li B, Tang T, Gu P, Yin X, Huangfu W, Chen P-Y, Yu S, Cao Y, Wang Y, Xie Y, Yang H (2016) Mnsim: simulation platform for memristor-based neuromorphic computing system. In: Proceedings of the conference on design, automation & test in Europe. European Design and Automation Association
48. Li B, Xia L, Gu P, Wang Y, Yang H, Merging the interface: power, area and accuracy co-optimization for rram crossbar-based mixed-signal computing system. In: 2015 52nd ACM/EDAC/IEEE design automation conference (DAC), pp 1–6
49. Liu B, Li H, Chen Y, Li X, Wu Q, Huang T (2015) Vortex: variation-aware training for memristor x-bar. In: Proceedings of the 52nd annual design automation conference. ACM, p 15
50. Prezioso M, Merrikh-Bayat F, Hoskins B, Adam G, Likharev KK, Strukov DB (2015) Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature 521(7550):61–64 2015.
51. Liu B, Li H, Chen Y, Li X, Huang T, Wu Q, Barnell M, Reduction and ir-drop compensations techniques for reliable neuromorphic computing systems. In: 2014 IEEE/ACM international conference on computer-aided design (ICCAD). IEEE, pp 63–70
52. Wen W, Wu C-R, Hu X, Liu B, Ho T-Y, Li X, Chen Y (2015) An EDA framework for large scale hybrid neuromorphic computing systems. In: Proceedings of the 52nd annual design automation conference. ACM, p 12

Chapter 6
NVRAM-Assisted Optimization Techniques for Flash Memory Management in Embedded Sensor Nodes

Duo Liu and Kan Zhong

Abstract Embedded sensor nodes are sensitize to battery lifetime and evidences show that DRAM-based main memory subsystem is the major contributor of the energy consumption of embedded sensor nodes. Due to the high density, byte-addressability, and low standby power consumption, non-volatile random access memories (NVRAMs), such as PRAM and STT-RAM, become promising main memory alternatives in embedded sensor nodes. On the other hand, NAND flash memory is widely adopted for storing collected data in embedded sensor nodes. However, both NVRAM and NAND flash memory have limited lifetime, how to optimize the management of NAND flash memory in NVRAM-based embedded sensor nodes while considering the endurance issue becomes quite important. In this chapter, we introduce a write-actively-aware NAND flash memory management scheme to effectively manage NAND flash memory while reducing the write activities to NVRAM-based main memory in embedded sensor nodes. The basic idea is to preserve each bit in flash mapping table, which is stored in NVRAM, from being inverted frequently during the mapping table update process. To achieve this, a two-level mapping mechanism is employed while considering the access behavior of IO requests, and a customized wear-leveling scheme is developed to evenly distribute the writes across the whole mapping table. Evaluation results show that the proposed technique can reduce the write activities significantly and achieve an even distribution of writes in NVRAM with low overhead.

Keywords Sensor node • Embedded systems • Flash translation layer • Flash memory • Non-volatile memory • Phase change memory • NVRAM • Wear-leveling • Endurance

D. Liu (✉) • K. Zhong
College of Computer Science, Chongqing University, No. 174 Shazhengjie, Shapingba, Chongqing 400044, China
e-mail: liuduo@cqu.edu.cn; kzhong1991@cqu.edu.cn

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_6

6.1 Introduction

Due to the limitation in size and cost, embedded sensor nodes are commonly equipped with a small battery, which has limited capacity, making the sensor nodes sensitize to battery lifetime. Recent researches report that DRAM-based main memory subsystem has become the major contributor of the embedded system's overall energy consumption [1, 2]. To solve this problem, a mount of researches argue that non-volatile random access memory (NVRAM), such as phase change memory (PCM) [3–10], spin-transfer torque random access memory (STT-RAM) [11–13], is a promising DRAM alternative [14]. However, compared to DRAM, NVRAM exhibits limited endurance (e.g., 10^6–10^8 for PRAM cells) and high write latency/energy [15]. These constraints impose challenges for using NVRAM as a complete replacement for DRAM. On the other hand, NAND flash memory has been widely used in embedded sensor nodes due to its attractive features, such as shock resistance, low power, and high density [16]. To manage flash memory, FTL is introduced to emulate NAND flash memory as a block device interface for file systems [17]. The FTL functions as translating logical addresses of I/O requests into physical addresses in NAND flash memory. To achieve this, FTL maintains a mapping table (i.e., metadata), which stores the mapping information between logical addresses and physical addresses. The mapping table is usually cached in main memory for better performance and written back to NAND flash periodically.

Over the past decade, many studies for FTL schemes have been proposed [18–27]. According to the granularity of mapping unit, there are three types of FTL schemes: page-level mapping, block-level mapping, and hybrid-level mapping [17]. Most of the previous work, however, have not yet explored the management mechanism of NAND flash memory in the emerging NVRAM-based embedded sensor nodes. Kim et al. [28] propose a page-level mapping FTL (*h*FTL) for managing NAND flash memory in the NVRAM-based embedded systems, where the page-level mapping table is stored in NVRAM and user data is stored in NAND flash memory. Nevertheless, their approach does not consider write activities of FTL mapping table in NVRAM, and the access behavior of I/O requests as well. As FTL mapping table is updated frequently in NVRAM, a huge number of unnecessary write operations on FTL mapping table will degrade the endurance of NVRAM. New techniques, therefore, are needed to eliminate unnecessary write operations on FTL mapping table and, at the same time, to enhance the endurance of NVRAM-based sensor nodes.

In this chapter, we introduce a write-activity-aware two-level FTL scheme, called **NV-FTL**, to effectively manage NAND flash memory and enhance the endurance of NVRAM-based embedded sensor nodes. Different from existing approaches [29–34], NV-FTL enhances the lifetime of NVRAM by making the management of NAND flash memory aware of write activities on underlying memory architecture. With NV-FTL, no change to the file systems or hardware implementation of NAND flash and NVRAM is required. Our basic idea is to preserve each bit in FTL mapping table, which is stored in NVRAM, from being

inverted frequently, i.e., we focus on minimizing the number of bit flips in an NVRAM cell when updating the FTL mapping table. NV-FTL employs a two-level mapping mechanism, which not only focuses on minimizing write activities of NVRAM but also considers the access behavior of I/O requests. To achieve this, in NVRAM, we use a page-level mapping table to handle not frequently updated random requests, and allocate a tiny buffer of block-level mapping table to record most frequently updated sequential requests. To further minimize write activities in NVRAM, NV-FTL actively chooses a physical block in NAND flash memory whose physical block number (PBN) incurs minimum number of bit flips. Consequently, the write activities are eliminated and the endurance of NVRAM is enhanced.

We conduct trace-driven experiments with both general purpose and mobile I/O workloads to show the effectiveness and versatility of NV-FTL. A representative FTL design *h*FTL [28] for NVRAM-based embedded systems is selected as a baseline scheme. The proposed NV-FTL is compared with *h*FTL in terms of NVRAM bit flips with various configurations. The experimental results show that our approach can achieve an average reduction of 93.10 % and a maximum reduction of 98.98 % in the maximum number of bit flips for an NVRAM-based embedded sensor nodes with 1 GB NAND flash memory. In addition, the results also show that NV-FTL can achieve an even distribution of bit flips in NVRAM when compared with the baseline scheme.

The rest of this chapter is organized as follows. Section 6.2 introduces the background and motivation. Section 6.3 presents our proposed NV-FTL technique. Section 6.4 reports the experimental results. Finally, in Sect. 6.5, we present the conclusion.

6.2 Background and Motivation

In this section, we first introduce the background knowledge of NVRAM-based embedded sensor node. Then we describe the issues of a representative FTL scheme. Finally, we present the motivation of our work.

6.2.1 NVRAM-Based Sensor Node

Figure 6.1 shows a typical NVRAM-based embedded sensor node. As shown, NVRAM is served as the sensor node's main memory and NAND flash memory is adopted as the storage media. Analog signals collected by various sensors are firstly convert into digital signals by the analog digital converter (ADC), and then the digital signals are processed by the CPU and stored in the storage system [35], in which the FTL mapping table is cached in NVRAM and sensor data are stored in NAND flash memory. In the storage system, the MTD layer provides primitive functions such as read, write, and erase operations. The FTL layer emulates the

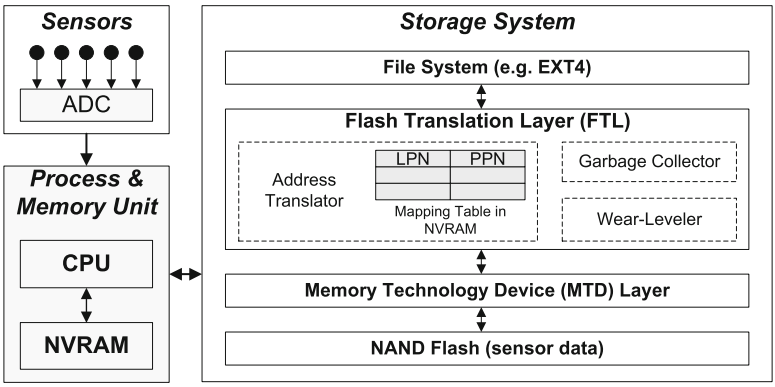


Fig. 6.1 Illustration of NVRAM-based embedded sensor node with NAND flash memory

flash memory as a disk device so that it can provide transparent storage service to file systems. Following the I/O requests, FTL translates addresses between logical page number (LPN) and physical page number (PPN), and keeps track of the mapping information by using an FTL mapping table in NVRAM. Then according to the mapping, data can be directly read from (write into) NAND flash memory.

Unlike NAND flash memory, NVRAMs support bit-addressability and in-place update. NVRAMs keep data by changing the physical state of its underlaying material without maintaining constant current. One promising candidate is PCM, which stores data by changing the state of the phase change material (e.g., GST). By ejecting electrical pulses to heat up the GST region, each PCM cell can switch between two states—amorphous and crystalline, which have high and low electrical resistance, respectively. Reading a bit from a PCM cell is accomplished by sensing the resistance level of the cell. To represent binary "1", a SET operation is performed to turn a PCM cell into the crystalline state by applying a moderate power, long duration pulses; To represent binary "0", a RESET operation is performed to turn a PCM cell into the amorphous state by applying a high power, short duration pulses. Both of these operations impose heat stress to PCM cells, and thus a PCM cell can only sustain a limited number of write (SET/RESET) operations (e.g., 10^6–10^8 for Micron P5Q PCM [36]). Other NVRAM like STT-RAM, memristor [37] also suffers from the endurance problem. In this chapter, we do not target at any specific NVRAM, we target at the optimization of NAND flash memory management in NVRAM-based sensor nodes since all the NVRAMs have the same problem—limited endurance.

6.2.2 A Representative FTL Scheme

In this section, we briefly revisit the *h*FTL scheme which is proposed for managing NAND flash memory in PCM-based embedded systems [28].

*h*FTL is based on page-level mapping scheme [18], but it is optimized for PCM-based embedded systems. *h*FTL stores metadata such as FTL mapping

table, physical page information, and physical block information in PCM. NAND flash memory is only used for storing user data from the file system, and the blocks in NAND flash memory are categorized into three types, i.e., garbage blocks, data blocks, and a buffer block. Different from the conventional page-level mapping FTL, *h*FTL uses a buffer block to store the newly arrived data. When the buffer block runs out of free pages, it is put into the data block list and another empty buffer block is allocated from the garbage block list. If there is not enough number of garbage blocks, a garbage collection operation is performed to reclaim a block from the data blocks. In *h*FTL, a page-level mapping table in PCM keeps track of mappings between LPN and PPN, in terms of the I/O requests. Consequently, the mapping table is updated frequently and thus imposes the endurance issue for PCM. A motivational example is illustrated in Fig. 6.2.

In the example, we presume that PCM is adopted as the main memory of sensor node, and there are four blocks in NAND flash memory, and each block has 8 pages. Therefore, a page-level mapping table in PCM has 32 entries to record the mapping information. To facilitate the comparison of *h*FTL and our NV-FTL scheme, the PPN, PBN, and the offset of each block are represented by binary number. We assume that each entry of the mapping table is empty at the beginning, and the binary number in an entry is the updated PPNs to reflect the updates of mapping. The I/O access requests of write operations (w) are listed in Fig. 6.2a. According to the given I/O requests, the status variation of the blocks in NAND flash memory is shown in Fig. 6.2b. For *h*FTL, when a write operation is performed, the corresponding content is first written to a free page of the current buffer block in a sequence order.

As shown, the first request is written to LPN (#18). A new buffer block (PBN #00) is allocated from the garbage block list, and the content *A* with the corresponding LPN (#18) is stored in the first page of current buffer block (PBN #00). Meanwhile, the mapping information of LPN (#18) and PPN (#00000) is stored into the mapping table shown in Fig. 6.2c. Note that PPN is the combination of PBN and the block offset. After serving the eighth request, buffer block (PBN #00) is full and becomes a data block. Likewise, the remaining garbage blocks (PBN #01, PBN #10, and PBN #11) are allocated as a buffer block, respectively, to serve the following write operations. Finally, when the content of *N*2 with the corresponding LPN (#29) is written into the last page of buffer block (PBN #11), all garbage blocks become data blocks and some entries of the mapping table have been updated by new PPNs for several times.

6.2.3 Motivation

In the motivational example, several update operations are performed in the FTL page-level mapping table. For instance, the 13th request updates the old content in the 1st page of data block (PBN #00) by setting that page invalid, and writes the new content to the current buffer block (PBN #01). Meanwhile, the corresponding

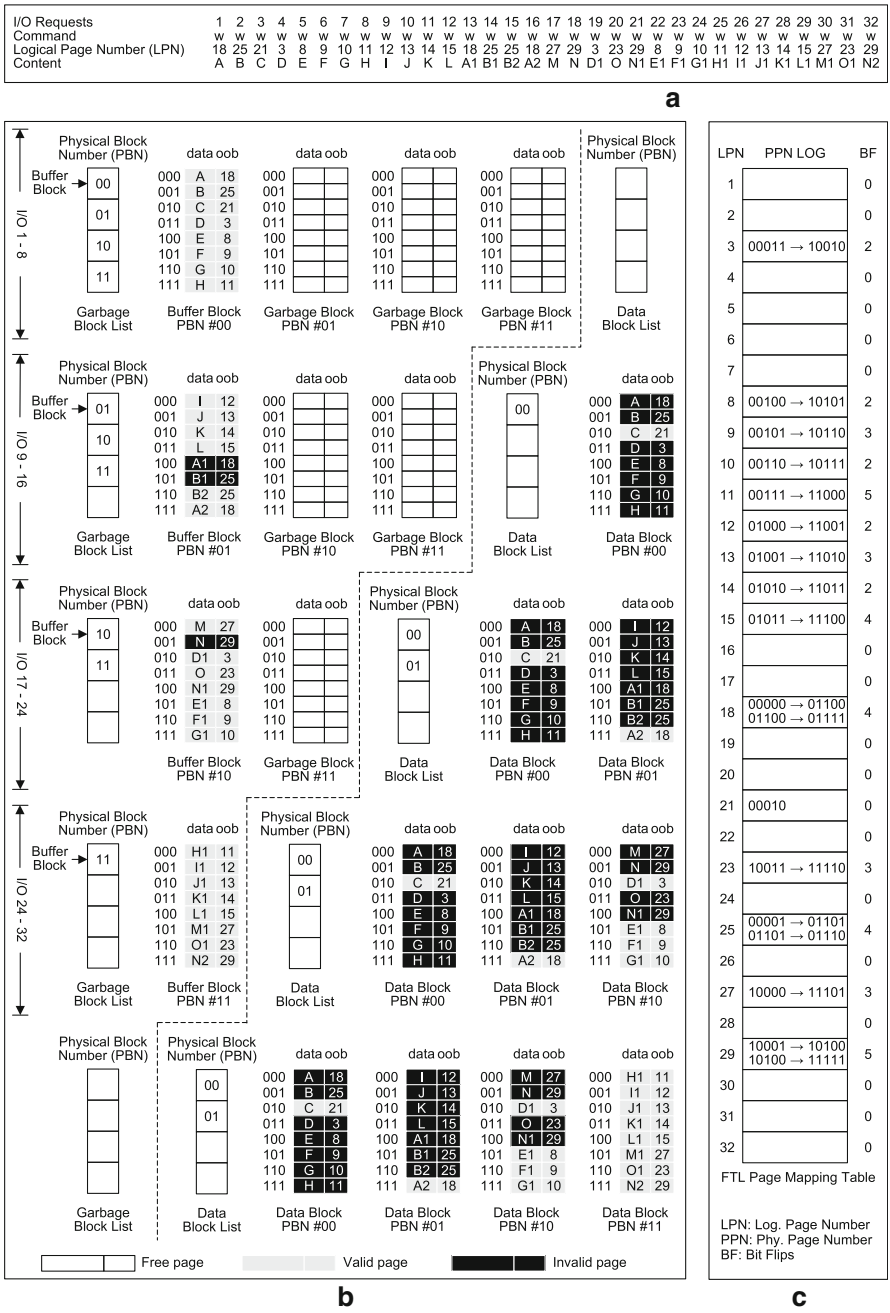


Fig. 6.2 Motivational example. (**a**) I/O access requests. (**b**) The status variation of blocks in NAND flash memory. (**c**) The status variation of FTL page-level mapping table in PCM

mapping information in the mapping table is updated as well. In Fig. 6.2c, we use the bit flips (BF), shown on the right side of the mapping table, to reflect the update frequency of each entry in the mapping table. As shown, the 11th and 29th entry have the maximum number of bit flips 5. Since NVRAM cell, like PCM, can only sustain limited number of write cycles, frequent update operations in mapping table will lead to the fast worn out of NVRAM. These observations motivate us to propose a write-activity-aware FTL to effectively manage NAND flash memory and, at the same time, to improve the endurance of NVRAM-based embedded sensor node.

As mentioned above, several hardware optimization techniques for NVRAM have been proposed [38–40] to tackle the redundant write activities by eliminating a write if its designated memory cell holds the same value. Then through utilizing such a fine-grained hardware feature, this work actively chooses mapping information (e.g., PBN) which is almost the same as the mapping to be updated in the mapping table, such that the number of write activities in NVRAM is minimized.

6.3 NV-FTL: Write-Activity-Aware FTL

In this section, we present the details of our NV-FTL, a write-activity-aware FTL, that can effectively enhance the endurance of the NVRAM-based embedded sensor node. We first present an overview of NV-FTL in Sect. 6.3.1. We then provide a detailed description of NV-FTL in Sect. 6.3.2.

6.3.1 Overview

The objective of NV-FTL is to reduce write activities in NVRAM-based embedded sensor node, and therefore, the endurance of NVRAM is enhanced. So the basic idea of NV-FTL is to preserve each bit in FTL mapping table, which is stored in NVRAM, from being inverted frequently, i.e., we focus on minimizing the number of bit flips in an NVRAM cell when updating the FTL mapping table. Different from the previous work [28], our NV-FTL adopts a two-level mapping mechanism, which not only focuses on minimizing write activities in NVRAM but also considers the access behavior of I/O requests. NV-FTL uses a page-level mapping table to record the mapping of write requests not frequently updated, and allocates a tiny buffer of block-level mapping table to cache the mapping of those most frequently updated write requests. With the consideration of write activities, once a block is needed for incoming write requests, NV-FTL actively chooses a physical block in NAND flash memory whose PBN incurs minimum number of bit flips.

By applying NV-FTL, the number of bit flips is reduced, and thus the number of write activities in NVRAM is minimized. Consequently, the endurance of the NVRAM-based embedded sensor node is enhanced.

6.3.2 NV-FTL Description

In general, a realistic I/O workload is a mixture of random and sequential requests. By separating the random requests from the sequential requests, we can not only obtain the access behavior but also handle those frequently updated write requests. Otherwise, without considering the access behavior of I/O workload, we cannot effectively manage NAND flash memory and may waste lots of blocks in garbage collection due to frequent update operations. Therefore, in NV-FTL, we design a behavior detector to separate the I/O workload into random and sequential requests, according to the length of each request in the I/O workload. The length is a user-defined threshold, which is determined by observing performance gains with different threshold values (e.g., 8, 16, and 32) in the experiments. For example, if the length of a request is smaller than 8, then this request is treated as a random request; Otherwise, if the length of a request is greater than or equal to 8, then it is treated as a sequential request.

Figure 6.3 shows the structure of NV-FTL. As shown, NV-FTL first separates the I/O workload into random requests and sequential requests. Then NV-FTL adopts a two-level FTL mechanism to handle these two cases as follows:

- For random requests: NV-FTL sequentially allocates physical pages from the first page of a physical block in NAND flash memory, so that all pages in blocks are fully utilized. Accordingly, NV-FTL adds LPN to PPN mapping of random requests into the page-level mapping table.
- For sequential requests: NV-FTL allocates physical pages based on block offset as most sequential requests usually occupy a whole block, so that all pages in blocks are fully utilized as well. Similarly, NV-FTL adds an LBN to PBN mapping of sequential requests into the block-level mapping table buffer.

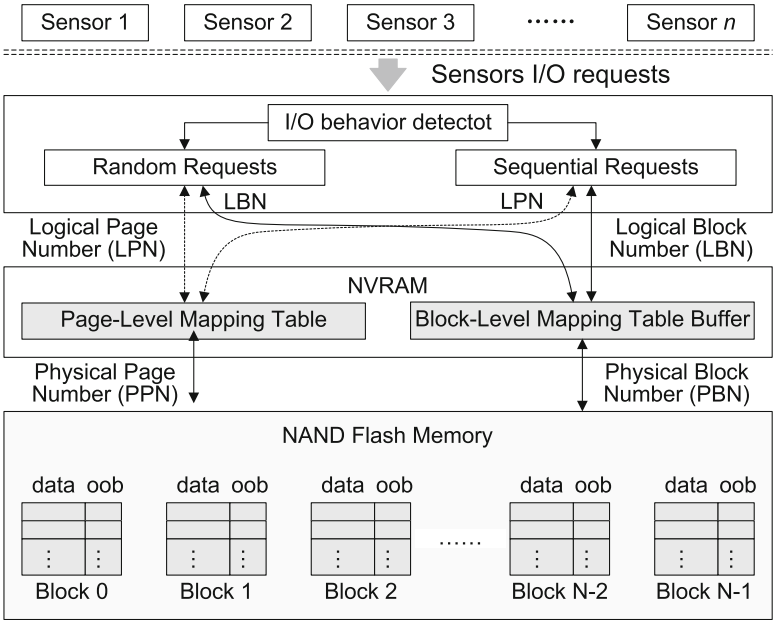


Fig. 6.3 Structure of NV-FTL

In NV-FTL, we only allocate a tiny buffer for temporary storing a part of the block-level mapping table. For example, the size of this block-level mapping buffer is set as 10 % of the size of the original block-level mapping table. Therefore, a replacement policy should be considered when the buffer is full. Similar as a cache, we only kick out the mapping of those not frequently updated blocks, while maintaining the mapping of frequent updated blocks. The kicked out mapping information is put into the page-level mapping table. If a block in NAND flash memory has N_p valid pages, and its corresponding block-level mapping is kicked out to page-level mapping table, then N_p entries in page-level mapping table should be filled with the corresponding LPN to PPN mapping for each page in the block. On the contrary, the page-level mapping of a block can be re-added into the block-level mapping table buffer, once the block is updated again by sequential write requests. Therefore, by observing the frequently updated requests, our technique can dynamically adjust the block-level mapping table buffer and the page-level mapping table, such that write activities of frequently updated requests are only buffered in block-level mapping table buffer which only contributes a small number of bit flips in NVRAM. The experimental results in Sect. 6.4 confirms this fact.

To further minimize write activities in NVRAM, a write-activity-aware strategy is proposed. In our technique, to allocate a new block for the write/update requests, the corresponding original physical block number (PBN) is first obtained from page-level mapping table (by dividing PPN with the number of pages in a block), or from block-level mapping table buffer with the requested LPN. Then according to the original PBN, we actively select a physical block in NAND flash memory whose PBN is almost the same as the original PBN, i.e., the new PBN incurs minimum number of bit flips if the original PBN is updated by the new PBN in the mapping table. As a result, a large number of redundant bit flips are reduced, and the endurance of NVRAM is enhanced.

Algorithm 6.3.1 The algorithm of NV-FTL

Require: I/O requests with random request or/and sequential request.
Ensure: Allocate pages for the I/O request.

1: Divide the I/O request into random writes or/and sequential writes according to a predefined threshold.
2: **if** Random write request arrives **then**
3: Obtain the *LBN* and *LPN* of the random write request.
4: **if** *LBN*'s mapping is not in block-level mapping table buffer or *LPN*'s mapping is not in page-level mapping table **then**
5: This is a new write, allocate a new block *PBN*, and write the contents into the block sequentially from the first page.
6: Add the mapping of (*LPN*, *PPN*) into the page-level mapping table.
7: **end if**

(continued)

Algorithm 6.3.1 (continued)

```
 8:  if LBN's mapping exists in block-level mapping table buffer or LPN's
     mapping exists in page-level mapping table then
 9:    This is an update, obtain the PBN of the updated block.
10:    if There exists enough space in the PBN block for the update
       request then
11:      Write the update contents in the left space of the PBN block sequen-
         tially, and invalid the old pages in the same block.
12:    else
13:      Actively find a new block whose block number is almost the same as
         PBN, write the update contents in the new block sequentially, and
         invalid the old pages in PBN block.
14:    end if
15:    Update block-level mapping table buffer or page-level mapping table.
16:  end if
17: end if
18: if Sequential write request arrives then
19:  Obtain the LBN and LPN of the sequential write request.
20:  if LBN's mapping is not in block-level mapping table buffer or LPN's
     mapping is not in page-level mapping table then
21:    This is a new write, allocate a new block PBN, and write the contents of
       the request into the block based on block offset.
22:    if The block-level mapping table buffer is full then
23:      Kick out least frequently used entry, add the kicked out mappings into
         page-level mapping table.
24:    end if
25:    Add the mapping of (LBN, PBN) into the block-level mapping table
       buffer.
26:  end if
27:  if LBN's mapping exists in block-level mapping table buffer or LPN's
     mapping exists in page-level mapping table then
28:    This is an update, obtain the PBN of the updated block.
29:    if There exists enough space in the PBN block for the update
       request then
30:      Write the update contents in the left space of the PBN block based on
         block offset, and invalid the old pages in the same block.
31:    else
32:      Actively find a new block whose block number is almost the same as
         PBN, write the update contents in the new block based on block offset,
         and invalid the old pages in PBN block.
33:    end if
34:    Update block-level mapping table buffer or page-level mapping table.
35:  end if
36: end if
```

Algorithm 6.3.1 shows the process of a write operation of NV-FTL. NV-FTL first divides the incoming I/O request into random writes or/and sequential writes according to a threshold. Then the random and sequential write requests are processed separately. For random write request (lines 2–17), if it is a new write, i.e., we cannot find its corresponding *LBN* or *LPN* mapping in the block-level mapping table buffer or page-level mapping table. So NV-FTL finds a new block *PBN*, and write the contents of the random write request into the allocated new block sequentially from the first page. After that, we add the (*LPN*, *PPN*) mapping into the page-level mapping table. If the random write request is an update, and there exists enough space in the updated block, then write the update contents into the left space of the block sequentially, and invalid the old pages in the same block. Otherwise, there does not exist enough space in the updated block, NV-FTL will actively find a new block whose block number is almost the same as *PBN*, and then write the update contents in the new block based on block offset. At last, we update the corresponding block-level mapping table buffer or page-level mapping table. For sequential write request (lines 18–36), we process it in the similar way as that for processing random write request.

Note that the block-level mapping table buffer is updated frequently by sequential write requests, so it may become very hot and lead to an uneven distribution of bit flips in NVRAM. To avoid this scenario and enhance NVRAM endurance, a wear-leveling method is integrated into NV-FTL. In NV-FTL, during a period of time (e.g., every 100 I/O requests), the block-level mapping table buffer is moved across the whole mapping table area (block-level and page-level mapping table) in NVRAM. With acceptable copy operations of mapping information, an even distribution of bit flips in NVRAM is obtained.

An example of NV-FTL is shown in Fig. 6.4. This example is based on the I/O requests and the NAND flash memory assumptions for the motivational example shown in Fig. 6.2. As shown, for the first random request with LPN (#18), we find a new block (PBN #00), and the content *A* is written sequentially into the first page (#00000) of block (PBN #00). For this request, there is no bit flip when updating the mapping table. It can be seen that *A* is updated by a new content *A*1 in the 13th request, and *A*1 is written into the physical page (#00010) according to the update policy of NV-FTL. When the 13th request arrives, we use the LPN (#18) to get the corresponding LBN (#10). Then we find the LBN (#10) is already in the block-level mapping table buffer, so the 13th request is an update to the old page in the block (PBN #00), then by checking the block (PBN #00), we know the old content *A* of this LPN (#18) is stored in the page PPN (#00000), thus this page is set as invalid. Since there exists enough space in block (PBN #00), the new update content *A*1 of LPN (#18) is written sequentially into the block.

It is noticed that the 5th to 12th requests form a sequential write, then we allocate a new block (PBN #11) for this request, and write the contents into each page of the block based on offset. The corresponding LBN to PBN mapping (01, 11) is added into the block-level mapping table buffer. Later, when the following 22nd to 29th sequential update requests arrive, then the old pages in the block (PBN #11) are invalid. Since we cannot find free block, the block (PBN #11) is erased, and the new

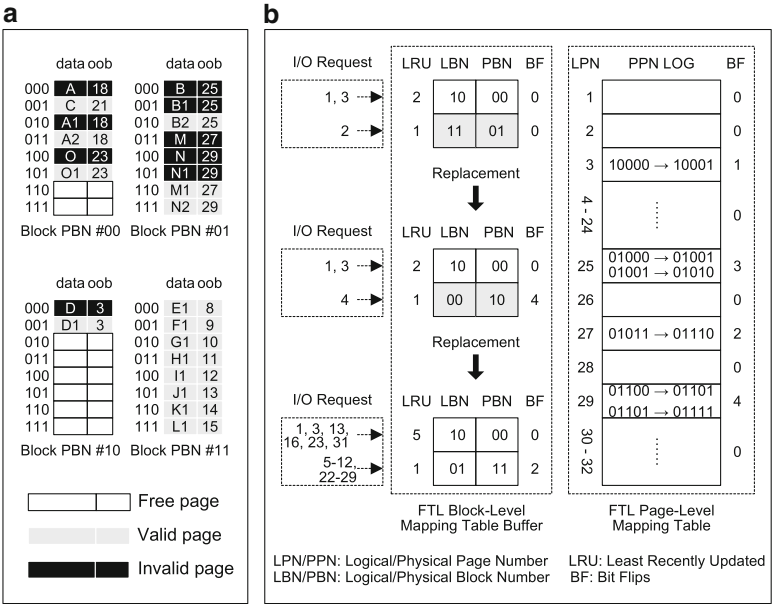


Fig. 6.4 Illustration of NV-FTL. (**a**) The status variation of blocks in NAND flash memory according to the access sequence in Fig. 6.2. (**b**) The status variation of FTL page-level mapping table and block-level mapping table buffer in NVRAM

update data E1 to L1 is written into this block based on offset. Finally, we update the block-level mapping table buffer, and the value of corresponding LRU is updated as well.

After processing all requests, we found that the total number of bit flips in NVRAM is 16 by our NV-FTL, while the total number of bit flips in NVRAM are 44 by *h*FTL. Our scheme achieves a reduction of 63.6 % in the total number of bit flips, which confirms that our approach can effectively reduce write activities in NVRAM. The experimental results in Sect. 6.4 also show that our scheme can effectively reduce the total number of bit flips.

6.4 Evaluation

To evaluate the effectiveness of the proposed NV-FTL, we conduct a series of experiments and present the experimental results with analysis in this section. We compare and evaluate our proposed NV-FTL scheme over the representative page-level FTL scheme, *h*FTL[28].

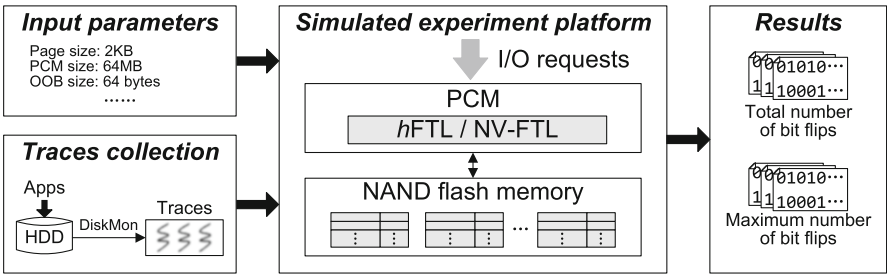


Fig. 6.5 The framework of simulation platform

6.4.1 Experimental Setup

Although our design does not target at any specifical NVRAM, in the evaluation, we assume that a particular NVRAM—PCM—is used as the main memory of embedded sensor node. The evaluation is through a trace-driven simulation. The framework of our simulation platform is shown in Fig. 6.5. In our experiment, we use the same experimental configuration adopted by *h*FTL [28], a 1 GB NAND flash memory and a 64 MB PCM are configured in our simulator. The traces along with various flash parameters, such as block size and page size, are fed into our simulation framework. The page size, number of pages in a block, and size of the OOB for each page are set as 2 KB, 64, and 64 Bytes, respectively. Therefore, the 1 GB NAND flash memory used in the experiment has 8,192 physical blocks. To fully evaluate our scheme, we further conduct the experiments on a 4 GB NAND flash memory with the same configurations. In addition, the threshold for distinguishing random and sequential requests is set as 8.

To represent the realistic I/O request patterns, we collected the traces from desktop running DiskMon [41] with an Intel Pentium Dual Core 2 GHz processor, a 200 GB hard disk, and a 2 GB DRAM. Among these traces, CopyFiles is a trace collected by copying files from hard disk to an external hard drive; DownFiles represents a trace collected by downloading files from a network server; Office represents a trace collected by running some office related applications; P2P represents a trace collected by running a P2P file-sharing application on an external hard drive; Table 6.1 summarizes our experimental platform and trace collection environment.

6.4.2 Results and Discussion

In this section, we present the experimental results with analysis. We first present the endurance impact of NV-FTL. Then we present the wear-leveling comparison of NV-FTL and the baseline scheme.

Table 6.1 Experimental setup

Hardware	CPU	Intel dual core 2 GHz
	Disk space	200 GB
	RAM	2 GB
Simulation environment	OS kernel	Linux 2.6.17
	Flash size	1 GB & 4 GB
	PCM	64 MB
Trace	OS	Windows XP (NTFS)
	Trace name generator	DiskMon
	Applications	Web applications, MSN, Word, Excel, PowerPoint, Media player, Emuler

6.4.2.1 NVRAM Endurance

The objective of this work is to reduce write activities to enhance the endurance of NVRAM-based embedded sensor node. Therefore, the endurance of NVRAM is one of the most important factors in analyzing the reliability of NVRAM-based embedded sensor node. The endurance of NVRAM is mainly affected by the worst case of bit flips in an NVRAM cell, i.e., the maximum number of bit flips in a NVRAM cell determines the endurance of NVRAM. For example, if PCM can only sustain 10^6 write cycles, then a PCM cell is worn out if it suffers from more than 10^6 bit flips. So our technique not only focuses on minimizing write activities in NVRAM but also reducing the maximum number of bit flips. Table 6.2 presents the results for the maximum and total number of bit flips among all PCM cells when managing 1 GB and 4 GB NAND flash memory embedded sensor node.

We observe that NV-FTL can significantly reduce write activities of PCM in comparison with the baseline scheme—*h*FTL. As shown in the table, for the embedded sensor node with 1 GB NAND flash memory, NV-FTL can achieve more than 60 % reduction of total number of bit flips. Similarly, for 4 GB NAND flash memory, NV-FTL also achieves a great amount of total number of bit flips when compared with the baseline scheme, proofing that NV-FTL can effectively preserve the PCM cells being converted frequently.

Moreover, in terms of maximum of bit flips, NV-FTL exhibits a similar tend find in the result of total number of bit flips. When compared with the baseline scheme, NV-FTL can reduce the maximum number of bit flips more than 90 and 80 % for 1 GB and 4 GB NAND flash memory, respectively. To some extend, the reduction in maximum number of bit flips can slow down the wearing out of certain NVRAM cells. Since NV-FTL can reduce both the total number of bit flips and maximum number of bit flips, we therefore conclude that NV-FTL can effectively prolong the endurance of NVRAM, such as PCM, making the NVRAM-based embedded sensor node has a longer longevity.

Table 6.2 The maximum and total number of bit flips of NV-FTL versus *h*FTL

PCM with 1 GB NAND flash memory					
Trace name	% of write	% of read	Total number of bit flips		
			hFTL	NV-FTL	NV-FTL over hFTL (%)
CopyFiles	78.75	21.25	559,496,658	293,866,292	47.48
DownFiles	71.88	28.12	1,756,464,372	568,987,257	67.61
Office	77.37	22.63	7,520,028,995	2,576,892,175	65.73
P2P	28.95	71.05	6,929,967,624	1,718,812,456	75.20
Average					64.00
Trace name	% of write	% of read	Maximum number of bit flips		
			hFTL	NV-FTL	NV-FTL over hFTL (%)
CopyFiles	78.75	21.25	9,977	519	94.80
DownFiles	71.88	28.12	21,945	567	97.42
Office	77.37	22.63	9,385	1,762	81.23
P2P	28.95	71.05	74,540	762	98.98
Average					93.10
PCM with 4 GB NAND flash memory					
Trace name	% of write	% of read	Total number of bit flips		
			hFTL	NV-FTL	NV-FTL over hFTL (%)
CopyFiles	78.75	21.25	122,605,530	93,456,325	23.77
DownFiles	71.88	28.12	842,401,436	191,579,924	77.26
Office	77.37	22.63	6,981,790,260	919,567,590	86.83
P2P	28.95	71.05	15,269,865,958	650,611,634	95.74
Average					70.90
Trace name	% of write	% of read	Maximum number of bit flips		
			hFTL	NV-FTL	NV-FTL over hFTL(%)
CopyFiles	78.75	21.25	10,461	2,076	80.15
DownFiles	71.88	28.12	20,857	3,923	81.19
Office	77.37	22.63	36,667	7,377	79.88
P2P	28.95	71.05	86,383	7,623	91.18
Average					83.10

6.4.2.2 PCM Wear-Leveling

Wear-leveling is another major concern in NVRAM-based embedded sensor node. A good wear-leveling not only prolong NVRAM-based embedded sensor node's longevity, but also increase its reliability since worn out cells may lead to corrupted data. Therefore, in Fig. 6.6, we plot the maximum number of bit flips among all mapping table entries for sensor node with 64 MB PCM and 1 GB NAND flash memory. For each subfigure, the *x*-axis denotes the number of page-level and block-level mapping table entries in PCM, while the *y*-axis shows the maximum number

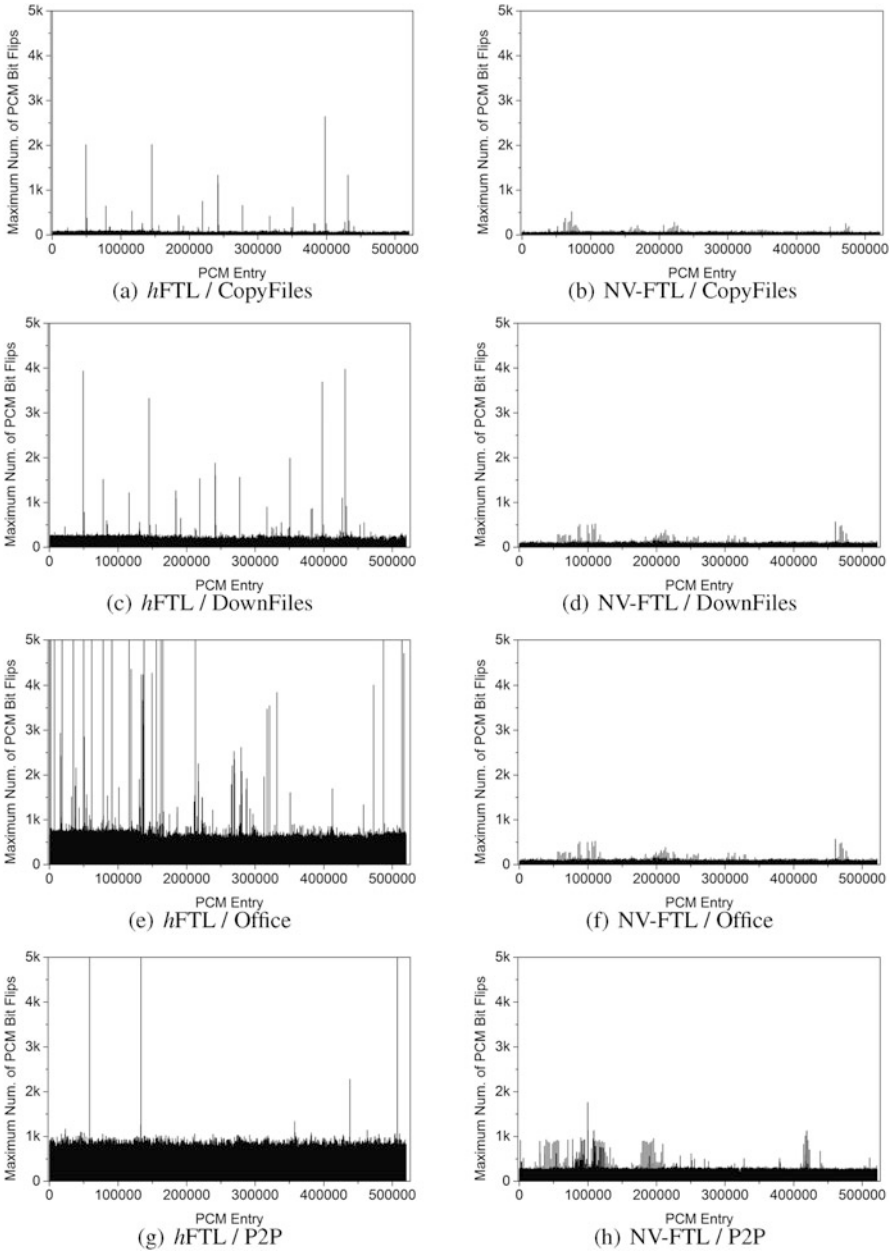


Fig. 6.6 The wear-leveling comparison of *h*FTL and NV-FTL in a PCM-based embedded sensor node with 1 GB NAND flash memory over four traces. (**a**) *h*FTL/CopyFiles, (**b**) NV-FTL/CopyFiles, (**c**) *h*FTL/DownFiles, (**d**) NV-FTL/DownFiles, (**e**) *h*FTL/Office, (**f**) NV-FTL/Office, (**g**) *h*FTL/P2P, (**h**) NV-FTL/P2P

of bit flips extracted from each mapping table entry. To present the distributions clearly, we restrict the maximum number of bit flips on *y*-axis to 5,000.

As shown in the figure, we observe that the write distribution of bit flips for *h*FTL varies a lot, and this will surely pose a threat to the endurance of PCM, making certain PCM cells worn out quickly, as well as other NVRAMs. However, compared to *h*FTL, NV-FTL distributes write activities more evenly among all PCM cells, especially for DownFiles and Office, the wear-leveling of which have a great improvements in NV-FTL. The results listed in Table 6.2 also illustrate this fact. In summary, NV-FTL achieves much better wear-leveling than the baseline scheme, leading to the NVRAM-based embedded sensor nodes have a better reliability.

6.5 Conclusion

In this chapter, we have proposed a write-activity-aware NAND flash memory management scheme NV-FTL which takes the first step to reduce write activities in NVRAM-based sensor node. In our NV-FTL, the performance improvement is achieved by preserving a bit in an NVRAM cell from being inverted frequently. Through a two-level mapping mechanism, and a write-activity-aware strategy, unnecessary write activities in NVRAM are directly eliminated. We conducted experiments on a set of realistic I/O workload collected from daily-life. For a sensor node with 64 MB PCM and 1 GB (4 GB) NAND flash memory, the experimental results show that the maximum number of bit flips among PCM cells can be reduced by 93.10 % (83.10 %) on average, and the total number of bit flips of all PCM cells can be reduced by 64.00 % (70.90 %) on average. Furthermore, the results show that NV-FTL can evenly distribute write activities among PCM cells in comparison with a representative baseline FTL scheme.

References

1. Duan R, Bi M, Gniady C (2011) Exploring memory energy optimizations in smartphones. In: Proceedings of the international green computing conference and workshops (IGCC '11), pp 1–8
2. Perrucci G, Fitzek F, Widmer J (2011) Survey on energy consumption entities on the smartphone platform. In: Proceedings of the IEEE 73rd vehicular technology conference (VTC '11), pp 1–6
3. Wong HSP, Raoux S, Kim S, Liang J, Reifenberg JP, Rajendran B, Asheghi M, Goodson KE (2010) Phase change memory," Proc IEEE 98(12):2201–2227
4. Cho S, Lee H, Flip-n-write: a simple deterministic technique to improve pram write performance, energy and endurance. In: Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO '09), pp 347–357

5. Hu J, Xue CJ, Zhuge Q, Tseng W-C, Sha EH-M (2013) Write activity reduction on non-volatile main memories for embedded chip multiprocessors. ACM Trans Embed Comput Syst 12(3):77:1–77:27
6. Lee BC, Ipek E, Mutlu O, Burger D (2009) Architecting phase change memory as a scalable DRAM alternative. In: Proceedings of the 36th annual international symposium on computer architecture (ISCA '09), pp 2–13
7. Qureshi MK, Karidis J, Franceschini M, Srinivasan V, Lastras L, Abali B (2009) Enhancing lifetime and security of PCM-based main memory with start-gap wear leveling. In: Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO '09), pp 14–23
8. Qureshi MK, Srinivasan V, Rivers JA (2009) Scalable high performance main memory system using phase-change memory technology. In: Proceedings of the 36th annual international symposium on computer architecture (ISCA '09), pp 24–33
9. Dhiman G, Ayoub R, Rosing T (2009) PDRAM: a hybrid PRAM and DRAM main memory system. In: Proceedings of the 46th annual design automation conference (DAC '09), pp 664–469
10. Ferreira AP, Zhou M, Bock S, Childers B, Melhem R, Mossé D (2010) Increasing PCM main memory lifetime. In: Proceedings of the conference on design, automation and test in Europe (DATE '10), pp 914–919
11. Hosomi M, Yamagishi H, Yamamoto T, Bessho K, Higo Y, Yamane K, Yamada H, Shoji M, Hachino H, Fukumoto C, Nagao H, Kano H (2005) A novel nonvolatile memory with spin torque transfer magnetization switching: spin-RAM. In: Proceedings of the IEEE international on electron devices meeting (IEDM '05), pp 459–462
12. Oboril F, Bishnoi R, Ebrahimi M, Tahoori M (2015) Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy. IEEE Trans Comput Aided Des Integr Circuits Syst 34(3):367–380
13. Wen W, Zhang Y, Chen Y, Wang Y, Xie Y (2014) PS3-RAM: a fast portable and scalable statistical STT-RAM reliability/energy analysis method. IEEE Trans Comput Aided Des Integr Circuits Syst (TCAD) 33(11):1644–1656
14. Xue CJ, Zhang Y, Chen Y, Sun G, Yang JJ, Li H (2011) Emerging non-volatile memories: opportunities and challenges. In: Proceedings of the seventh IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis (CODES+ISSS '11), pp 325–334
15. International Technology Roadmap for Semiconductors (2007) Process integration, devices, and structures (2007 edition). http://developer.intel.com
16. Xie Y (2011) Modeling, architecture, and applications for emerging memory technologies. IEEE Des Test Comput 28(1):44–51
17. Chung T-S, Park D-J, Park S, Lee D-H, Lee S-W, Song H-J (2009) A survey of flash translation layer. J Syst Archit 55(5–6):332–343
18. Ban A (1995) Flash file system. US patent 5,404,485
19. Ban A (1999) Flash file system optimized for page-mode flash technologies. US patent 5,937,425
20. Wu C-H, Kuo T-W (2006) An adaptive two-level management for the flash translation layer in embedded systems. In: Proceedings of the 2006 IEEE/ACM international conference on computer-aided design (ICCAD '06), pp 601–606
21. Chang Y-H, Hsieh J-W, Kuo T-W (2007) Endurance enhancement of flash-memory storage systems: an efficient static wear leveling design. In: Proceedings of the 44th annual conference on design automation (DAC '07), pp 212–217
22. Wang Y, Liu D, Wang M, Qin Z, Shao Z, Guan Y (2010) RNFTL: a reuse-aware NAND flash translation layer for flash memory. In: Proceedings of the ACM SIGPLAN/SIGBED 2010 conference on languages, compilers, and tools for embedded systems (LCTES '10), pp 163–172
23. Wang Y, Liu D, Qin Z, Shao Z (2011) An endurance-enhanced flash translation layer via reuse for NAND flash memory storage systems. In: Proceedings of the conference on design, automation and test in Europe (DATE '11), pp 1–6

24. Qin Z, Wang Y, Liu D, Shao Z (2011) A two-level caching mechanism for demand-based page-level address mapping in NAND flash memory storage systems. In: Proceedings of the 17th IEEE real-time and embedded technology and applications symposium (RTAS '11), pp 157–166
25. Qin Z, Wang Y, Liu D, Shao Z (2010) Demand-based block-level address mapping in large-scale NAND flash storage systems. In: Proceedings of the eighth IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis (CODES/ISSS '10), pp 173–182
26. Qin Z, Wang Y, Liu D, Shao Z, Guan Y (2011) MNFTL: an efficient flash translation layer for MLC NAND flash memory storage systems. In: Proceedings of the 48th design automation conference (DAC '11), pp 17–22
27. Liu D, Wang Y, Qin Z, Shao Z, Guan Y (2011) A space reuse strategy for flash translation layers in SLC NAND flash memory storage systems. IEEE Trans Very Large Scale Integr (VLSI) Syst 20(6):1094–1107
28. Kim JK, Lee HG, Choi S, Bahng KI (2008) A PRAM and NAND flash hybrid architecture for high-performance embedded storage subsystems. In: Proceedings of the 8th ACM international conference on embedded software (EMSOFT '08), pp 31–40
29. Wang J, Dong X, Xie Y, Jouppi N (2013) i2WAP: improving non-volatile cache lifetime by reducing inter- and intra-set write variations. In: IEEE 19th international symposium on high performance computer architecture (HPCA '13), pp 234–245
30. Joo Y, Niu D, Dong X, Sun G, Chang N, Xie Y (2010) Energy- and endurance-aware design of phase change memory caches. In: Proceedings of the conference on design, automation and test in Europe (DATE '10), pp 136–141
31. Qureshi M, Franceschini M, Lastras-Montano L (2010) Improving read performance of phase change memories via write cancellation and write pausing. In: IEEE 16th international symposium on high performance computer architecture (HPCA '10), pp 1–11
32. Hu J, Xue CJ, Tseng W-C, He Y, Qiu M, Sha EH-M (2010) Reducing write activities on non-volatile memories in embedded CMPs via data migration and recomputation. In: Proceedings of the 47th design automation conference (DAC '10), pp 350–355
33. Hu J, Xue CJ, Zhuge Q, Tseng W-C, Sha EH-M (2013) Write activity reduction on non-volatile main memories for embedded chip multiprocessors. ACM Trans Embed Comput Syst 12(3):77:1–77:27
34. Ferreira A, Childers B, Melhem R, Mosse D, Yousif M (2010) Using PCM in next-generation embedded space applications. In: 2010 16th IEEE real-time and embedded technology and applications symposium (RTAS '10), pp 153–162
35. Akyildiz I, Su W, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. IEEE Commun Mag 40(8):102–114
36. Micron Technology, Inc. (2011) Micron phase change memory. http://www.micron.com/products/pcm/
37. Strukov DB, Snider GS, Stewart DR, Williams RS (2008) The missing memristor found. Nature 453:80–83
38. Zhou P, Zhao B, Yang J, Zhang Y (2009) A durable and energy efficient main memory using phase change memory technology. In: Proceedings of the 36th annual international symposium on computer architecture (ISCA '09), pp 14–23
39. Lee B, Zhou P, Yang J, Zhang Y, Zhao B, Ipek E, Mutlu O, Burger D (2010) Phase-change technology and the future of main memory. IEEE Micro 30(1):143–143
40. Yang B-D, Lee J-E, Kim J-S, Cho J, Lee S-Y, Yu BG (2007) A low power phase-change random access memory using a data-comparison write scheme. In: IEEE international symposium on circuits and systems (ISCAS '07), pp 3014–3017
41. DiskMon for Windows (2006) http://technet.microsoft.com/en-us/sysinternals/bb896646.aspx

Part III
Sensors for Image Capture and Vision Processing

Chapter 7
Artificially Engineered Compound Eye Sensing Systems

Young Min Song, Hyun Gi Park, Gil Ju Lee, and Ju Sung Park

Abstract Understanding of light sensing organs in biology creates opportunities for the development of novel optic systems that cannot be available with existing technologies. The insect's eyes, i.e., compound eyes, are particularly notable for their exceptional interesting optical characteristics, such as wide fields of view (FOV) and nearly infinite depth of field. The construction of man-made sensing systems with these characteristics is of interest due to potential for applications in micro air vehicles (MVAs), security cameras, clinical endoscopes, and new approaches to navigation and sensing. Mimicking of such compound eyes has been evolving for the last few decades, which starts from simple fabrication of stand-alone microlens arrays (MLAs). Recent work has yielded significant progress in the realization of artificial compound eyes, including multiple lens arrays together with sensing pixel arrays with a hemispherical geometry. In this chapter, we discuss a complete set of materials, design layouts, integration schemes, and operating principles for sensing systems that mimic compound eyes. Certain concepts extend recent advances in flexible electronics that provide previously unavailable options in design.

Keywords Compound eyes • Microlens • Image sensors • Photodetectors • Biomimetics

7.1 Introduction

Animal visions play a significant role in the survival through locating food, navigation, and identification of mating suitability. Currently, biological image-capturing optic systems are attracting a great deal of interest among scientists and

Y.M. Song (✉)
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
e-mail: ymsong81@gmail.com

H.G. Park • G.J. Lee • J.S. Park
Department of Electronics Engineering, Pusan National University, 2 Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 609735, South Korea

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_7

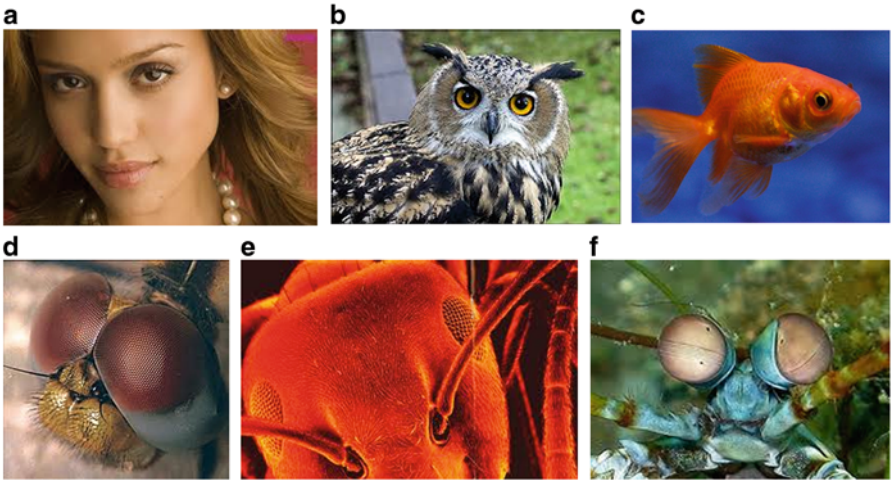


Fig. 7.1 Examples of (**a**–**c**) camera-type eyes and (**d**–**f**) compound eyes in the biological world. Eyes in (**a**) human, (**b**) bird, (**c**) fish, (**d**) fly, (**f**) ant, and (**g**) shrimp

engineers due to their sophisticated structures and functionalities. Biological eyes have shown remarkable structural complexity integrating distinct components across many length scales, nanometer through centimeter. The functionality comes from the organization of functional molecules, cells, and other biomaterials. There are many examples of vision systems that offer high visual acuity, high sensitivity to motion, excellent photosensitivity in low-light environments, wide fields of view (FOV), polarization perception enhancement, aberration correction, and depth of field [1, 2]. Figure 7.1 shows examples of the variety of animal eyes. The diversity of visual systems found in nature provides a variety of design options with desirable operational properties, particularly compared to designs found in conventional imaging system.

Biological eyes can be described, much like conventional imaging technologies, as an integrated set of front-end primary optics (the optical components that collect and direct light) and a back-end processor (e.g., the components that capture and process the collected visual information), as well as additional optical components that serve to improve overall system function. The structures of most animal eyes can be described as a variation of one of approximately eight to ten major eye structures and these are commonly categorized by two major eyes [1]: single lens eyes or compound eyes, as shown in Fig. 7.2. Single lens (or camera-type) eyes have a single integrated lens system, which focuses incoming light onto common photoreceptors in the back-end structure. Compound eyes, on the other hand, have multiple lenses per eye.

The human eye is a single lens imaging system similar to photographic or digital camera systems. The eye consists of a flexible lens for focusing, a variable pupil for fast sensitivity adaptation, and the retina for light detection. The human eye is regarded as a special version of a pinhole camera, where Fresnel diffraction is

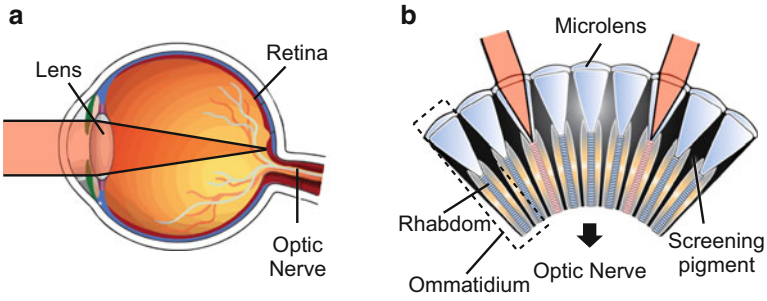


Fig. 7.2 Anatomy of (**a**) a single lens eye and (**b**) a compound eye

compensated by the introduction of a focusing system [3]. The diameter of the pinhole or iris is variable from 1.8 to 8 mm for bright light and dark viewing. The lens of the human eye is spherically overcorrected by graded index effects and aspherical surface profiles. The refractive index of the lens is higher in the central part of the lens; the curvature of the lens becomes weaker toward the edge. This correction offsets the undercorrected spherical aberration of the outer surface of the cornea. Contraction or relaxation of muscles varies the focal length.

The retina contains nerve fibers, light-sensitive rod and cone cells, and a pigment layer. There are about 6 million cones, about 120 million rods, and only about 1 million nerve fibers. The cones of the fovea (center of sharp vision) are 1–1.5 μm in diameter and about 2–2.5 μm apart. The rods are around 2 μm in diameter. In the outer portions of the retina, the sensitive cells are more widely spaced and are multiply connected to nerve fibers (several hundreds to a fiber). The field of vision of an eye approximates an ellipse about 150° high by about 210° wide. The angular resolution or acuity $\Delta\Phi$ is about 0.6–1 min of arc for the fovea [4].

Compound eyes, which can be found in arthropods (i.e., insects and crustaceans), are made up of multiple lenses per eye, while camera-type eyes have a single lens. Figure 7.3 shows five distinct types of compound eyes in nature. These eyes are divided into two main classes: apposition compound eyes (Fig. 7.3a) and superposition compound eyes (Fig. 7.3b–e). In the apposition eye, each microlens-photoreceptor unit (i.e., ommatidium) is optically isolated from its neighbors. Each ommatidium has a single positive microlens producing an image of a relatively large sector of the environment. The rhabdom, which measures the light intensity, has a narrow FOV and its role in imaging resembles a single rod in the camera-type eye. The principle focusing element is the crystalline cone, which is positioned between the microlens and rhabdom. The corneal lens provides only a minor focusing strength. Black screening pigments form opaque walls between adjacent ommatidia to avoid stray light. Apposition eyes have some hundreds up to tens of thousands of these ommatidia packed in nonuniform hexagonal arrays. The distribution of ommatidia on a hemispherical surface allows the apposition compound eyes to have an extremely wide FOV while the total volume consumption keeps small.

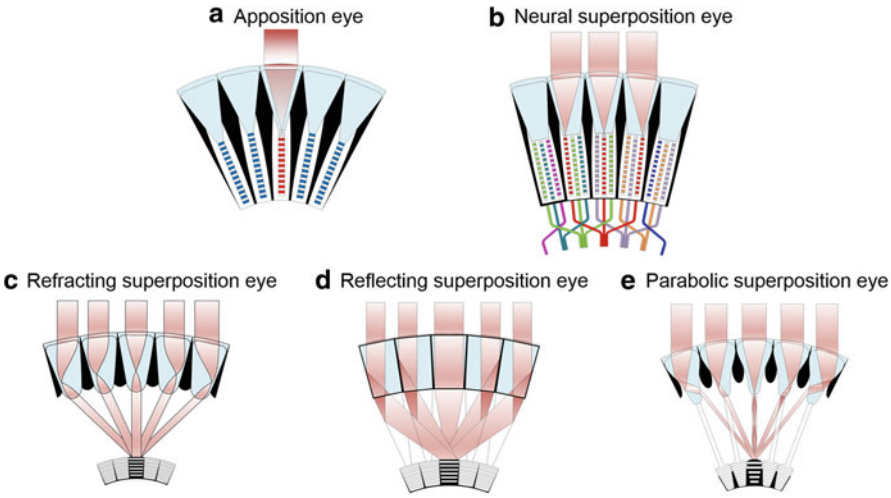


Fig. 7.3 Five different types of compound eyes

The superposition compound eye has optically cooperating ommatidia, so that a bright image is produced by the combined action of many identical units. This type of eyes has primarily evolved on nocturnal insects and deepwater crustaceans. The light from multiple facets combines on the surface of the photoreceptor layer to form a single erect image of the object. In neural superposition eyes (Fig. 7.3b), the rhabdoms are split into seven separate light guides, each with its own optical axis. Both the six summed signals and the seventh single signal produce an overall "neural" image in the eye that does not differ in principle from that of other compound eye. This provides redundant sampling and increased photosensitivity and minimizes the loss of visual acuity.

In refractive superposition eyes (Fig. 7.3c), light is refracted through multiple ommatidial lenses and is focused on a small portion of a common photoreceptor. Between the lenses and photoreceptors, there is a section of unpigmented and transparent cells in order to share photons efficiently from the adjacent ommatidia. The refractive superposition eye demonstrates improved photosensitivity by as much as three orders of magnitude compared to simple eyes. The number of arthropod groups with refracting superposition eyes is large. Moths, beetles, and crustaceans such as krill are on the list. Another variant is the reflecting superposition eyes (Fig. 7.3d) found in decapod crustaceans such as shrimp and lobster. The facets of reflecting compound eyes have long rectangular walls, which act as biological mirrors that reflect light to the retina. In these cases, a superposition image can be formed without any lenses at all. The last example is parabolic superposition eyes (Fig. 7.3e) found in many crabs and a few mayflies. The structure involves ordinary lenses, cylindrical lenses, parabolic mirrors, and light guides. The mechanism is the most complicated and relies on both reflection and refraction. Detailed optical systems and operating principles can be found in other literature [1, 2].

7.2 Artificial Compound Eyes

7.2.1 *Planar-Type Compound Eye Imaging Systems*

Artificial implementation of compound eyes has attracted a great deal of research interest due to their exceptionally wide FOV, high sensitive to motion, and nearly infinite depth of field, which exhibits a huge potential for medical, industrial, and military applications. The use of miniaturized, arrayed optical components fabricated by using semiconductor planar processing technologies has been proposed to mimic the natural compound eyes. Since the commercially available image sensors such as CCD or CMOS are fabricated on planar wafers, a thin monolithic objective based on the compound eye concept has to be a planar structure as well.

Various technical approaches were reported for planar-type compound eye imaging systems. Duparre et al. reported an artificial apposition compound eye, which consists of a microlens array (MLA) positioned on a substrate, preferably with optical isolation of the channels, and optoelectronic detector arrays (Fig. 7.4) [5]. The MLAs have a diameter D, focal length f, and pitch p_L, as illustrated in Fig. 7.4b. Pinhole arrays with a diameter d and pitch p_P are positioned in the microlenses' focal plane on the spacing structure's backside. The pitch difference enables the different viewing directions of each optical channel. Each channel's optical axis points in a different direction in object space with the optical axes of the channels directed outward if the pitch of the receptor array is smaller than that of the MLA. If the pitch of the MLA is smaller than that of the receptor array, the image is inverted. A pinhole array can be used to narrow the photosensitive area of the detector pixels if they are not small enough for the required resolution.

Tanida et al. proposed the concept of thin observation module by bound optics (TOMBO) inspired by compound eyes [6]. In the TOMBO system, only a small

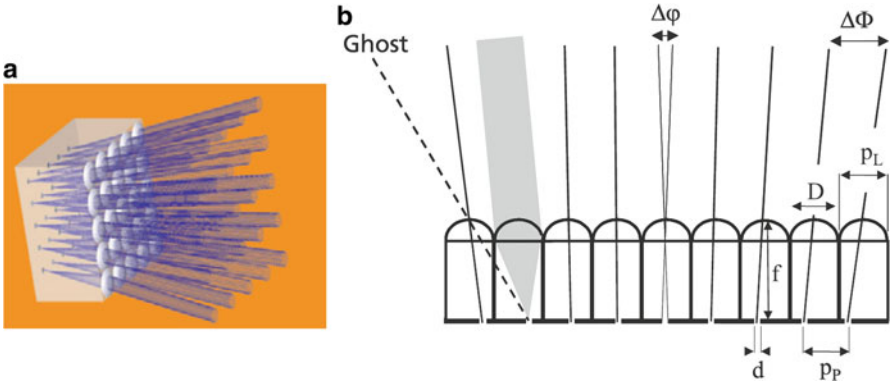


Fig. 7.4 Planar artificial apposition compound eye. (**a**) Three-dimensional model of the artificial apposition compound eye showing the focusing microlens array (MLA). (**b**) Cross-sectional view of (**a**) with important parameters. (Reproduced by permission of the International Society for Optics and Photonics [5])

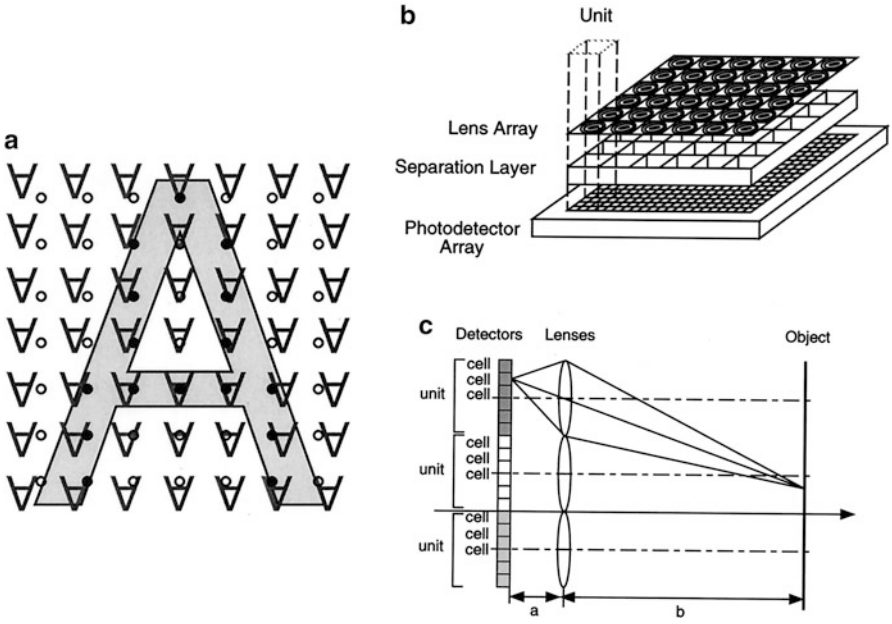


Fig. 7.5 (**a**) Erect image retrieved by sampling of multiple images. (**b**) TOMBO system structure and (**c**) optical system. (Reproduced by permission of The Optical Society of America [6])

number of channels are used, but each cell of the system has a matrix of associated photoreceptors which pick up all the information of the microimages. The microlenses are centered with the matrix of photodetectors capturing the microimages (Fig. 7.5a). The difference in the microimages is the result of the different radial positions of the corresponding channel within the array. For close objects, the information content of the overall image calculated from all the microimages is much larger than that of the single microimages (Fig. 7.5b, c). However, there is a burden of extraordinary image processing. Other imaging types based on microelectromechanical systems (MEMS) technology or an array of elementary motion detectors (EMDs) in a radial arrangement were also reported [7, 8].

In a more complex approach, mimicking a superposition compound eye, a stack of three MLAs, as shown in Fig. 7.6, was used [9]. The lenses of the first array image demagnify and invert subimages of the object into the plane of the second lens array. The lenses of the third array image these inverted subimages to the image sensor plane. The lenses of the second array serve as field lenses, imaging the pupil of the first array to the corresponding entrance pupils of the third lens array. As shown in Fig. 7.6a, size, focal length, and distances in the system are optimized to obtain correct superposition of the individual subimages in the plane of the image sensor. The parallel transfer of different parts of an overall FOV with strong demagnification by separated optical channels allows the superposition eye to have a collective space bandwidth product (SBP) which is equal to the sum of the

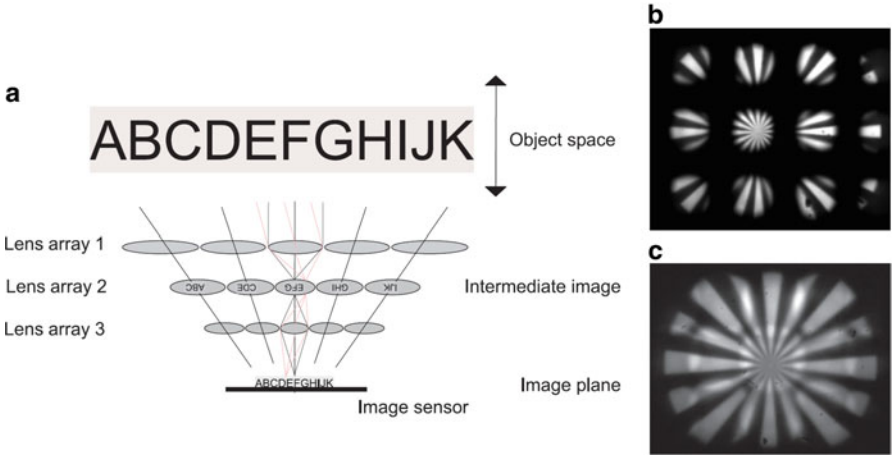


Fig. 7.6 (**a**) Superposition-type compound eye imaging systems consisting of three layers of MLAs. (**b**) Image of a radial star test pattern at an object distance of 41 cm from the compound eye. (**c**) Image captured at the image plane 120 μm further away from the compound eye. (Reproduced by permission of The Optical Society of America [9])

individual channel's ones. Consequently, the superposition eye has the potential for much higher resolution than the experimentally demonstrated artificial apposition compound eyes.

Figure 7.6b, c shows the images of a radial star pattern, captured at different axial positions from the superposition eye. It can be observed that the matching of the image plane of the individual telescopes with the position of the perfect annexation of the partial images is particularly critical. This is mainly influenced by the correspondence of the axial position of the intermediate images with the position of the field apertures. It is demonstrated that one overall image is generated by the transfer of different image sections through separated channels with a strong demagnification. One of the drawbacks is the complexity of lens configuration, which is much higher as compared with the apposition compound eye. Future applications are, for example, large-object field microscopes. If it becomes possible to build ultrathin and flexible cameras, a large range of possible products comes into sight.

7.2.2 Hemispherical Compound Eye Lens Arrays and Optic Components

Since the abovementioned imaging systems were fabricated on the planar substrates, achieving a wide FOV in those structures has been hindered mainly due to the inherent flatness of the arrayed optical components. Several research groups have tried to demonstrate MLAs on hemispherical or curvy surfaces. One of the

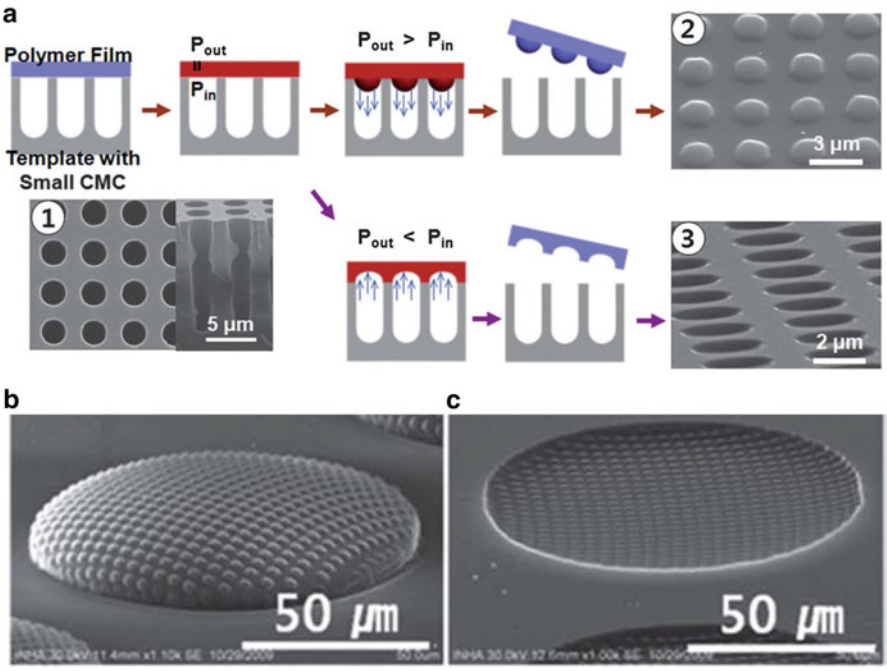


Fig. 7.7 (**a**) The fabrication process of the MLA. Top and side views of the silicon template used to make the MLA (1), oblique scanning electron microscope images of convex MLA (2) and concave MLA (3). (**b**–**c**) Oblique SEM images of convex-on-convex (**b**) and convex-on-concave (**c**). (Reproduced by permission of The Royal Society of Chemistry [10])

techniques to create compound eye lens structures is the fabrication of lens-on-lens arrays (Fig. 7.7) [10]. Lens-on-lens arrays are small concave polymer structures that sit on top of larger concave structures. To start, cylindrical microchannels with a diameter of 2–3 μm are formed in a silicon substrate, as depicted in Fig. 7.7a. A polycarbonate film was sealed to the microchannel array and heated, and a pressure difference between the inside and outside of the microchannels was applied. This process was then repeated with a larger set of microchannels to complete lens-on-lens arrays. The resulting lens-on-lens array can be fabricated with any combination of concave and convex lenses as well. Figure 7.7b, c shows the fabricated examples of lens-on-lens arrays with shapes of (b) convex-on-convex and (c) convex-on-concave. The organization, dimensions, and height of the small and large lenses can be controlled through the fabrication of the template.

Another approach uses irreversible thermomechanical deformation to form a curved MLA [11]. A planar array of concave silica microlenses was fabricated using femtosecond laser pulses followed by HF treatment, as shown in Fig. 7.8. Silica was then used as a mold to form a planar array of convex lenses out of polymethyl methacrylate. The array was then heated and bent around a glass

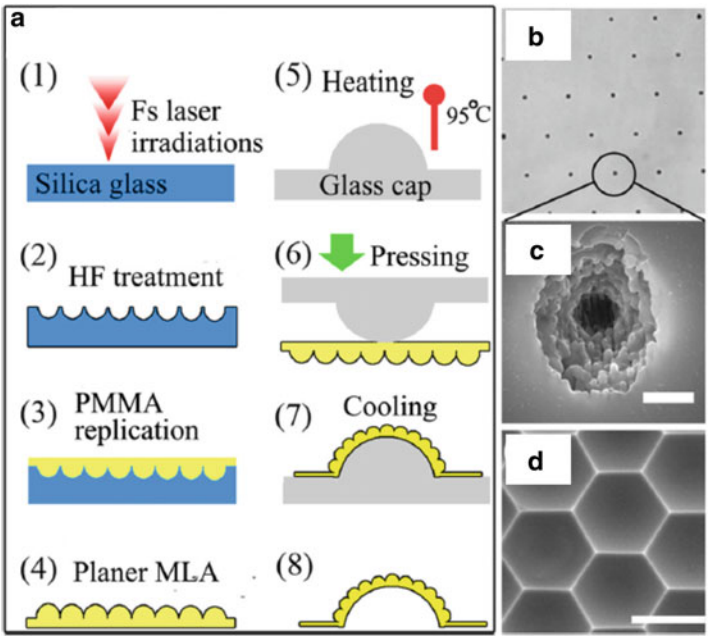


Fig. 7.8 (**a**) Fabrication process steps for MLAs on a hemispherical surface. (**b**) Optical image of laser exposure spots on glass substrate. (**c**) SEM image of single exposure spot. (**d**) SEM image of the morphology of lens array template. (Reproduced by permission of the American Institute of Physics [11])

hemisphere. The resulting array demonstrates a 140° FOV, which is significant improvement over planar lens arrays.

More advanced example is a hemispherical polymer dome with a set of artificial ommatidia, which consists of a refractive polymer microlens, a light-guiding polymer cone, and a self-aligned waveguide to collect light with a small angular acceptance [12]. The ommatidia are omnidirectionally arranged along a hemispherical surface such that they provide a wide FOV similar to that of natural compound eyes. The spherical configuration of the microlenses is achieved by reconfigurable microtemplating that enables polymer replication using the deformed elastomer membrane with microlens patterns. More importantly, the formation of polymer waveguides and cuvette-shaped cones, those are self-aligned with microlenses, is also realized by a self-writing process in a photosensitive polymer resin. These 3D polymer optical systems have the potential for a broad range of optical applications, such as data storage, medical diagnostics, surveillance imaging, and light-field photography.

The compound eyes in nocturnal insects present a fascinating object for biomimetic studies due to their well-organized hierarchical structures, consisting of subwavelength structures (SWSs) with a tapered profile on the MLAs [13]. The former acts as a homogeneous medium with a graded refractive index to reduce

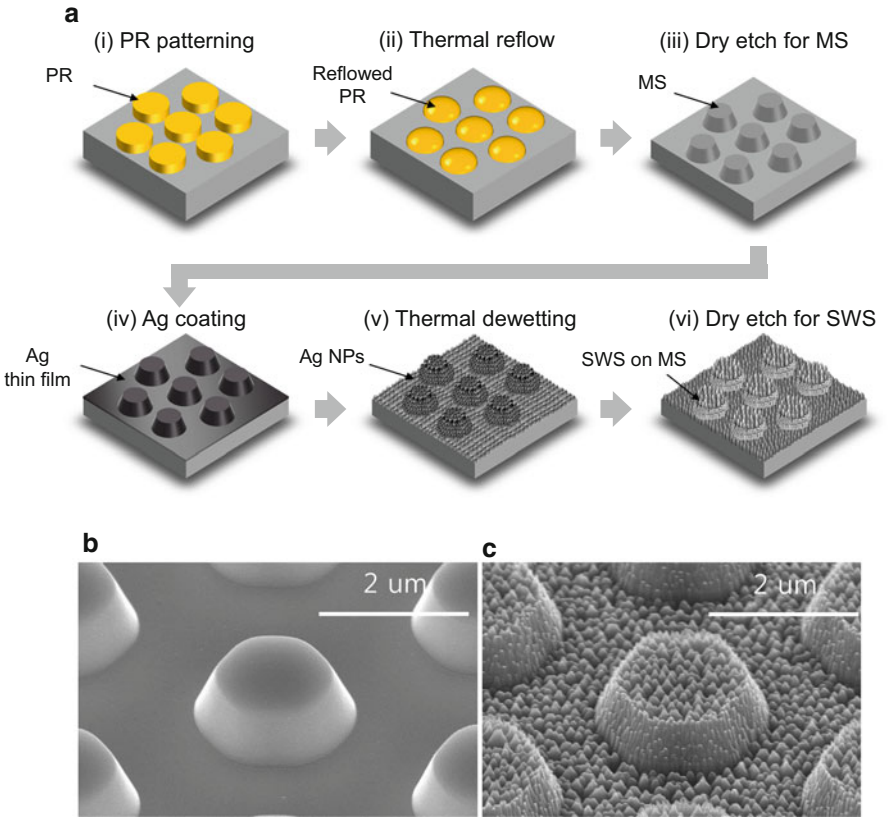


Fig. 7.9 (**a**) Fabrication procedure for the subwavelength structure (SWS)/MS architecture. (**b**–**c**) Tilted-angle view of SEM images for the fabricated sample with (**b**) an MS and (**c**) an SWS/MS on a gallium phosphide (GaP) substrate. (Reproduced by permission of The Optical Society of America [13])

Fresnel reflection at the surface, and the latter focuses the incident light toward the photoreceptor cells. The combination of micro- and nanostructures also shows antifogging effects as well as efficient light collecting properties. Song et al. demonstrated integrated ommatidium-like structures on the semiconductor materials to enhance the optical efficiency in the optoelectronic devices. The fabrication procedure for these structures is depicted in Fig. 7.9a. First, hexagonally patterned microstructures (MSs) were fabricated on a gallium phosphide (GaP) substrate by a dry etch process using thermally reflowed photoresist (PR) masks (Fig. 7.9b). For SWS fabrication, Ag nanoparticles were grown on the entire surface by a thermal dewetting of Ag thin films. These nanoparticles were used as an etch mask to define tapered nanostructure on the micro-patterned substrate. The final nanostructures were randomly distributed on the MS arrays, with an average distance of 150 nm and a height of 120 nm (Fig. 7.9c). These structures can be formed on other substrates to enhance the optical efficiency of various optoelectronic devices [14, 15].

7.2.3 *Curved Image Sensors and Hemispherical Compound Eye Imaging Systems*

There also have been advanced researches beyond fabricating stand-alone lens arrays to create complete imaging systems that integrating lenses, photodetectors, and other electrical and/or optical components. Creating curved photodetector arrays is an extremely significant challenge to the demonstration of bioinspired vision systems, including single lens and compound eye systems. Recently, researchers at the University of Michigan have developed the organic photodetctor focal plane arrays (FPAs) on a three dimensionally curved surface that mimics the size, function, and architecture of the human eye [16]. To create an imaging system mimicking the natural eye requires the fabrication of photodetector arrays onto a curved surface that matches the curvature of a single lens. They overcome these challenges by direct transfer technique employing elastomeric stamps and cold welding (Fig. 7.10a), which dramatically simplify lens design without degrading the FOV, focal area, illumination uniformity, and image quality.

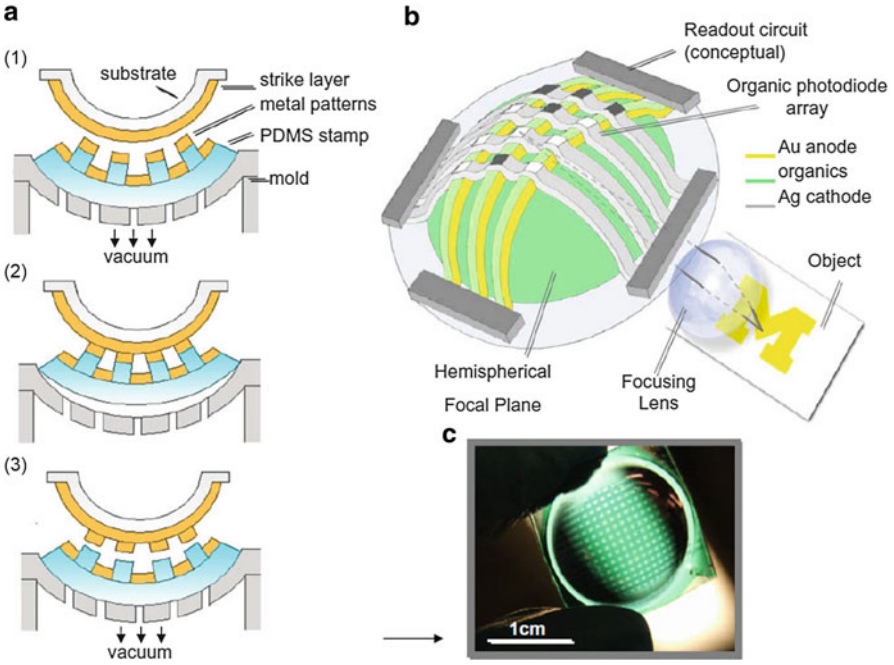


Fig. 7.10 (**a**) Hemispherical organic photodetector focal plane array (FPA) and its incorporation into a simple imaging system. (**b**) Photograph of a completed hemispherical focal plane with an 11×13 array of photodetectors on a 1 cm radius plastic hemisphere. (**c**) Process sequence for fabricating a hemispherical FPA. (Reproduced by permission of the Elsevier publishing group [16])

Figure 7.10b shows the schematic of the completed imaging system and its incorporation into a simple imaging system. The passive matrix FPA consists of two perpendicular electrode stripe arrays placed above and below continuous layers of organic semiconductor materials forming the active photodetection regions. Individual photodetectors are defined at crossings of the upper and lower stripes, where device readout is realized by probing the appropriate row and column electrodes. Figure 7.10c shows an example of a completed hemispherical FPA consisting of double heterojunction photodiodes. The direct material transfer technique avoids introduction of excessive strain into heterogeneous material layers, thus allowing for the fabrication of high performance organic electronic devices with micrometer scale dimensions on curved surfaces with radii 1 cm or less.

Similar imaging systems were also reported from University of Illinois at Urbana Champaign [17, 18]. The curved photodetector arrays, for demonstrating artificial human eye camera, were fabricated using conventional planar fabrication methods, but it was transferred onto the hemispherical substrate by using a transfer printing technique. A 16×16 array of photodetectors, blocking diodes, and flexible metal interconnects was fabricated on a silicon on insulator (SOI) wafer, and then transferred onto a radially stretched PDMS film. After releasing the PDMS, hemispherical FPAs can be achieved. Transfer printing onto a matching hemispherical glass substrate with an adhesive, adding a hemispherical cap with integrating lens and interfacing to external control box complete the human eye camera system.

Research on curved image sensor was then extended to the electronic eye camera system with adjustable zoom capability [19]. Key idea is the use of both tunable lens and tunable detector arrays. Figure 7.11a shows a schematic illustration of such camera, including the tunable lens (upper) and tunable detector (lower) modules. The lens consists of a fluid-filled gap between a thin PDMS membrane and a glass window, to form a plano-convex lens with 9 mm diameter and radius of curvature

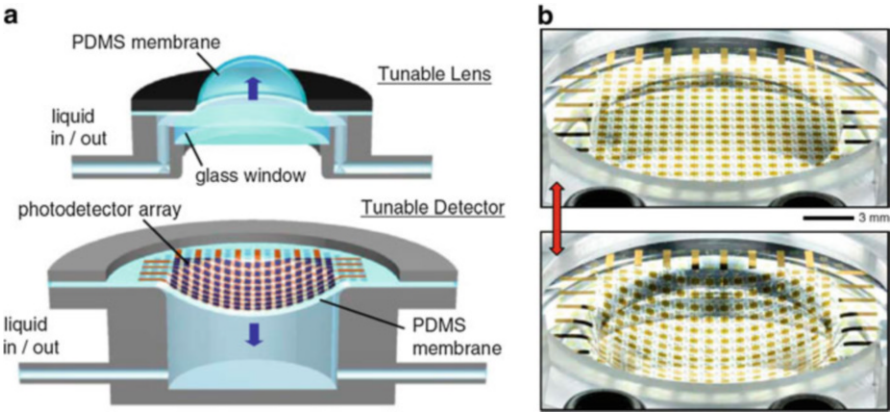


Fig. 7.11 (**a**) Hemispherical image sensor with a tunable lens (*top*) and a tunable detector array (*bottom*). (**b**) Tilted view of the photodetector array in flat (*top*) and deformed (*bottom*) state. (Reproduced by permission of Proceedings of the National Academy of Sciences [19])

that is adjustable with fluid pressure. The tunable detector consists of an array of interconnected silicon photodiodes and blocking diodes (16×16 pixels) mounted on a thin PDMS membrane, in a mechanically optimized, open mesh serpentine design. This detector sheet mounts on a fluid-filled cavity; controlling the pressure deforms the sheet into concave or convex hemispherical shapes with well-defined, tunable levels of curvature.

Figure 7.11b shows tilted views of a photodetector array in its initial, flat configuration without applied pressure (upper frame) and in a concave shape induced by extracting liquid out of the chamber with negative applied pressure (lower frame). The PDMS film is elastomeric and reversibly deformable, creating a flexible optoelectronics array capable of high strain without loss of performance. Narrow metal lines encapsulated with thin films of polyimide on top and bottom provide ribbon-type interconnects between unit cells, in a neutral mechanical plane that isolates the metal from bending induced strains. These features enable the photodetector array to accommodate large strains during deformation by hydraulic tuning systems. The photodetector surface deforms to a hemispherical shape due to water extraction, which implies a uniform spacing between photodetectors. Certain concepts extend recent advances in stretchable electronics that is the technology for building electronic circuits, being made stretchable, rather than flexible.

A challenge in building digital cameras with the hemispherical, compound apposition layouts of arthropod eyes is that essential design requirements cannot be met with existing planar sensor technologies or conventional optics. Recently, as a natural extension of recent advances in stretchable electronics and hemispherical photodetector arrays, Song et al. have demonstrated arthropod-inspired cameras with nearly full hemispherical shape with 160° FOV [20]. Their surfaces are densely populated by imaging elements (i.e., artificial ommatidia), which are comparable in number (180) to those of the eyes of fire ants and dark beetles. The devices combine elastomeric compound optical elements with deformable arrays of thin silicon photodetectors into integrated sheets that can be elastically transformed from the planar geometries to hemispherical shapes for integration into apposition cameras.

Figure 7.12a shows illustrations of an array of elastomeric microlenses and supporting posts joined by a base membrane (above) and a corresponding collection of silicon photodiodes and blocking diodes interconnected by filamentary serpentine wires and configured for matrix addressing (below). On the left in Fig. 7.12a, these two subsystems are shown in planar geometries. Aligned bonding of these two subsystems places each photodiode at the focal position of a corresponding microlens to yield integrated imaging systems. Similar to the concept of tunable cameras, hydraulic actuation can deterministically transform the planar layout into a full hemispherical shape without any change in optical alignment or adverse effect on electrical or optical performance. Figure 7.12b shows an image of a representative system after hemispherical deformation. A complete apposition camera consists of this imager, combined with a perforated sheet of black silicon in order to prevent stray light (Fig. 7.12c). By analogy to natural compound eyes, each microlens, supporting post, and photodiode corresponds to a corneal lens,

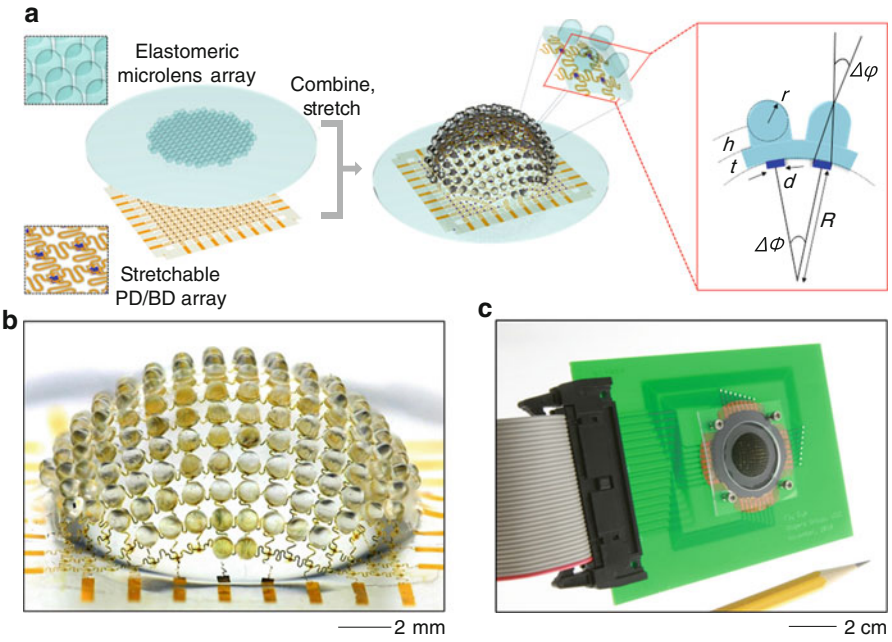


Fig. 7.12 (**a**) Fabrication methods for hemispherical compound eye camera consisting of optical and electrical subsystems. (**b**) Optical micrograph of the integrated forms of lens/detector arrays after deformation. (**c**) Photograph of the completed camera mounted on a printed circuit board. (Reproduced by permission of Nature Publishing Group [20])

crystalline cone, and rhabdom, respectively. The black perforated sheet serves as the black screening pigment, which can be found in those apposition-type compound eyes. These device configurations seem to be applicable to other types of compound eyes, such as refracting/reflecting superposition eyes and neural superposition eyes.

Figure 7.13a depicts operating principles of a hemispherical, apposition compound eye camera through quantitative ray-tracing results for the simple case of an 8×8 ommatidia. Each microlens generates an image of the object with characteristics determined by the viewing angle. Overlap of a portion of each image with the active area of a corresponding photodiode generates a photocurrent at this location of the array. Improved resolution can be realized by scanning the camera, as shown in the left frame of Fig. 7.13a. Figure 7.13b presents pictures, rendered on hemispherical surfaces, of soccer ball patterns captured at three different polar angles (i.e., $-50°$, $0°$, and $50°$) relative to the center of the camera. This is the first demonstration of compound eye imaging, which enables exceptionally wide-angle FOV, without off-axis aberrations. All three images reveal comparable clarity without blurring or aberrations. The researchers also presented the imaging results showing the nearly infinite depth of field, which results from the short focal length of microlens and the nature of image formation in compound eyes.

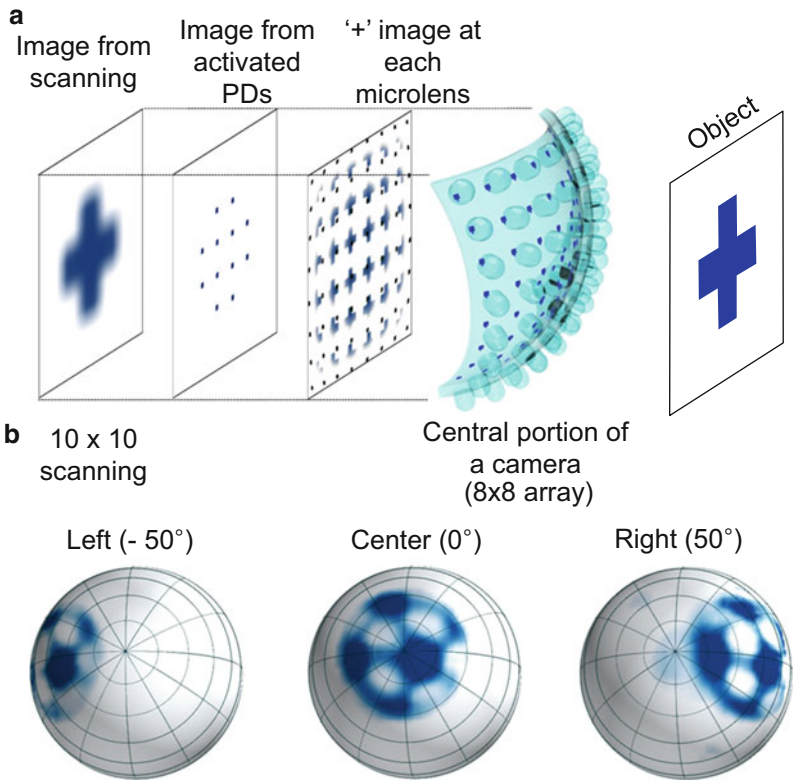


Fig. 7.13 (**a**) Conceptual view of image formation of hemispherical, apposition-type compound eye camera. (**b**) Pictures of a soccer ball captured at three different polar angles, i.e., −50°, 0°, and 50°, respectively. Images are rendered on the hemispherical surface. (Reproduced by permission of Nature Publishing Group [20])

One of the drawbacks of such compound eye imagers is inferior spatial resolution compared to that of camera-type eyes. In order to overcome these limitations, Lee et al. recently proposed COMPUtational compound EYE (COMPU-EYE), a new compound eye design that increases acceptance angles and uses a modern digital signal processing technique [21]. The proposed COMPU-EYE yielded a fourfold improvement in spatial resolution. Another powerful mode of such system is that each area is observed by multiple ommatidia. Thus, in the COMPU-EYE systems, damaged or disjointed ommatidia do not have a significant influence on the overall observation.

Another type of artificial compound eyes, featuring a panoramic FOV in a very thin package, has been developed through the CurvACE (Curved Artificial Compound Eyes) research project in the European Union [22]. Figure 7.14 shows schematic illustration of "CurvACE" design and assembly. In this approach, three separate planar array layers, i.e., microlenses, photodetectors, and electromechanical interconnects, were fabricated and integrated into a curved optical system. The

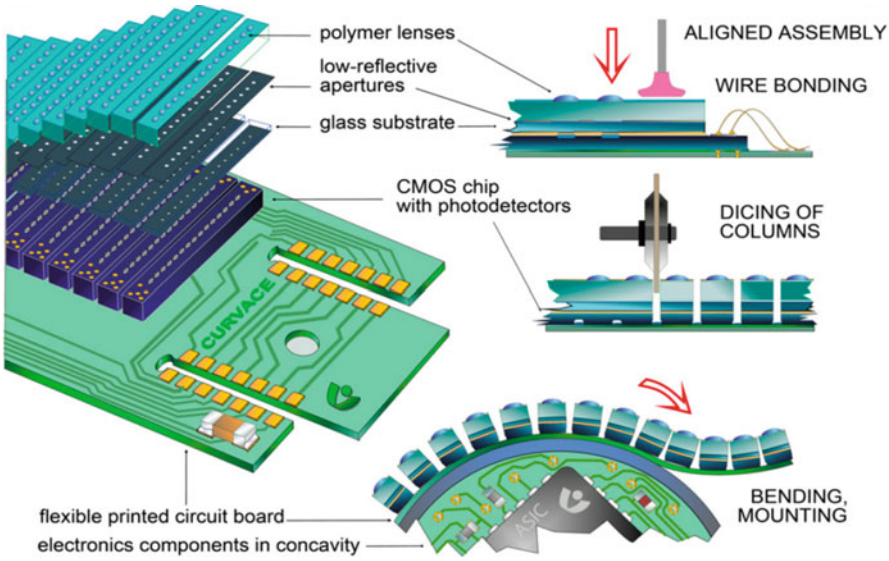


Fig. 7.14 Fabrication procedure for "CurvACE (Curved Artificial Compound Eye)" system. (Reproduced by permission of Proceedings of the National Academy of Sciences [22])

MLAs are molded on a glass carrier, which focuses light precisely onto the sensitive areas of a silicon-based photodetector layer. This layer contains an array of analog very-large-scale integration (VLSI) photodetectors as well as additional circuitry for signal processing. A flexible electromechanical interconnection layer, formed by a polyimide printed circuit board, physically supports the ensemble and transfers the output signals from the individual ommatidia to the processing units. These three individual layers were then aligned and integrated. The ommatidial layers are cut down to the bottom layer using a high-precision dicing. Since the interconnect layer is flexible, this step allows the entire array to be flexible and bendable.

The final imager, shown in Fig. 7.14a, is light and exhibits good hemispherical FOV ($180^\circ \times 60^\circ$, depicted in Fig. 7.14b), high temperature resistance, and local adaptation to illumination. The fabricated prototype features several similarities with the eye of the fruit fly Drosophila in terms of spatial resolution, acceptance angle, the number of ommatidia, local light adaptation, crosstalk prevention, and signal acquisition bandwidth. A fully spherical CurvACE could be realized by fabricating and individually bending several ommatidial arrays with one ommatidium per column along the meridians of a sphere to measure optic flow omnidirectionally (Fig. 7.15).

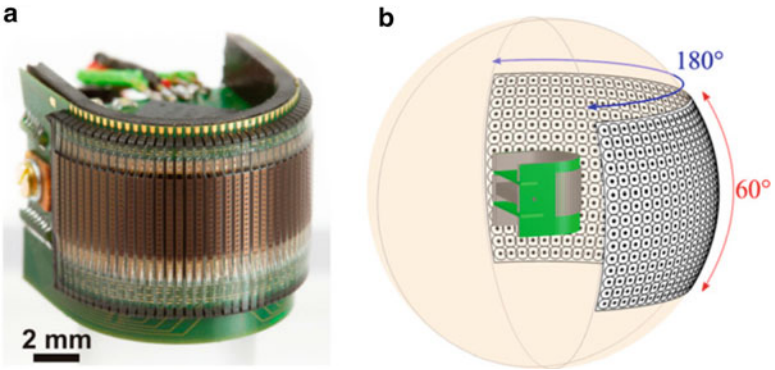


Fig. 7.15 (**a**) Representative "CurvACE" prototype. (**b**) Illustration of the panoramic fields of view (FOV) of the fabricated imaging system. (Reproduced by permission of Proceedings of the National Academy of Sciences [22])

7.3 Summary

In this chapter, we reviewed recent approaches for mimicking compound eye imaging systems, offering the wide FOV, nearly infinite depth of field, and fast response to motion. We first summarized basic anatomies and operating principles for five different types of compound eyes (i.e., one apposition type and four superposition types) found in arthropod's eyes. We categorized the artificial compound eyes into three different types: planar-type compound eyes, hemispherical compound eye optics (without imager), and hemispherical compound eyes. Examples in three different types of artificial compound eyes illustrated the rapid progress of these concepts. In particular, it is shown that recent advances in flexible electronics enabled a realization of a complete set of well-matured, hemispherical compound eye imagers that incorporate all of the functional organs, such as cornea, crystalline cones, rhabdom, and screening pigments. Imaging characteristics of these compound eye sensors/imagers also illustrated the powerful mode of compound eye sensing concepts. The size, resolution, and optical efficiency of artificial compound eyes are still not on the level of commercial product. However, future works will provide brilliant ideas for cost-effective mass production of high resolution, small form package of these imaging systems with enhanced optical efficiency.

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

References

1. Land MF, Nilsson D-E (2002) Animal eyes. Oxford University Press, Oxford
2. Warrant E et al (2006) Invertebrate vision. Cambridge University Press, New York
3. Hoffman C (1980) Die optische Abbildung. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig
4. Adelson EH, Bergen JR (1991) The plenoptic function and the elements of early vision. In: Landy M, Movshon J (eds) Computational model of visual processing. MIT Press, Cambridge, p 3
5. Duparre JW et al (2006) Novel optics/micro-optics for miniature imaging systems. Proc. SPIE 6196, photonics in multimedia, 619607, Apr 2006
6. Tanida J et al (2001) Thin observation module by bound optics (TOMBO): concept and experimental verification. Appl Opt 40(11):1806
7. Hoshino K et al (2001) A one-chip scanning retina with an integrated micromechanical scanning actuator. J MEMS 10:492
8. Franceschini N et al (1992) From insect vision to robot vision. Philos Trans R Soc B 337:283
9. Duparre JW et al (2005) Microoptical telescope compound eye. Opt Express 13(3):889
10. Park BG et al (2012) Micro lens-on-lens array. Soft Matter 8:1751
11. Liu H et al (2012) Fabrication of bioinspired omnidirectional and gapless microlens array for wide field-of-view detections. Appl Phys Lett 100:133701
12. Jeong KH et al (2006) Biologically inspired artificial compound eyes. Science 312(5773):557
13. Song YM et al (2011) Multifunctional light escaping architecture inspired by compound eye surface structure: from understanding to experimental demonstration. Opt Express 19:A157
14. Leem JW et al (2013) Biomimetic artificial Si compound eye surface structures with broadband and wide-angle antireflection properties for Si-based optoelectronic applications. Nanoscale 5:10455
15. Kang EK et al (2015) Improved light extraction efficiency of GaN-based vertical LEDs using hierarchical micro/subwavelength structures. Jpn J Appl Phys 54:06FH02
16. Xu X et al (2008) Direct transfer patterning on three dimensionally deformed surfaces at micrometer resolutions and its application to hemispherical focal plane detector arrays. Org Electron 9:1122
17. Ko HC et al (2008) A hemispherical electronic eye camera based on compressible silicon optoelectronics. Nature 454:748
18. Jung I et al (2010) Paraboloid electronic eye cameras using deformable arrays of photodetectors in hexagonal mesh layouts. Appl Phys Lett 96:021110
19. Jung I et al (2011) Dynamically tunable hemispherical electronic eye camera system with adjustable zoom capability. Proc Natl Acad Sci U S A 108(5):1788
20. Song YM et al (2013) Digital cameras with designs inspired by the arthropod eye. Nature 497:95
21. Lee W-B et al (2016) COMPU-EYE: a high resolution computational compound eye. Opt Express 24(3):2013
22. Floreano D et al (2013) Miniature curved artificial compound eyes. PNAS 110(23):9267

Chapter 8
Intelligent Vision Processing Technology for Advanced Driver Assistance Systems

Po-Chun Shen, Kuan-Hung Chen, Jui-Sheng Lee, Guan-Yu Chen, Yi-Ting Lin, Bing-Yang Cheng, Guo-An Jian, Hsiu-Cheng Chang, Wei-Ming Lu, and Jiun-In Guo

Abstract Intelligent vision processing technology has a wide range of applications on vehicles. Many of these applications are related to a so-called Advanced Driver Assistance System (ADAS). Collaborated with cameras, Pedestrian and Motorcyclist Detection System (PMD), Lane Departure Warning System (LDWS), Forward Collision Warning System (FCWS), Speed Limit Detection System (SLDS), and Dynamic Local Contrast Enhancement (DLCE) techniques can help drivers notice important events or objects around. This chapter gives an in-depth exploration for these intelligent vision processing technologies from the viewpoints of methodology development, algorithm optimization, and system implementation on embedded platforms. More precisely, this chapter tends to first give a survey and overview for newly appeared state-of-the-art intelligent vision processing technologies for ADAS, and then highlights some significant technologies including PMD, LDWS, FCWS, SLDS, and DLCE developed in System on Chip (SoC) Laboratory, Fong-Chia University, Taiwan, and intelligent Vision System (iVS) Laboratory, National Chiao Tung University, Taiwan. Besides, implementation and verification of the above ADAS technologies will also be presented. In summary, the proposed PMD design achieves 32.5 frame per second (fps) for 720×480 (D1) resolution on an AMD A10-7850K processor by using heterogeneous computing. On an automotive-grade Freescale i.MX6 (including 4-core ARM Cortex A9, 1 GB DDR3 RAM, and Linux environment) platform, the proposed LDWS, FCWS, and SLDS designs, respectively, achieve 33 fps, 32 fps, and 30 fps for D1 resolution. Finally, the proposed DLCE system is realized on a TREK-668 platform with an Intel Atom 1.6 GHz processor for real-time requirement of 50 fps at D1 resolution.

P.-C. Shen • J.-S. Lee • G.-Y. Chen • Y.-T. Lin • B.-Y. Cheng • G.-A. Jian
H.-C. Chang • J.-I. Guo
Department of Electronics Engineering, National Chiao Tung University, Hsinchu City, Taiwan

K.-H. Chen (✉) • W.-M. Lu
Department of Electronics Engineering, Feng Chia University, Taichung City, Taiwan
e-mail: kuanhung@fcu.edu.tw

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_8

Keywords ADAS • FCWS • Intelligent vision processing • LDWS • Pedestrian detection

8.1 Introduction

According to the survey report of Gartner in 2015, autonomous vehicles attract the highest expectation among many other popular applications. Autonomous cars adopt sensors such as RAdio Detection And Ranging (RADAR), Light Detection And Ranging (LiDAR), and cameras to understand environment around. Among these sensors, cameras are rather inexpensive in cost consideration and mature in manufacturing aspect. However, we need an elaborated intelligent processing system to analyze the visual contents to construct sensing ability for autonomous cars. In addition, design trend for car safety moves from passive ways to active ones. In the USA, rear view monitoring becomes a standard equipment for new cars. Car manufacturers such as BMW, Lexus, and Infiniti have launched Around View Monitoring (AVM) adoption. In EU, Lane Departure Warning System (LDWS) is already a standard equipment for vehicles. Besides, car manufacturers such as BMW, Audi, Volvo, and Mercedes-Benz provide advanced options, e.g., Adaptive Cruise Control (ACC), Adaptive Front lighting System (AFS), Driver Status Monitoring (DSM), Blind Spot Detection (BSD), and so on, to customers. Moreover, search engine vendor Google also devotes to develop self-driving cars. Google self-driving cars adopt a LiDAR sensor for detecting objects around from a 3D-space viewpoint and millimeter wave RADAR sensors for distant object detection. As for pedestrian and bicycle detection, vision technology is adopted. A car electronics vendor worth mentioning is Mobileye, which provides vision Systems on Chip (SoCs) for safety driving to car manufacturers. Based on solid foundation of ARM and digital signal processor (DSP) experience, TI has also launched TDA2x/3x SoCs for Advanced Driver Assistance Systems (ADAS). These above facts reveal that intelligent vision processing technology for ADAS attracts high attention recently and will probably become critical in the upcoming 5–10 years. Accordingly, this chapter tends to give a survey and overview for newly appeared intelligent vision processing technology for ADAS in Sect. 8.2, and highlights some significant technologies including Pedestrian and Motorcyclist Detection System (PMD), LDWS, Forward Collision Warning System (FCWS), Speed Limit Detection System (SLDS), and Dynamic Local Contrast Enhancement (DLCE) developed in our laboratory in Sect. 8.3. Besides, implementation and verification of the above ADAS technology are also presented in Sect. 8.4. Finally, we end this chapter in Sect. 8.5, i.e., the conclusion.

8.2 Existing ADAS Systems

Machine learning leads the trend to support multiple-object detecting and distinguishing. How to generate good machine learning samples and simplify the complex architecture are still challenges to real-time processing with limit computing resource. Certain unique features are extracted to recognize one kind of object, which reduces the computation complexity with minor sacrifice of detection diversity. With inclement weather against techniques, ADAS systems perform even better. This section reviews the related works of intelligent vision processing technology for ADAS which can be categorized into PMD, LDWS, FCWS, SLDS, and DLCE in the following sub-sections.

8.2.1 Pedestrian and Motorcyclist Detection System

Various techniques have been developed for detecting moving objects commonly seen on the road, including pure pedestrians [1–21], pure vehicles [22–29], and multiple kinds of objects [30, 31]. Enhancing the detection rate and lowering down the false alarm rate are the two main objectives of all existing vision-based objects detection designs. The designs [30, 31] are dedicated on multiple kinds of objects detection, while the remainders focused on single kind of object, either pedestrian or motorcyclist detection. All of these existing designs are verified by using software models only and simulated on personal computers. One of the contributions of this work is to present the experience of implementing a vision-based multiple moving objects detection system on a portable platform. Matching models are required to enable machines to detect objects in images. These matching models can be approximately classified into two types, i.e., with either global features or local features. Features of the interested objects are extracted to train the classifier. Using global features of the objects is beneficial to achieving high detection rate, while using local features of the objects can solve the occlusion problems.

In [1], the key insight is that one may compute finely sampled feature pyramids at a fraction of the cost, without sacrificing performance, i.e., features computed at octave-spaced scale intervals are sufficient to approximate features on a finely sampled pyramid for a broad family of features. Extrapolation is inexpensive as compared to direct feature computation. The work [2] presented a spatialized random forest (SRF) approach, which can encode an unlimited length of high-order local spatial contexts. By spatially random neighbor selection and random histogram-bin partition during the tree construction, the SRF can explore much more complicated and informative local spatial patterns in a randomized manner. In [3], the authors had evaluated their system on a data set specifically for pedestrian detection from a moving vehicle, and they have shown that it is able to outperform other fast detection methods in both speed and accuracy. This is due to: (1) the use of a Coarse to Fine (CtF) procedure for fast image scan; (2) the use of object parts to

simulate local deformations; (3) the evaluation of detections with missing resolutions; and (4) the introduction of an additional feature that balances out scores with missing resolutions and gives possibly high scores also to small detections, which are very important in the context of driving assistance. The work [4] proposed a decomposition-based human localization model dealing with this issue in three steps, i.e., a stable upper-body is firstly detected, then a set of bigger bounding boxes are extended, from which the most appropriate instance is distinguished by a discriminative Whole Person Model (WPM). The work [5] presented a method for characterizing tiny images of pedestrians in a surveillance scenario, specifically, for performing head orientation and body orientation estimation, employing arrays of covariance as descriptors, named Weighted ARray of COvariances (WARCO). The design [7] addressed the problem of ascertaining the existence of objects in an image. In the first step, the input image is partitioned into non-overlapping local patches, then the patches are categorized into two classes, namely natural and man-made objects to estimate object candidates. Then, a Bayesian methodology is employed to produce more reliable results by eliminating false positives. To boost the object patch detection performance, they exploit the difference between coarse and fine segmentation results. The design [8] proposed a representation for scenes containing relocatable objects that can cause partial occlusions of people in a camera's field of view. The authors formulated an occluder-centric representation, called a graphical model layer, where a person's motion in the ground plane is defined as a first-order Markov process on activity zones, while image evidence is aggregated in 2D observation regions that are depth-ordered with respect to the occlusion mask of the relocatable object. The work [9] improved on the successful Evolution COnstructed (ECO) features algorithm by employing speciation during evolution to create more diverse and effective ECO features. Speciation allows candidate solutions during evolution to compete within niches rather than against a large population.

The aforementioned literature provide solid foundation for researchers to develop their works. However, there is still a gap to be filled before one can achieve accurate moving objects detection for intelligent automobiles with real-time performance on portable platforms.

8.2.2 Lane Departure Warning System

In the basic system flow of LDWS, there are two main steps of detecting the lanes. One is lane-mark generation, and the other is lane model fitting.

At lane-mark generation stage, most papers, such as [32, 33], and [34, 35], used canny edge detector, which can keep a good performance even in low contrast weather conditions, to extract the lane-mark. However, the computing burden of canny edge detector is more than the one of brightness thresholding. With this reason, the paper [36] used brighter region to extract lane-mark based on Charge-Coupled Device (CCD) camera parameter regulation skill.

Straight line is the most famous and common lane model, which can be detected with Hough transform [33, 34, 36] or Weight Least Square Regression (WLDR) [37], because the lane-mark appears straight near the vehicles. In order to include curves, curve lane models are adopted in some papers. Lindner et al. [32] used the fitting value of Hamacher function to decide whether the line can be added or not. Yoo et al. [35] used the quadratic curve model to represent the lane-mark as shown in Eq. (8.1), where x and y are coordinate values, c_0 is the curvature of the lane, m is the slope, and b is the offset of the lane.

$$y = \frac{1}{2}c_0x^2 + mx + b \tag{8.1}$$

8.2.3 Forward Collision Warning System

In basic system flow of FCWS, Sun et al. [38] claimed two steps for detecting vehicles, i.e., Hypothesis Generation (HG) and Hypothesis Verification (HV). Generally speaking, the computing time in HV is more than that in HG, so eliminating most of free-driving space in HG by apparent vehicle features is needed. Next, some famous methods of HG and HV are introduced in the following paragraphs.

For shadow feature in HG, Kumar [39] used edge and gray value to capture shadow features. To begin with, free-driving space is extracted by edge segmentation and then they compute the mean and standard deviation to calculate threshold of shadow. The intensity-value difference in the vertical direction is examined to see whether or not there exists a transition from brighter intensity to darker intensity. The candidates are revealed in HG by shadow features.

For symmetry feature in HG, Teoh et al. [40] used symmetry characteristic to generate vehicle candidates. Canny operator is adopted to find reliable edge to be the basis of symmetry calculation. They select a pair of proper width and height to calculate symmetry depending on different scanning lines.

For tail-light feature in HG, Fossati et al. [41] used color, shape, area, and position information as a criterion of pairing algorithm. With low-exposure camera, they can take advantage of accurate color information in Hue–Saturation–Value (HSV) space to decide which type the light-object belongs to. After pairing tail-light objects, they estimated the forward vehicles position by using a pinhole model.

For Support Vector Machine (SVM) in HV, Teoh et al. [40] used two-pattern classifier based on a linear SVM to differentiate between vehicles and non-vehicles. In fact, they extract Edge Orientation Histogram (EOH) to be their features of the classifier by quantizing the gradient of each pixel into eight bins. Khairdoost et al. [42] applied lots of methods to SVM in order to raise classifier accuracy. First, Pyramid Histogram of Oriented Gradient (PHOG) is adopted to produce more

features for the classifier. Second, they used Principle Component Analysis (PCA) to eliminate redundant PHOG features. Third, genetic algorithm is utilized to find the weighting of PHOG-PCA features.

8.2.4 Speed Limit Detection System

A basic speed limit signs detection flow can be generally divided into three parts, which include speed limit signs detection, to firstly locate the potential candidates for speed limit signs, speed limit signs verification, to verify if the candidates from the previous stage are indeed speed limit signs, and speed limit signs recognition, to classify critical information from the speed limit signs [54, 55].

8.2.4.1 Speed Limit Signs Detection

The goal of this phase is to select the potential signs by locating where they appear. There are two major kinds of speed limit signs, which are circular and rectangular as shown in Fig. 8.1 over the world.

Radial symmetric transform [44, 45] is one of popular shape detection algorithms in the sign location process, which aims to detect the center of n-side regular polygons in gray-scale images through the radial symmetric feature.

8.2.4.2 Speed Limit Signs Verification

In this phase, the candidates from the previous stage are verified whether they are speed limit signs or not by checking their contents.

AdaBoost learning with Haar-like features, a machine learning architecture proposed by Viola–Jones, is adopted [46]. Cascaded geometric detectors are defined such as area, solidity which is the ratio between the number of ROI

Fig. 8.1 Common types of speed limit signs over the world

background pixels and the total number of ROI pixels, vertices relative positions which detect rotated and non-rotated objects, and dimensional ratios which discard non-symmetric objects and maintain rectangular shape objects [47].

8.2.4.3 Speed Limit Signs Recognition

Here, the speed limit signs are classified to recognize the actual speed limit digits inside the signs. A binary classifier SVM is adopted by the tree structure with rotation invariant features, which is generated by Fourier transformed input image to classify speed limit signs [48]. Digits features are also considered one of efficient manners, e.g., blob, which is defined as a closed region, and breach, which is defined as the open region [49, 50].

8.2.5 Inclement Weather Processing Technology (DLCE)

Currently, there is no one-size-fits-all solution to inclement weathers. Each inclement weather is considered separately. Many de-fog technologies are proposed in recent years, which can be categorized into multi-frame-need/single-frame-need or image enhancement/physical model recovery. Some multi-frame-need methods [56, 57] generated a good result but those are not able to be adopted in dynamic scenes. Thus, single-frame-need methods were developed, with an image enhancement-based idea. Solving fog influence with single frame is accompanied with huge challenges. No matter using cost function and Markov Random Fields [58] or Independent Component Analysis (ICA) [59] has some disadvantages, i.e., unnatural image, and limited applied image. A Dark Channel Prior (DCP) [60, 61] was proposed for de-fogging foggy images in 2009. This assumption obtained wonderful de-fogging results but required high computational complexity and relied on precision of the dark channel computation, that is, the result may fail once the dark channel is wrong. Night is also one of most encountered inclement weathers. Histogram Equalization (HE) [62] is a common way used in image processing for inclement weathers, which is usually applied to image enhancement for single-camera high dynamic range (HDR) processing. However, HE could not reveal comprehensive details and did not cover the consideration for local conditions. Therefore, Adaptive Histogram Equalization (AHE) [63, 64] and Contrast Limited Adaptive Histogram Equalization (CLAHE) [65] were proposed to improve this weakness of HE. A better result is generated by adopting these methods. Since low contrast is a common phenomenon among different inclement weathers, these contrast enhancement algorithms are quite often to be adopted in this field.

8.3 Advanced ADAS System

Vision processing technology has a wide range of applications on vehicles (Fig. 8.2). Collaborated with front cameras, PMD, LDWS, FCWS, Stop-and-Go, Traffic Light Detection (TLD), and SLDS techniques can help drivers notice pedestrians, motorcyclists, vehicles, traffic lights, and speed limit signs in front of the way. With proper vehicle control intervention, FCWS and Stop-and-Go techniques can guarantee drivers further driving safety. Side cameras capture videos for BSD Systems (BSDS). Meanwhile, wide-view video stitching can combine several video sources captured from different cameras around the car and provide a panoramic view with viewing angle up to 360° for drivers. Furthermore, HDR technology helps drivers see clear in scenes with high variation in lighting, e.g., when going in/out tunnels and facing strong lights in opposite direction at night. Inside the cars, driver dangerous behavior detection system can help remind drivers to drive the cars properly. Besides, hand tracking technology help drivers control the in-car equipment by a more convenient and safer way, i.e., hand gesture.

In the following, we introduce more details on several significant intelligent vision processing technologies for ADAS including PMD, LDWS, FCWS, SLDS, and DLCE.

8.3.1 Pedestrian and Motorcyclist Detection System

Machine learning algorithms are widely used in pattern recognition such as face detection and license plate recognition. They own high flexibility and good accuracy in detecting the target objects for specific applications. The classifiers used in

Fig. 8.2 The application scenario of our intelligent vision processing technology developed in NCTU and FCU, Taiwan

machine learning algorithms are specifically trained based on the collected samples for the target. However, they often generate false alarms as patterns similar to target objects appear in the background, which is also the major design challenge when applying machine learning algorithms on object detection. On the other hand, larger color contrast between the object and the background is helpful for detection. Hence, the positive samples used to train the classifier should be representative. To enhance the robustness of multiple moving object detection, we propose several rules in capturing training samples, and then employ AdaBoost algorithm to detect pedestrians, and people riding motorcycles commonly seen on the road.

Due to the great diversity of clothing and posture, pedestrian detection has always been a challenging task. Besides, information from the background may also influence the detection result. Therefore, we propose to choose pedestrian samples by following the rules below: each sample includes one pedestrian only, let the boundary of the sample close to the pedestrian, and include both samples which include global features and local features. In addition, the training sample selection method for motorcycle detection is somewhat different to that for pedestrian detection because motorcycles move much faster than pedestrian. According to the distance, we present sample selection ways for detecting farther motorcycles and nearer ones, respectively. Considering that farther motorcycles appear to be smaller than the nearer ones, we adopt samples with global features to enhance both the detection accuracy and detection distance. On the contrary, nearer motorcycles appear to be larger in size than the farther ones, we adopt samples with local features to decrease the false detection rate and avoid detection miss due to occlusion. Moreover, we adopted vehicle samples of local features for training the classifier to detect both nearer and farther vehicles because vehicles are rather larger than motorcycles in terms of size.

The performance of the classifier is not always enhanced as the number of samples is increased. The training phase for obtaining a discriminative objects classifier may be time consuming. Therefore, we propose a multi-pass self-correction procedure for effectively achieving the goal. First, we have a classifier trained by using conventional one-pass procedure. We then use this classifier to detect the internal samples from the adopted database and keep the correctly detected samples only for retraining the classifier. After that, we use the classifier to detect more samples beyond the adopted database and keep the correctly detected samples only when retraining the classifier. This method not only can avoid wrecking the feature of the basic classifier, but also can emphasize versatile features of different objects for improving detection performance. To improve the detection performance indicates both increasing the detection accuracy and lowering down the false detection rate. From the performance evaluation, we find that the proposed self-correction procedure not only can save 25 % effective sample selection time but also can lower down 50 % false detection rate due to the effective samples collected. The time saving is evaluated according to the working time saving of an operator who is familiar with the sample selection process including sample capturing and classifier training. More details of the above methods can be found in [30].

Furthermore, we adopt heterogeneous computing with OpenCL for realizing this object detection design on a multicore CPU and GPU platform. By utilizing the

techniques of scale parallelizing, stage parallelizing, and dynamic stage scheduling on AdaBoost algorithm, windows load unbalance problem and scale load unbalance problem are solved. Consequently, the proposed object detection design achieves 32.5 frame per second (fps) at 720 × 480 (D1) resolution on an AMD A10-7850K processor.

8.3.2 Lane Departure Warning System and Forward Collision Warning System

Traffic accidents may cause great damage on people's lives and wealth. In Taiwan, there are 273,449 traffic accidents (i.e., the sum of A1 traffic accidents, which causes people die within 24 h, and A2 ones, which causes people injury or die beyond 24 h) in 2013. About 20.4 % of A1 traffic accidents happened due to fatigue driving or drivers' inattention. Therefore, it is inevitable to develop some driver assistance functions to help reminding the drivers to be aware of the dangerous driving conditions. Among the ADAS functions, LDWS and FCWS are two major technologies.

In real driving environments, there are lots of problems, which may cause ADAS functions to fail, such as inclement weathers and complicated scenes. To conquer the problem of inclement weather, dynamic threshold, which combines local threshold with global threshold and change the threshold automatically, is proposed to avoid illumination variation. To reduce the effect of different scenes, multiple frame approval, which accumulates the frequency of the desired objects based on the position of the frame in order to eliminate static objects and the false alarm caused by windshield wiper, is adopted when the vehicle speed is over 40 km/h. With the aforementioned solutions, we are able to overcome the design challenges of the two popular ADAS functions, i.e., LDWS and FCWS, to be applied in real world driving environments, which is one of the major contributions of this chapter.

Hence, we propose the design, verification, and vision radar system integration of two popular ADAS functions including LDWS and FCWS. A dynamic threshold method (including local threshold and global threshold) is adopted to improve the accuracy of lane detection and vehicle detection in various weather conditions. Multiple frame approval is adopted to conquer the effect of some tough scenes, such as static objects, signs appearing temporarily, or interference of the windshield wiper. The proposed system is implemented on automotive-grade Freescale i.MX6 (including 4-core ARM Cortex A9, 1 GB DDR3 RAM, and Linux environment) with a USB webcam to capture the video. Under the D1 resolution, the performance of the proposed LDWS achieves 33 fps, while the performance of the proposed FCWS achieves 32 fps, and the performance of integrated application achieves 22 fps.

8.3.3 Speed Limit Detection System

In recent years, car cam recorders have become more and more popular. Thanks to it, the footages of car accidents, dangerous driving, or other critical road-side events can be entirely recorded, which can help justify who is the perpetrator and be the evidence for illegal driving. Furthermore, the information of traffic signs is able to be extracted from footages. To realize multiple-country SLDS on embedded systems, we simplify the proposed algorithms computing resource with template databases and digital features.

Currently, color-based algorithms are adopted in most of the SLDSs along with training samples for machine learning algorithms. This approach needs to not merely adjust training samples based on different cameras, but also take long-period training time causing it less practical chance. A low complexity shape-based speed limit sign locating algorithm, adaptive threshold and multiple-country-suitable speed-limit-digit recognition algorithm are proposed. Our multiple-country SLDS maintains good detection rate under inclement weathers. The proposed algorithm reaches 150 fps on the Intel i7-2600 3.4 GHz CPU desktop with D1 resolution on desktop and 30 fps with D1 resolution on Freescale i.MX 6 platform.

8.3.4 Inclement Weather Processing Technology (DLCE)

Many key functions in ADAS were proposed and expected to execute with normal weather conditions in the recent years. Capturing a low contrast image and shooting a color-faded image might cause failure of these systems. The technologies proffered to deal with inclement weather [56–68] are framed narrowly, which are not one-size-fits-all. The contrast of vision decrease exists at nights, foggy days, cloudy days, and rainy days with our observations. Thus, exploiting the idea of AHE, we propose a so-called Dynamic Local Contrast Enhancement (DLCE) technique, which can strengthen the image quality in the most inclement weather conditions, improve unnatural over-enhancement image quality, and reduce noise existed in the image. DLCE technique is designed and implemented on an embedded platform to verify its correctness and robustness. Without specific hardware and software optimizations, the proposed DLCE system is realized on TREK-668 platform with ATOM 1.6 GHz in real-time for both requirements of 120 fps 352×288 (CIF) resolution and 50 fps D1 resolution.

8.4 Implementation Issues

Both detection performance and real-time implementation on embedded system are our important targets. Through using OpenCL programming language, a heterogeneous system architecture can speed up the processing performance of PMD. Those

who tend to detect targets with unique appearances, i.e., LDWS, FCWS and SLDS, can well recognize the targets by analyzing specific features, e.g., shadows, tail lights, and shape. Furthermore, these detection systems keep good precision rate in various weather by adopting the extended idea of DLCE. The implementation of proposed PMD, LDWS, FCWS, SLDS, and DLCE is illustrated in this section as follows.

8.4.1 Pedestrian and Motorcyclist Detection System

AdaBoost algorithm has been widely used in face recognition. Due to its simplicity and regularity, it has also been adopted to detect other targets such as pedestrians, motorcyclists, and vehicles. Haar-like features are usually utilized alone with AdaBoost algorithm. Besides, developers find that a skill called integral imaging can reduce the redundant computation of Haar-like features and therefore can accelerate the detection speed. Moreover, AdaBoost classifier contains multiple weak classifiers in a cascade manner. Only when a candidate passes all weak classifiers, it is recognized as a targeted object.

Before a superior AdaBoost classifier can be obtained, one should collect enough positive samples and negative samples and train the classifier. The selection of training samples affects the detection performance of the resulting classifier obviously. By adopting the sample selection rules and multi-pass self-correction procedure presented in Sect. 8.3.1, we can train a superior AdaBoost classifier efficiently. After that, we can utilize OpenCV to implement the AdaBoost classifier for PMD.

Although OpenCV contains many useful subroutines, the version performs only 4.76 fps on the Panda board for 320×240 (QVGA) video format. On the other hand, we have also implemented the proposed design as a C model to improve the real-time performance. The C version can achieve 30.3 fps for QVGA video format on the Panda board, i.e., 6.37 times of performance speedup is obtained. Our method achieves a detection rate of 91.8 % with only 3.3 % false alarm rate for multiple objects detection. The detection performance of our method is obviously better than that of existing methods.

In addition, we also adopt heterogeneous computing with OpenCL for realizing this PMD design on a multicore CPU and GPU platform. Figure 8.3 illustrates the flow chart of the proposed heterogeneous computing system for PMD. Scaling up of detecting windows for different sizes of pedestrians is required. The big scaling factors have less number of windows than small scaling factors have. Besides, the amount of windows for detecting near objects is much less than that for detecting far objects. Meanwhile, using GPU to process small amount of windows is inefficiency due to the induced memory latency. Hence, we propose to execute the far-distance PM detection in parallel on GPU and finish the near-distance PM detection in parallel on CPU to maximize the resource utilization of both CPU and GPU cores. Besides, the non-PM windows are rejected at the earlier stages in

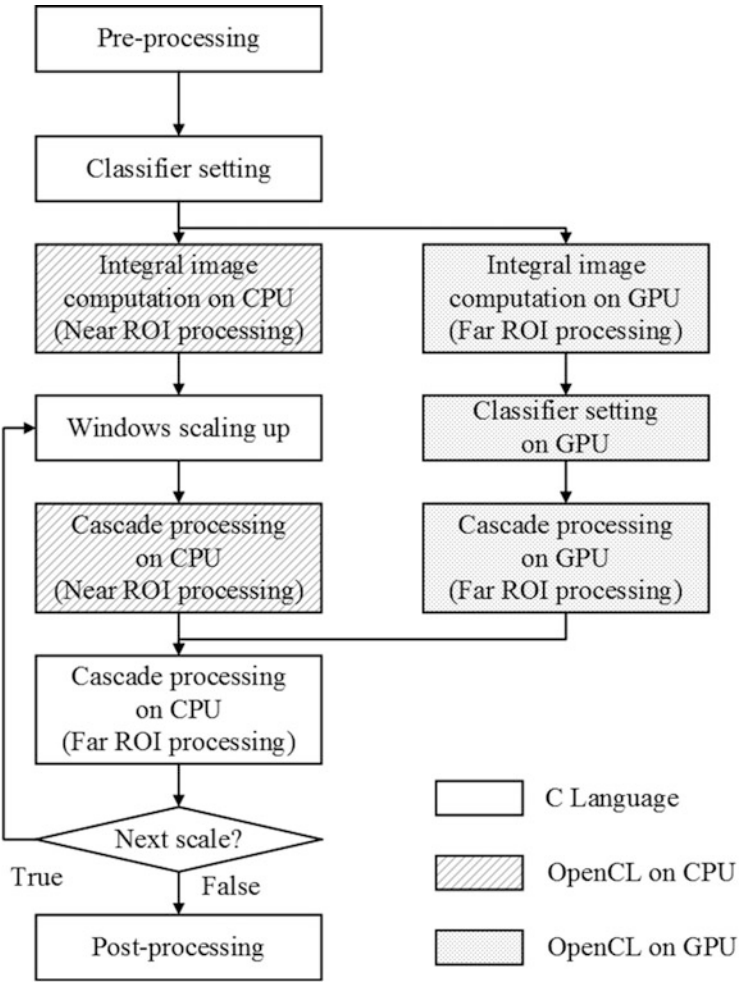


Fig. 8.3 The flowchart of proposed heterogeneous computing system for Pedestrian and Motorcyclist Detection System (PMD)

the cascade process of AdaBoost algorithm, and only few windows go through all the stages. Over 95 % of candidate windows are dropped after the first five stages and only 0.03 % of candidate windows finish the whole cascade process. Because the parallelism of the rest candidate windows is not suitable for GPU parallel processing, GPU passes the rest of candidate windows to CPU to complete the whole cascade process in the proposed system. Meanwhile, GPU can process the next scale of pedestrian detection. Moreover, a dynamic stage scheduling is proposed to keep both the CPU and GPU busy in different situations. Once the processing time on GPU is less than that on CPU, more stages then are allocated for GPU in the next scale computation and vice versa. By utilizing these techniques

of scale parallelizing, stage parallelizing, and dynamic stage scheduling on AdaBoost algorithm, windows load unbalance problem and scale load unbalance problem are solved. Consequently, the proposed PMD design can achieve 32.5 fps at D1 resolution on an AMD A10-7850K processor.

8.4.2 *Lane Departure Warning System*

As shown in Fig. 8.4, Conditional Dynamic Threshold (CDT) is adopted to extract brighter region as lane-marks. Line-thinning speeds up the Hough transform procedure, then line collection reduces lots of lines not belonging to lane-marks. Our algorithm can solve the problems of detecting outer lanes and avoiding the effect of windshield wiper by the occurrence frequency in finite state machine.

CDT is applied at the first stage, that is, more conditions are added into the equation to conquer the effect of the different weathers, such as day, nightfall, or night. The weather condition is based on the sky region as shown in Fig. 8.5.

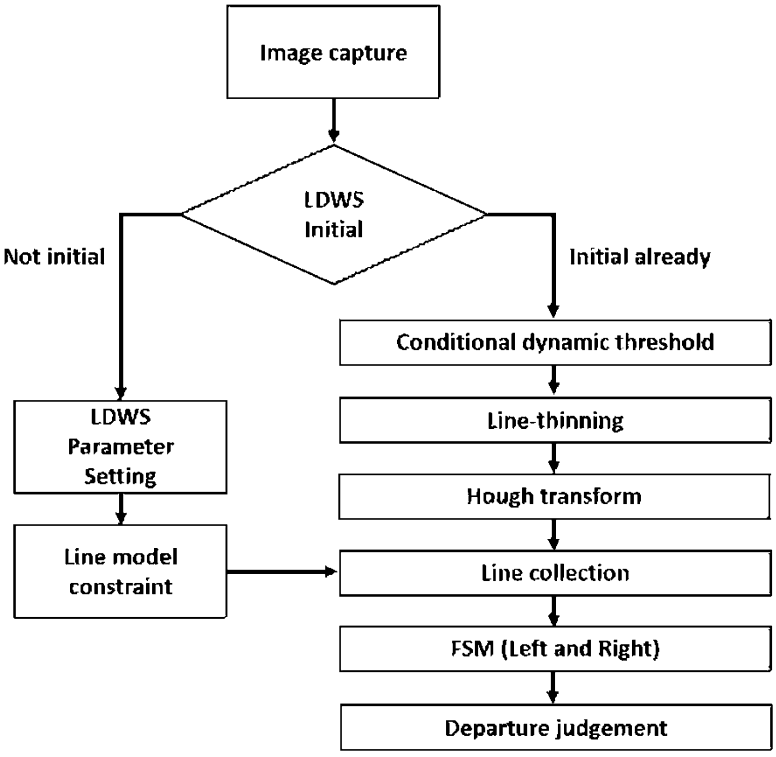


Fig. 8.4 The flowchart of proposed Lane Departure Warning System (LDWS)

Fig. 8.5 Sky region of judging weather condition

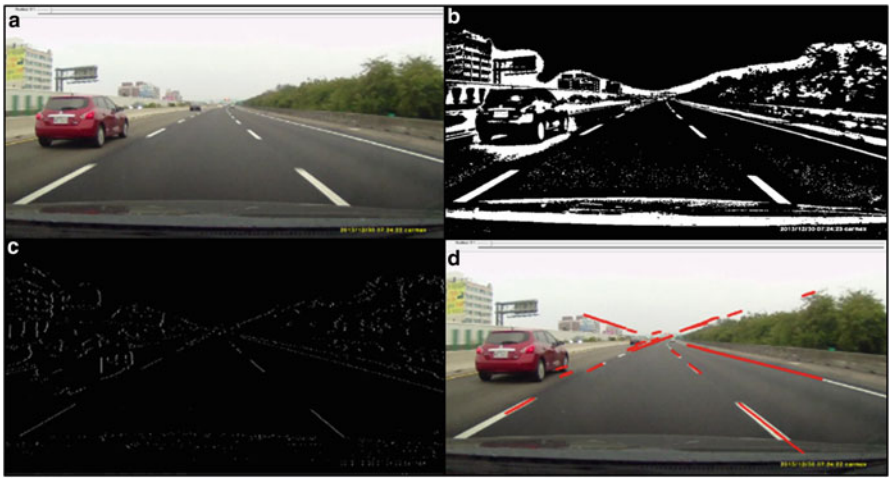


Fig. 8.6 Line detection result images: (**a**) original image, (**b**) after CDT, (**c**) after line-thinning, and (**d**) after Hough transform

After thresholding of CDT, line-thinning is executed to speed up Hough transform, as shown in Fig. 8.6.

The line model is composed of the middle point of the vehicle in x coordinate, and the vanishing point in y coordinate. With our observation, we find that the distance from the point to the line, belonging to the lane, will be always less than the threshold, 50 pixels in D1 resolution, and the constraint is shown in Eq. (8.2).

$$\begin{aligned} &p(x_0, y_0) \ \ and \ \ L : ax + by + c = 0 \\ &D(p, L) = \frac{|a \cdot x_0 + b \cdot y_0 + c|}{\sqrt{a^2 + b^2}} < 50 \end{aligned} \tag{8.2}$$

This system is integrated with the FCWS in order to enhance driving safety, whose results are shown at the end of Sect. 8.4.3.

8.4.3 *Forward Collision Warning System*

As shown in Fig. 8.7, weather judgment is applied and benefits dynamic threshold decision to extract brighter region as tail-light features and darker region as shadow features. We use vertical and horizontal edge to verify features in spatial domain. The occurrence frequency reduces the false detections caused by different scenes or windshield wiper in temporal domain since the changes in each frame of different scenes and windshield wiper are stronger than the one of lane is.

Vehicle model based on certain position takes advantage of judging whether the vehicle size is correct or not. The proposed vehicle model expresses the situation, that is, the farther the vehicle is, the smaller it will be, and vice versa, as shown in Eq. (8.5), which is deduced from pinhole camera extended equation, Eq. (8.3), and an equation from [43], i.e., Eq. (8.4), by eliminating variable D. For parameter meaning, F_c is camera focal, and y_h is the vanishing point. We assume that the

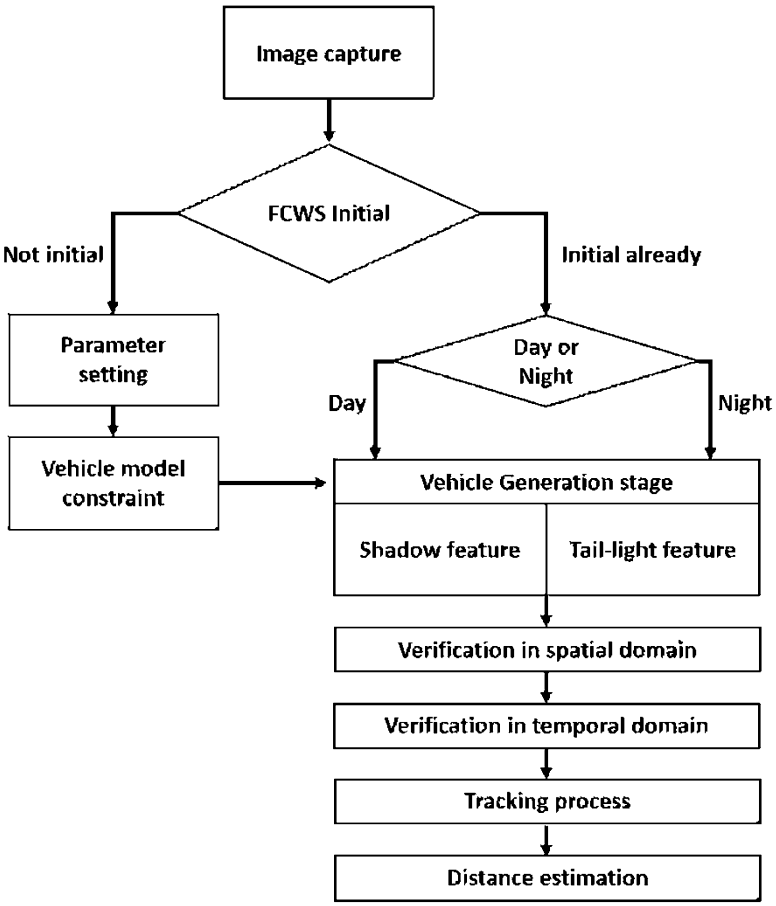


Fig. 8.7 The flowchart of proposed Forward Collision Warning System (FCWS)

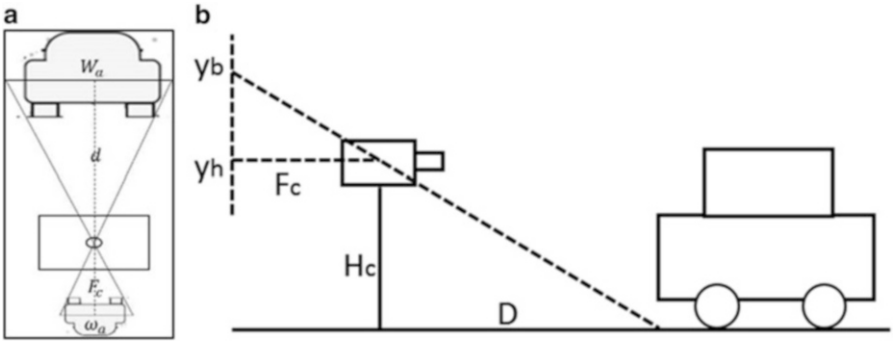


Fig. 8.8 Distance estimation: (**a**) pinhole model and (**b**) equation from [43]

camera height (H_c) is 1.5 m, and the average width of vehicles (W_a) is 2 m. The bottom width of the vehicle must be in the range of generated vehicle model on the corresponding image y-axis as demonstrated in Fig. 8.8.

$$\mathrm{D} = \mathrm{F_c} \cdot \frac{W_a}{\omega_a} \tag{8.3}$$

$$\mathrm{D} = \frac{\mathrm{F_c} \cdot H_c}{y_b - y_h} \tag{8.4}$$

$$\omega_a = (y_b - y_h)^* \frac{W_a}{H_c} \rightarrow \omega_a = (y_b - y_h)^* 1.3 \tag{8.5}$$

Light objects can be obtained by Connecting Component Labels (CCL) in the image after dynamic threshold. Rule-pairing is to find tail-light features with four rules, i.e., horizontal position, area, motion, and vehicle model. If two of the light objects satisfy the four rules, the tail-light feature is formed.

Four 5-min videos are picked up to show our integrated achievements (i.e., LDWS and FCWS) in this chapter. Forward vehicles within the range between 5 and 40 m at day, and between 5 and 30 m at night, and lateral vehicles within the range between 10 and 30 m at both day and night should be detected in our specification. We sample each frame per second in order to reduce the counting process of detection rate and false alarm, defined in Eq. (8.7). As shown in Table 8.1, we can achieve 90.15 % detection rate and 5.96 % false alarm rate averagely, and each result of the video sequence would be displayed in Fig. 8.9.

$$\begin{cases} vehicles(real\ \ answer) \\ non-vehicles(real\ \ answer) \end{cases} \begin{cases} vehicles(our\ \ judgement) & a \\ non-vehicles(our\ \ judgement) & b \\ vehicles(our\ \ judgement) & c \\ non-vehicles(our\ \ judgement) & d \end{cases} \tag{8.6}$$

$$Detection\ \ rate = \frac{a}{a+b}, \quad False\ \ alarm\ \ rate = \frac{c}{a+c}$$

Table 8.1 The integrated system experiment results

No	Weather	Scene	Detection rate	False alarm
1	Day	Highway	99.62 % (0265/0266)	001.48 % (004/0269)
2	Day	Highway	96.88 % (0249/0257)	002.35 % (006/0255)
3	Day	City	98.49 % (0196/0199)	009.25 % (020/0216)
4	Night	Highway	76.07 % (0248/0326)	006.06 % (016/0264)
5	Night	City	80.85 % (0114/0141)	016.17 % (022/0136)
Day			98.33 % (0710/0722)	004.05 % (030/0400)
Night			77.51 % (0362/0467)	005.26 % (038/0722)
Average			90.15 % (1072/1189)	005.96 % (068/1140)

Fig. 8.9 The integrated system experiment results

On Freescale i.MX6 with Logitech C920 webcam input, LDWS achieves 33 fps at D1 resolution, FCWS achieves 32 fps at D1 resolution, and the integrated system reaches 22 fps at D1 resolution.

8.4.4 Speed Limit Detection System

8.4.4.1 Shape Detection

Figure 8.10 shows the flowchart of proposed SLDS, which starts from shape detection after getting image. The voting process is based on the gradient of each pixel. The direction of gradient can form a vote. The vote generated from each pixel follows the symmetric axes, which cause the highest in the center of the shapes.

Sobel operator is used to generate horizontal and vertical gradients, where each selected pixel is represented with its absolute magnitude, and the gradient vector is denoted as $\mathbf{g}(\mathrm{p})$. The direction of $\mathbf{g}(\mathrm{p})$ can be formulated with the horizontal gradient G_x and the vertical gradient G_y into an angle as shown in Eq. (8.7).

$$\mathbf{g}(\mathrm{p}) = \tan^{-1}\frac{G_y}{G_x} \tag{8.7}$$

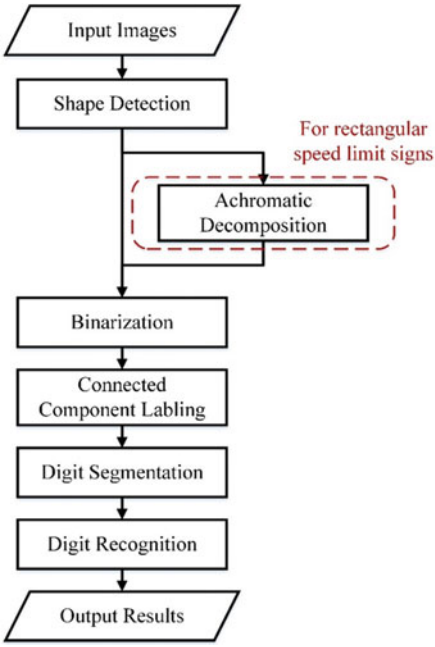


Fig. 8.10 Proposed speed limit detection algorithm

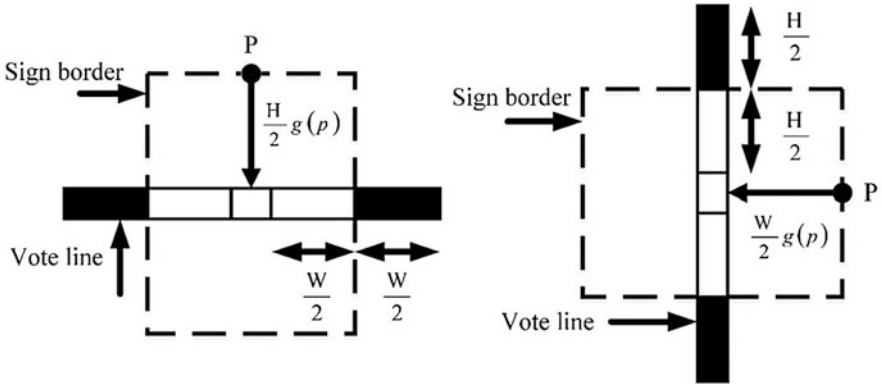


Fig. 8.11 The voting process for both horizontal and vertical vote

The voting image is firstly initialized to zero, and accumulates vote-number by horizontal and vertical voting line, which is generated by each pixel with positive and negative votes as illustrated in Fig. 8.11.

The centers of sign candidate will receive higher vote-number as shown in Fig. 8.12, and then several sign candidates are picked up with designed constraints for rest judging processes.

Fig. 8.12 The result after the voting process

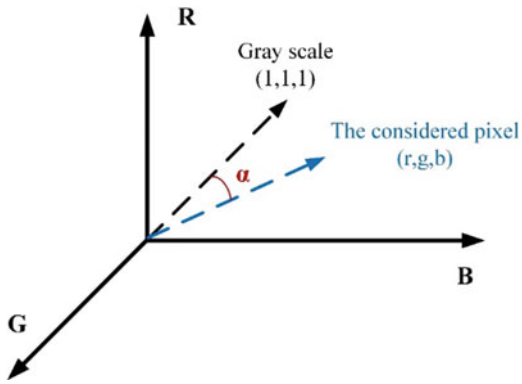


Fig. 8.13 The schematic of the RGB model and the angle α

8.4.4.2 Achromatic Decomposition

We use the inner product between (1,1,1), where gray scale is along, and each considered pixel to check the angle α between these two vectors to apply the decomposition in RGB domain as illustrated in Fig. 8.13, where each considered pixel in vector form of (r,g,b). The cosine function of α, which is equal to the inner product is shown in Eq. (8.8).

$$\cos\alpha = \frac{(1,1,1)\cdot(r,g,b)}{|(1,1,1)|\times|(r,g,b)|} = \frac{r+g+b}{\sqrt{3}\times\sqrt{r^2+g^2+b^2}} \tag{8.8}$$

8.4.4.3 Binarization and Digit Segmentation

Otsu threshold is adopted in days and adaptive threshold is adopted in nights. Using the fact that the speed limit of rectangular speed limit signs is two-digit, the pairing rules sizes and positions are proposed as below:

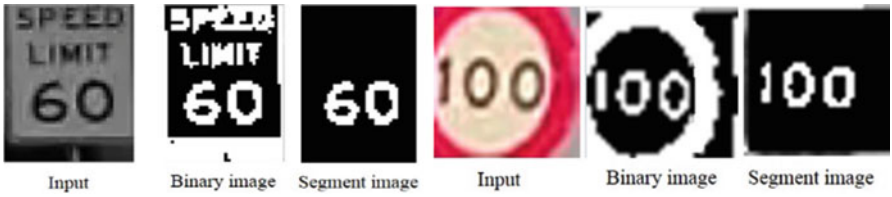


Fig. 8.14 The example of digit segmentation results

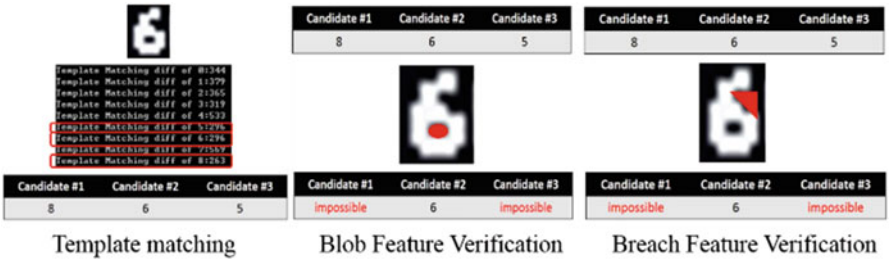


Fig. 8.15 The example of digit recognition results

1. The areas of the digit candidates should be similar;
2. The positions of the digit candidates should be close enough;
3. The density of the pixels inside the digit candidates should be similar.

The pairing steps of circular speed limit signs are similar but looser to one of the rectangular speed limit signs because two-digit and three-digit speed limits are included in circular speed limit signs (Fig. 8.14).

8.4.4.4 Digit Recognition

The extracted digits are firstly classified with the built-in different-font templates and selected out three possible numbers. After it, the blob, the closed region inside the digit, and breach features are applied to verify the final number.

After calculating the Sum of Absolute Difference (SAD) between the target digit candidate and the built-in templates, we select three possible digit numbers with less matching difference. With possible digit candidates, we adopt union row, which gathers several rows as a union row, to detect the blob feature, and then verify the digit. The pixel value of a union row is the union value of all rows in the union row. Then, for each union row, we count the number of lines in white pixels. A blob is formed only if the number of lines is the sequence of "1, 2, . . ., 2, 1". Similarly, we count the number of pixels where the white pixel first appears from both the right and the left in each column to half of width of digit candidates to detect breach, where there are a series number of pixels that are larger than half of digit height (Fig. 8.15).

Table 8.2 The accuracy for two different major types of signs

	Rectangular	Circular
Total video frames count	3482	2832
Total speed limit sign count	24	25
Detected signs	23	24
Detection accuracy (%)	95.83	96.00
Total detected signs frames count	81	80
Total detected signs and correctly classified frames	78	77
Recognition accuracy (%)	96.30	96.25
Overall accuracy (*detection accuracy* * *recognition accuracy*) (%)	92.28	92.40

Table 8.3 The details of rectangular speed limit detection

Video sequence number	1	2	3	4	5	6	7
Weather	Day	Day	Day	Night	Night	Rain	Rain
Number of signs	4	4	3	4	2	2	3
Detected signs	4	4	3	3	2	2	3
Missed signs	0	0	0	1	0	0	0
Number of frames with sign detection	15	17	10	9	7	9	12
Number of correct speed limit recognition	15	16	10	8	6	9	12
Number of wrong speed limit recognition	0	1	0	1	1	0	0

Table 8.4 The details of circular speed limit detection

Video sequence number	1	2	3	4	5	6	7
Weather	Day	Day	Day	Day	Night	Night	Rain
Number of signs	4	6	5	5	2	2	1
Detected signs	4	6	5	5	1	2	1
Missed signs	0	0	0	0	1	0	0
Number of frames with sign detection	12	20	15	16	4	9	4
Number of correct speed limit recognition	12	19	15	15	4	9	3
Number of wrong speed limit recognition	0	1	0	1	0	0	1

Freescale i.MX6 is chosen as the target embedded platform and the proposed algorithm is first designed in C++ with Visual Studio platform on an Intel i7-2600 3.40 GHz CPU desktop running Windows 7 with 8 GB memory. It reaches 150 fps at D1 resolution in average for both rectangular and circular speed limit signs on desktop and 30 fps at D1 resolution on Freescale i.MX6. The accuracy for two different major types of signs is listed in Table 8.2. Tables 8.3 and 8.4 also show the accuracy under different weather conditions for different types of signs.

As listed in Table 8.5, the proposed system provides high accuracy and efficient performance in applications of speed limit detection. It can reach real-time implementation on embedded systems with low computing resource, support different types of multiple-country speed limit signs, and support different digit fonts thanks to adopting the blob and breach features. Figure 8.16 shows SLDS results in various weather conditions and countries.

Table 8.5 The comparisons among other works and our system

	[51]	[52]	[46]	[53]	Our system	
CPU	2.13 GHz dual-core laptop	1.67 GHz Intel Atom 230 and a NVIDA GeForce 9400M GSGPU	2.16 GHz dual-core laptop	2.13 GHz dual-core laptop	Intel® Core™ i7-2600 CPU 3.4 GHz	Freescale i. MX6 with 4-core 1 GHz Cortex-A9 CPU
Accuracy (%)	90	88	96.25	90.9	92.3	
Video resolution	640 × 480	640 × 480	700 × 400	Image only	720 × 480	
Frame rate on PC (fps)	20	33	25	7.7 (130 ms)	150	30
Real-time on embedded system	X	O	X	X	O	
Supporting all types of speed limit sign	O	X	X	X	O	
Supporting different fonts of speed limit signs	X	X	X	X	O	

Fig. 8.16 The overall results for speed limit signs detection

8.4.5 *Inclement Weather Processing Technology (DLCE)*

In the proposed DLCE system, a contrast of a pre-defined block in the image is calculated and is used to limit the height of the histogram dynamically. Figure 8.17 shows the flowchart of the proposed system.

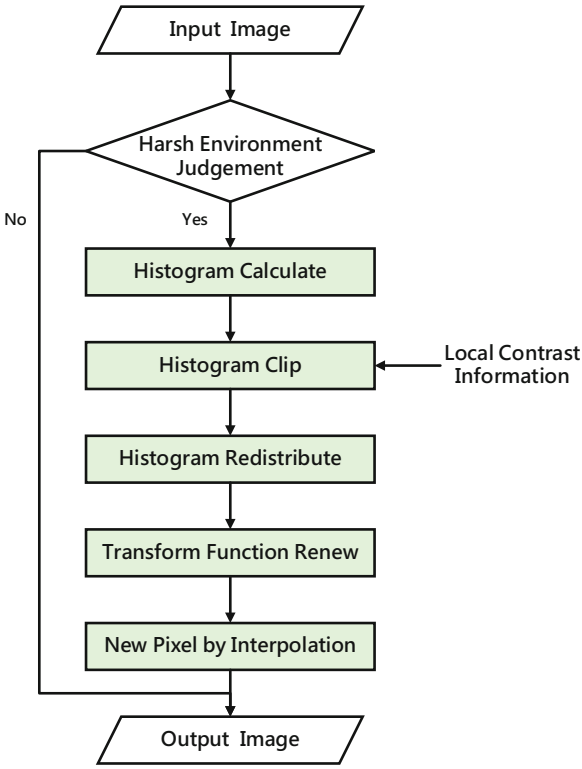


Fig. 8.17 Flowchart of the proposed Dynamic Local Contrast Enhancement (DLCE) method

Two main goals at harsh environment judgment phase are: (1) distinguishing harsh environment, and (2) obtaining the local contract information, as shown in Fig. 8.18. The luminance difference of pixel D(x,y) is calculated by the following equation:

$$\mathrm{D}(x, y) = |f(x, y) - A(x, y)| \tag{8.9}$$

where f(x,y) is luminance of pixel (x,y) and A(x,y) is the average of luminance in block $\mathrm{W} \times \mathrm{W}$. Eventually, the contrast with image size $\mathrm{M} \times \mathrm{N}$ is generated as Eq. (8.10).

$$\text{Contrast} = \frac{\sum_{x=0}^{M} \sum_{y=0}^{N} D(x, y)}{M \times N} \tag{8.10}$$

The luminance difference of pixel determines the value at which the histogram is clipped according to the following equations:

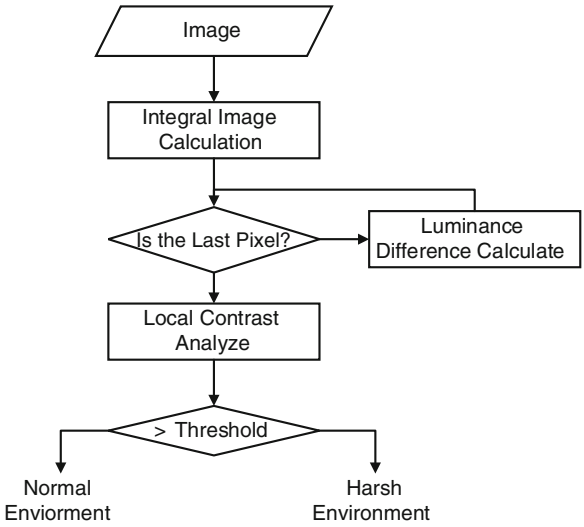


Fig. 8.18 Flowchart of the harsh environment detection

$$\text{clip} = (1-\beta)*H_{avg} + \alpha*H_{avg} \tag{8.11}$$

$$\beta = \left(\frac{\sum_{x=0}^{M}\sum_{y=0}^{N}D(x,y)}{M \times N}\right)/255 \tag{8.12}$$

β, H_{avg}, and α represent dynamic control parameter, average of histogram, and intensity parameter to bridle image adjustment, respectively. Mapping function Re (x) updates depend on β after clipped pixel is distributed, which benefits these techniques to fit the current scene. (Note: CDF is cumulative distribution function.)

$$\text{Re}(x) = \left(H'_{max} - H'_{min}\right) \times CDF(x) + H'_{min} \tag{8.13}$$

$$H'_{max} = H_{max} - \beta*(H_{max} - H_{min}) \tag{8.14}$$

$$H'_{min} = H_{min} + \beta*(H_{max} - H_{min}) \tag{8.15}$$

The proposed method has been implemented on TREK-668 embedded platform with 1.6 GHz Intel Atom N2600 CPU and tested under various inclement weather conditions. The performance of the prototype is able to achieve 50 fps at D1 video input.

Figures 8.19 and 8.20 present the quality of DLCE and Table 8.6 proves the statistics compared with Global Histogram Equalization (GHE), CLAHE [65], and DCP [60]. The average of higher contrast values, calculated by Eq. (8.10), as in Table 8.6 proves that DLCE method provides videos with clearer scene. Besides, local contrast is also improved when adopting the proposed method.

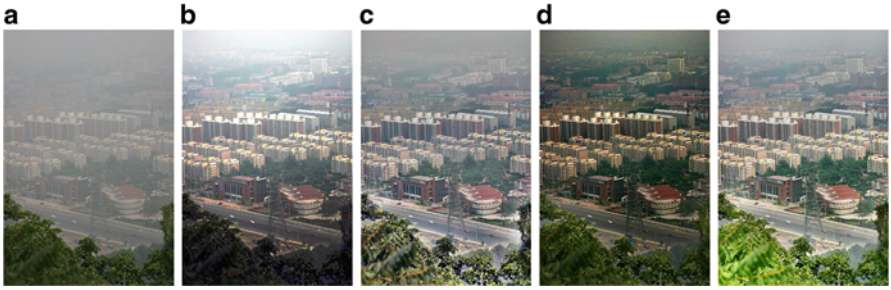


Fig. 8.19 Foggy day results with different methods: (**a**) original, (**b**) Global Histogram Equalization (GHE), (**c**) Contrast Limited Adaptive Histogram Equalization (CLAHE), (**d**) Dark Channel Prior (DCP), and (**e**) DLCE

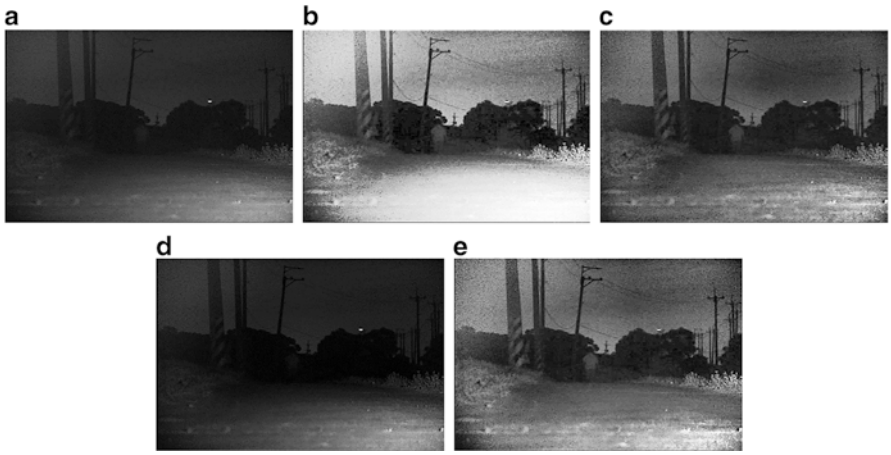


Fig. 8.20 Night results with different methods: (**a**) original, (**b**) GHE, (**c**) CLAHE, (**d**) DCP, and (**e**) DLCE

Table 8.6 Contrast values with various methods

	Original	GHE	CLAHE [65]	DCP [60]	DLCE
Fog day	9	18	25	19	24
Night	4	23	15	6	17

DLCE benefits ADAS in gaining better image quality at inclement weather conditions. More pedestrians are detected, better lane detection is generated, and farther scene can be seen while DLCE is applied, as demonstrated in Fig. 8.21 and Table 8.7.

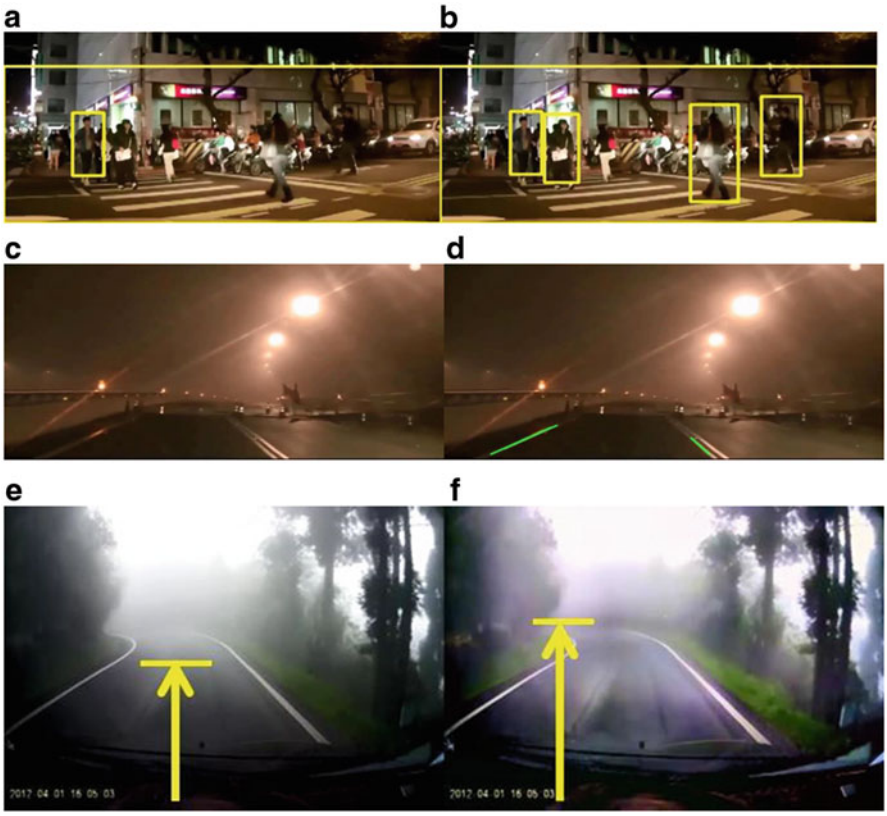


Fig. 8.21 ADAS results with adopting DLCE (**a**–**b**) PD, (**c**–**d**) LDWS, (**e**–**f**) car recorder, and (**b**, **d**, **f**) DLCE is adopted

Table 8.7 Increasing rate after adopting DLCE

	Number of detected lane LDWS	Number of detected lane LDWS + DLCE	Increasing rate of detected lane (%)
Rainy day	533	834	56.47
Foggy day	1100	1710	55.45
Night	912	1098	20.39

8.5 Conclusion

ADAS becomes quite important in these years for smart vehicle development. Moreover, having an autonomous car is not a far dream human beings immerse in. Vision-based object detection is an intuitive detection method similar to human visual perception, which is much low-cost compared with other detection methods

such as RADAR or LiDAR. However, current vision-based objects detection methods still suffer from several challenges such as high false alarm rate and unstable detection rate which limit their value in practical applications. In addition, luminance variation and weather change induce tougher challenges to vision-based object detection. Therefore, more research and development effort are necessary in this area.

In this chapter, we have introduced a set of rules in selecting proper training samples and presented a multi-pass self-correction training procedure to achieve an effective multiple moving objects detection system with high accuracy and low false alarm rates. Besides, FCWS and LDWS are proposed and implemented on Freescale i.MX6 with a monocular camera, which provides safety information to drivers. The dynamic threshold method conquers the problems resulted from different weather conditions, and the multiple frame approval reduces the effect of different scenes and windshield wiper. Moreover, the proposed LDWS and FCWS algorithms are integrated together in order to produce comprehensive information for day and night highway driving, which detects not only the forward vehicle, but also the potential cut-in vehicles. The implementation shows that the integrated system can achieve real-time processing for video input with D1 resolution. Furthermore, speed limit signs, which are significant traffic signs beside the road, are regarded as the primary targets to recognize. A detection system is designed to not only work under different weather conditions, but also achieve real-time processing performance based on single webcam. With more and more demands on the driving safety, future works can focus on using the blob and breach features on recognition of licenses plates and supporting other traffic signs based on the shape detection which can make good use of videos from car cam recorders. In addition, DLCE method can help ADAS system obtain better results at inclement weather conditions. The significant improvements of DLCE are proved in terms of detection accuracy of LDWS and pedestrian detection system. The proposed method has been implemented on TREK-668 embedded platform with 1.6 GHz Intel Atom N2600 CPU and tested with many inclement weather conditions. The performance of the prototype is able to achieve 50 fps at D1 video input.

Currently, these works can extend the functions of car cam recorders, making it actively ensure the safety of drivers. Although there are lots of challenges ahead, the dream of autonomous car will eventually come true by continuous efforts on enhancing existing functions and enlarging the ability of detecting object.

References

Pedestrian, Motorcyclist, and Vehicle Detection System (PMD)

1. Dollar P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
2. Ni B, Yan S, Wang M, Kassim AA, Qi T (2013) High-order local spatial context modeling by spatialized random forest. IEEE Trans Image Process 22(2):739–751

3. Pedersoli M, Gonzalez J, Hu X, Roca X (2014) Toward real-time pedestrian detection based on a deformable template model. IEEE Trans Intell Transp Syst 15(1):355–364
4. Bing S, Su S, Li S, Yun C, Ji R (2013) Decomposed human localization in personal photo albums. Proc. visual communications and image processing (VCIP), p 1–6
5. Tosato D, Spera M, Cristani M, Murino V (2013) Characterizing humans on Riemannian manifolds. IEEE Trans Pattern Anal Mach Intell 35(8):1972–1984
6. Yang Y, Ramanan D (2013) Articulated human detection with flexible mixtures of parts. IEEE Trans Pattern Anal Mach Intell 35(12):2878–2890
7. Choi J, Jung C, Lee J, Kim C (2014) Determining the existence of objects in an image and its application to image thumbnailing. IEEE Trans Signal Process Lett 21(8):957–961
8. Ablavsky V, Sclaroff S (2011) Layered graphical models for tracking partially occluded objects. IEEE Trans Pattern Anal Mach Intell 33(9):1758–1775
9. Lillywhite K, Lee D-J, Tippetts B (2012) Improving evolution-constructed features using speciation for general object detection. Proc. IEEE workshop applications of computer vision (WACV), p 441–446.
10. Satpathy A, Jiang X, Eng H-L (2014) Human detection by quadratic classification on subspace of extended histogram of gradients. IEEE Trans Image Process 23(1):287–297
11. Farhadi M, Motamedi SA, Sharifian S (2011) Efficient human detection based on parallel implementation of gradient and texture feature extraction methods. Proc. machine vision and image processing (MVIP), p 1–5
12. Enzweiler M, Gavrila DM (2011) A multilevel mixture-of-experts framework for pedestrian classification. IEEE Trans Image Process 20(10):2967–2979
13. Wang X, Wang M, Li W (2014) Scene-specific pedestrian detection for static video surveillance. IEEE Trans Pattern Anal Mach Intell 36(2):361–374
14. Prest A, Schmid C, Ferrari V (2012) Weakly supervised learning of interactions between humans and objects. IEEE Trans Pattern Anal Mach Intell 34(3):601–614
15. Yao BZ, Nie BX, Liu Z, Zhu S-C (2014) Animated pose templates for modeling and detecting human actions. IEEE Trans Pattern Anal Mach Intell 36(3):436–452
16. Wu J, Hu D (2014) Learning effective event models to recognize a large number of human actions. IEEE Trans Multimedia 16(1):147–158
17. Yao B, Li F-F (2012) Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Trans Pattern Anal Mach Intell 34 (9):1691–1703
18. Farhadi A, Sadeghi MA (2013) Phrasal recognition. IEEE Trans Pattern Anal Mach Intell 35 (12):2854–2865
19. Rakate GR, Borhade SR, Jadhav PS, Shah MS (2012) Advanced pedestrian detection system using combination of Haar-like features, Adaboost algorithm and Edgelet-Shapelet. IEEE International Computational Intelligence and Computing Research Conf. (ICCIC), Coimbatore, Dec 2012, p 1–5
20. Gressmann M, Palm G, Lohlein O (2011) Surround view pedestrian detection using heterogeneous classifier cascades. IEEE International Intelligent Transportation Systems Conf. (ITSC), Washington, DC, Oct 2011, p 1317–1324
21. Prioletti A, Mogelmose A, Grisleri P, Trivedi MM, Broggi A, Moeslund TB (2013) Part-based pedestrian detection and feature-based tracking for driver assistance: real-time, robust algorithms, and evaluation. IEEE Int Intell Transp Syst 14(3):1346–1359
22. Li Y, Li B, Bin T, Yao Q (2013) Vehicle detection based on the and- or graph for congested traffic conditions. IEEE Trans Intell Transp Syst 14(2):984–993
23. Lv Y, Yao B, Wang Y, Zhu S-C (2012) Reconfigurable templates for robust vehicle detection and classification. Proc. IEEE workshop applications of computer vision (WACV), p 321–328
24. Enzweiler M, Hummel M, Pfeiffer D, Franke U (2012) Efficient Stixel-based object recognition. Proc. IEEE conf. intelligent vehicles symposium (IV), p 1066–1071

25. Niknejad HT, Takeuchi A, Mita S, McAllester D (2012) On-road multivehicle tracking using deformable object model and particle filter with improved likelihood estimation. IEEE Trans Intell Transp Syst 13(2):748–758
26. Feris R, Datta A, Pankanti S, Sun M-T (2013) Boosting object detection performance in crowded surveillance videos. Proc. IEEE workshop applications of computer vision (WACV), p 427–432
27. Sivaraman S, Trivedi MM (2013) A review of recent developments in vision-based vehicle detection. Proc. IEEE conf. intelligent vehicles symposium (IV), p 310–315
28. Chunpeng W, Lijuan D, Jun M, Faming F, Xuebin W (2009) Detection of front-view vehicle with occlusions using AdaBoost. IEEE International Engineering and Computer Science, ICIECS 2009, Wuhan, 19–20 Dec 2009, p 1–4
29. Bobo D, Wei L, Pengyu F, Chunyang Y, Xuezhi W, Huai Y (2009) Real-time on-road vehicle and motorcycle detection using a single camera. IEEE international industrial technology, Gippsland, Feb 2009, VIC, p 1–6
30. Chen K-H, Ju T-F, Lu W-M, Guo J-I (2014) Vision-based multiple moving objects detection for intelligent automobiles. Int J Electr Eng 21(6):201–213
31. Choi MJ, Torralba A, Willsky AS (2012) A tree-based context model for object recognition. IEEE Trans Pattern Anal Mach Intell 34(2):240–252

Lane Departure Warning System

32. Lindner P, Blokzyl S, Wanielik G, Scheunert U (2010) Applying multi-level processing for robust geometric lane feature extraction. Proc. 2010 I.E. conference on multisensor fusion and integration for intelligent systems (MFI), Sept 2010, p 248–254
33. Lin Q, Han Y, Hahn H (2010) Real-time lane detection based on extended edge-linking algorithm. Proc. 2010 2nd international conference on computer research and development, May 2010, p 725–730
34. Chang CY, Lin CH (2012) An efficient method for lane-mark extraction in complex condition. Proc. international conference on intelligence and computing and international conference on autonomic and trusted computing, Sept 2012, p 330–336
35. Yoo H, Yang U, Sohn K (2013) Gradient-enhancing conversion for illumination-robust lane detection. IEEE Trans Intell Transp Syst 14(3):1083–1094
36. Ge PS, Guo L, Xu GK, Zhang RH, Zhang T (2012) A real-time lane detection algorithm based on intelligent CCD parameters regulation. Publishing in Discrete Dynamics in Nature and Society, Discret Dyn Nat Soc p 1–16
37. Huo CL, Yu YH, Sun TY (2012) Lane departure warning system based on dynamic vanishing point adjustment. Proc. of 2012 I.E. 1st conference on consumer electronics, p 25–28

Forward Collision Warning System

38. Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: a review. IEEE Trans Pattern Anal Mach Intell 28(5):694–711
39. Kumar U (2013) Vehicle detection in monocular night-time gray-level videos. Proc. 28th international conference in image and vision computing, New Zealand, Nov 2013, p 214–219
40. Teoh S, Brunl T (2012) Symmetry-based monocular vehicle detection system. Mach Vis Appl 23:831–842

41. Fossati A, Schönmann P, Fua P (2011) Real-time vehicle tracking for driving assistance. Mach Vis Appl 22(2):439–448
42. Khairdoost N, Monadjemi SA, Jamshidi K (2013) Front and rear vehicle detection using hypothesis generation and verification. Signal Image Process 4(4):31–50
43. Park K-Y, Hwang S-Y (2014) Robust range estimation with a monocular camera for vision-based forward collision warning system. Sci World J 2014:1–9

Speed Limit Detection System

44. Barnes N, Loy G (2006) Real-time regular polygonal sign detection. Springer Tracts Adv Robot 25:55–66
45. Loy G, Bames N (2004) Fast shape-based road sign detection for a driver assistance system. Proc. IEEE/RSL international conference on intelligent robots and systems, 28 Sept–2 Oct 2004
46. Keller CG, Sprunk C, Bahlmann C, Giebel J, Baratoff G (2008) Real-time recognition of U.S. speed signs. Proc. 2008 I.E. intelligent vehicles symposium, The Netherlands, 4–6 June 2008
47. Abukhait J, Zyout I, Mansour AM (2013) Speed sign recognition using shape-based features. Int J Comput Appl 84(15):32–37
48. Liu W, Lv J, Gao H, Duan B, Yuan H, Zhao H (2011) An efficient real-time speed limit signs recognition based on rotation invariant feature. Proc. 2011 I.E. intelligent vehicles symposium (IV), Baden-Baden, 5–9 June 2011
49. Lin H-H (1998) Recognition of printed digits of low resolution. Thesis, Institute of Computer and Information Science, National Chiao Tung University
50. Sharma R, Jain A, Sharma R, Wadhwa J (2013) Character and digit recognition aided by mathematical morphology. Int J Comput Technol Appl 4(5):828–832
51. Moutarde F, Bargeton A, Herbin A, Chanussot L (2007) Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular traffic signs recognition system. Proc. 2007 I.E. intelligent vehicles symposium, Istanbul, 13–15 June 2007
52. Glavtchev V, Muyan-Özçelik P, Ota JM, Owens JD (2011) Feature-based speed limit sign detection using a graphics processing unit. Proc. 2011 I.E. intelligent vehicles symposium (IV), Baden-Baden, 5–9 June 2011
53. Torresen J, Bakke JW, Sekanina L (2004) Efficient recognition of speed limit signs. Proc. 2004 I.E. intelligent transportation systems conference, Washington, DC, 3–6 Oct 2004
54. Chen L, Li Q, Li M, Mao Q (2011) Traffic sign detection and recognition for intelligent vehicle. Proc. 2011 I.E. intelligent vehicles symposium (IV), Baden-Baden, Germany, 5–9 June 2011
55. Zaklouta F, Stanciulescu B (2014) Real-time traffic sign recognition in three stages. Robot Auton Syst 62(1):16–24

Inclement Weather Processing Technology (DLCE)

56. Schechner YY, Narasimhan SG, Nayar SK (2001) Instant dehazing of images using polarization. Proc IEEE Conf Comput Vis Pattern Recognit 1:325–332
57. Shwartz S, Namer E, Schechner Y (2006) Blind haze separation. Proc IEEE Conf Comput Vis Pattern Recognit 2:1984–1991

58. Tan RT (2008) Visibility in bad weather from a single image. Proc. IEEE conf. on computer vision and pattern recognition, CVPR 2008, June 2008, p 1–8
59. Fattal R, 2008 (2008) Single image dehazing. ACM Trans Graph 27(3):1–9
60. He K, Sun J, Tang X (2009) Single image haze removal using dark channel prior. Proc. IEEE conf. on computer vision and pattern recognition, CVPR 2009, June 2009, p 1956–1963
61. He K, Sun J, Tang X (2013) Guided image filtering. IEEE Trans Pattern Anal Mach Intell 35 (6):1397–1409
62. Park G-H, Cho H-H, Choi M-R (2008) A contrast enhancement method using dynamic range separate histogram equalization. IEEE Trans Consum Electron 54(4):1981–1987
63. Zhiming W, Jianhua T (2006) A fast implementation of adaptive histogram equalization. 2006 8th international conference on signal processing, vol. 2
64. Pizer SM, Amburn EP, John DA, Robert C, Ari G, Trey G, Bart Ter Haar R, John BZ (1987) Adaptive histogram equalization and its variations. Proc Comput Vis Graph Image Process 39 (3):355–368
65. Zuiderveld K (1994) Contrast limited adaptive histogram equalization. Publishing in Graphics gems. In: Graphics gems IV. p 474–485
66. Narasimhan SG, Nayar SK (2003) Contrast restoration of weather degraded images. IEEE Trans Pattern Anal Mach Intell 25:713–724
67. Xu H, Guo J, Liu Q, Ye L (2012) Fast image dehazing using improved dark channel prior. Proc. IEEE conf. on information science and technology, ICIST, March 2012, p 663–667
68. Lv X, Chen W, Shen I (2010) Real-time dehazing for image and video. Proc. Pacific conf. on computer graphics and applications, PG, Sept 2010, p 62–69

Part IV
Smart Sensors for Biomedical and Health Monitoring

Chapter 9
Implantable Optical Neural Interface

Sang Beom Jun and Yoonseob Lim

Abstract For more than several decades, due to the rapid development of sophisticated electronics, the electrical neural interface has become the most popular method for recording and modulating neural activity in nerve systems. The electrical neural interface has been successfully applied to implantable neural prosthetic systems such as cochlear implant, deep brain stimulation system, artificial retina, and so on. Recently, in order to overcome the limitations of electrical methods for neural interface, novel optical technologies have been developed and applied to neuroscience research. Overall, the optical neural interfaces can be categorized into the intrinsic and the extrinsic methods depending on the modification of natural nerve system. For example, infrared neural stimulation (INS) and optogenetic neural stimulation are the typical methods for intrinsic and extrinsic neural interfaces, respectively. In addition to the optogenetic stimulation, it is also possible to monitor neural activity from specific neurons genetically modified to express activity-correlated fluorescence signals. Therefore, the optogenetic neural recording enables the detection of activity from specific types of neurons. Despite the fascinating advantages, to date, the use of the optical neural interface is limited only to the neuroscience research not for clinical purposes. In this chapter, the state-of-art technologies in optical neural interface are reviewed from the aspects of both neurobiology and engineering. In addition, the challenges to realize the clinical use of optical neural interfaces are discussed.

Keywords Optical neural interface • Infrared neural stimulation • Optogenetics

S.B. Jun (✉)
Department of Electronics Engineering, Ewha Womans University,
Seoul 120-750, South Korea

Department of Brain & Cognitive Sciences, Ewha Womans University,
Seoul 120-750, South Korea
e-mail: juns@ewha.ac.kr

Y. Lim
Center for Robotics Research, Robotics and Media Institute, Korea Institute of Science and Technology, Seoul, South Korea

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_9

9.1 Introduction

Since Santiago Ramón y Cajal demonstrated the richness of anatomical structures in nervous system, neuroscientists have been trying to understand the relationship between anatomical structures and functions of neural circuits. It would be undoubtable that the interaction of neural populations in different areas of brain determines the behavior [1]. It is also known that there exist many distinct cell types that play different roles in perception, decision-making, and memory formation even from neighboring neural populations [1–4]. For long time, electrophysiological techniques have been the most popular research tools for monitoring and manipulating the activity of a single or population of cells in the target brain area because the ionic current flow in the nerve system can be easily detected by using relatively simple electronics.

The remarkable progress in the electrical engineering since 1950s has also accelerated the development of various electrophysiological methodologies such as patch clamps, multichannel extracellular recording, silicon-based microelectrode arrays, and so on. The advances in electrophysiology have benefited not only the neuroscience research but also the neural interfaces for clinical purposes, which make interconnection between nerve systems and external electrical systems in order to help people with neurological diseases or disabilities. For example, the cochlear implant system is the first successful implantable neural prosthetic system commercialized from 1980s [5]. To date, the cochlear implant systems have been implanted to more than several hundreds of thousands people with profound hearing loss. Deep brain stimulation system is the second most successful implantable neural prosthetic device, which delivers repetitive electrical stimulation pulses in the deep brain region in order to restore normal motor functions of patients with Parkinson's disease, essential tremor, obsessive–compulsive disorder, and so forth. Recently, artificial retina system enabled the restoration of damaged vision for blind people. As the knowledge on the neurological diseases has accumulated and the technologies have rapidly developed in relevant engineering fields such as electronics, mechanics, and biomaterials, the implantable neural prosthetics are being an emerging technology for treatments of a variety of neurological diseases including psychiatric symptoms as well as sensory and motor disabilities.

While the commercialized neural prosthetic systems take advantages of electrical stimulation on the population of neurons, the recent state-of-arts technologies in neural interface aim to realize the connection with individual neurons to achieve precise monitoring and control of neuronal functions. These attempts have been undergone as brain–computer interface (BCI) or brain–machine interface (BMI). Recently, it was shown that it is feasible to precisely control the fine movement of a robot arm utilizing the information from individual neurons in motor cortices of quadriplegia patients.

In order to make the connection with a number of individual neurons, it is essential to use a multichannel electrode array that is fabricated via semiconductor thin-film processes. For more than several decades, various types of high-density

microelectrode arrays have been developed and applied to neuroscience researches. There have also been numerous evidences that monitoring electrical activity of single or multiple neuron and electrical stimulation can be used to understand and manipulate the function of neural circuits or even create artificial sensory perceptions [6–9]. To date, the conventional multichannel electrode array can detect the signals from a large number of neurons via up to several hundred channels. However, the electrically recorded neural signals do not convey the information on the neuronal cell types even though different types of neurons have distinct roles in the signaling in the brain. In addition, electrical stimulation also has a limitation of non-specific activation of different neuronal cell populations of the brain because electrical current can diffuse away toward nearby cells. Accordingly, there has been a growing need for different approaches for monitoring and manipulating the activities of specific neuronal cell types in different brain areas.

Electrical neural interfaces are not the only method for detecting or modulating the neural activity. Even though the electrical method has been the main technology for connection with nerve systems, there have also been a number of attempts to employ optical methods for neural interface. In terms of optical detection of neuronal activity, various phenomena accompanying neuronal activation have been identified and detected via optical means. Depending on whether the phenomena are intrinsic biological ones or extrinsically modified ones, the optical detections are divided into intrinsic and extrinsic optical neural recordings. Intrinsic optical signal (IOS) detection is to measure endogenous physiological changes associated with neuronal activity [10]. Labeling with fluorescence is not necessary because IOS takes advantages of reflection, transmittance, or scattering of incident light in blood flow, oxygenation of hemoglobin, cellular volume, and the membrane potential responding to physiological changes in the brain [11–17]. Although the intrinsic signals are mostly obtained from the surface of brain, fiber-optic techniques enable the detection from deeper brain regions combined with laser Doppler flowmetry (LDF), near-infrared (NIR) spectrometer, functional optical coherence tomography (fOCT), and surface plasmon resonance (SPR).

The intrinsic optical method is not applied only to monitoring the neural activity but also to neural modulation. For example, infrared neural stimulation (INS) enables to control neural activities by applying the light energy to the target neural cells [18–21]. While the exact underlying mechanism is not clarified yet, it was shown that infrared light can excite or inhibit neuronal cells depending on the thermal gradient and capacitance change across the cell membrane [22–27]. However, since the conventional intrinsic optical modulation uses infrared wavelength for longer penetration depth, it is inevitable to cause tissue damage due to strong water absorption [28]. In order to overcome this problem, it is proposed to use plasmonic nanoheater combined with INS. The surface plasmonic resonance interaction between metallic nanoheater and incident infrared light elevates the local temperature near the nanoheaters located at neuronal membrane. This intrinsic technique has a potential for optical activation or inhibition of neuronal metabolisms as well as modulation of the neuronal activity [29–32]. Even though the intrinsic optical measurement or modulation of neural activity provides a valuable

tool to monitor the activation of single neuron or brain region, there are drawbacks to become widespread for neuroscience research such as low signal to noise ratio, expensive and large size systems for miniaturization.

Another approach to achieve optical neural interface is to use exogenous optical probes or actuator based on fluorescence proteins. Over the last few decades, in neuroscience and biomedical engineering fields, the introduction of fluorescence has enabled the rapid advance in fluorescent microscopy for imaging biological structures. In the early studies, fluorescence microscopy is used mainly for morphological observation of tissues and cells. The introduction of green fluorescent protein (GFP) and the identification of gene sequences have accelerated the development of advanced microscopy technologies like two-photon microscopy and microendoscopy [33–35]. Combining genetic targeting techniques such as modification of genome, viral vector injection, or cell transfections, the fluorescence optics has become an essential tool for identifying specific type of cells or subcellular structures.

The combination of fluorescence and genetics has not only advanced the morphological imaging but also provided powerful tools for neural interface. The key principle is to optically modulate or monitor the activity of genetically selected neurons in the brain using light with specific wavelength, which is called optogenetics. Optogenetics is defined as a technique to control cells in living tissue, typically nervous tissues that are genetically modified to express light-sensitive ion channels such as neural activity-dependent fluorescence markers or to introduce light-sensitive opsins, light-sensitive proteins, which can control opening of channels on the membrane. The recent development of genetically encoded Ca^{2+}-activity dependent fluorescence has changed the paradigm of neuroscience research from electricity-based monitoring to more dynamic experiments where electrical activity of hundreds of neurons can be indirectly measured through two-photon microscopy at the same time [36–40]. In recent studies, head-mountable and miniaturized optical instruments were invented so that neuronal activities in the animal brain can be measured while animal behaves freely during behavior tasks [41–51].

The intrinsic optical technologies enabled a number of novel biomedical research as well as applications for biomedical imaging such as neural infrared spectroscopy (NIRS) or optical coherence tomography (OCT). However, in terms of monitoring or modulating individual neuronal activities, they are still far from the implantable neural interfaces. Compared with intrinsic optical methods, the extrinsic optogenetic methods can provide cell-type specificity and a higher efficiency for both stimulation and monitoring due to genetic modification and use of fluorescence. Due to these advantages, even though it has been only a single decade since the optogenetic method was first applied to neuroscience research, the related technologies are rapidly developing in various aspects such as opsins, fluorescence indicators, implantable systems, gene delivery methods, and even ethical issues for clinical applications. This chapter is organized to introduce the current technologies for implantable optical neural interfaces mostly based on the optogenetic technology. First, we are going to review the different types of microscopy technology for

monitoring activities of neurons. Second, optical modulation technique will be reviewed in terms of biological and engineering aspects. Last, the existing challenges in various aspects are discussed to achieve the implantable optical system for clinical purpose.

9.2 Optical Monitoring of Neural Activity

As the optogenetic techniques enabled cell type-specific targeting, the activity-dependent fluorescence proteins such as calcium indicators and voltage-sensitive fluorescence proteins have become very popular in neuroscience studies [52]. Combined with the genetic modification, these extrinsic optical signals are also becoming a powerful means to examine neuronal signaling from targeted neurons in culture, brain slice, and living animal [53, 54]. However, in order to monitor the neuronal activity from specific brain regions not from neuronal cultures or slices, it is crucial to use specialized fluorescence detection or imaging systems. Recently, various types of implantable fluorescence detection systems have been proposed for in vivo animal experiments. Overall, there are two different strategies for the instrumentation. One is to detect fluorescence intensity through implanted optic fibers. The other approach is to use fluorescence endoscopy techniques. Since the former is detecting the fluorescence intensity near the individual optic fiber tips implanted in the brain region of interest, it is suitable for detection of deep brain regions. The latter is useful to acquire the fluorescence time-lapse images in a specific visual field. For both approaches, it is obvious that the ideal implantable optical monitoring requires the development and the improvement of the optical probes converting action potentials into the fluorescence change, as well as the development of the implantable systems for fluorescence detection. In this section, the basic mechanisms and the current state of the in vivo extrinsic optical monitoring are reviewed in terms of the activity-dependent optical probes and the instrumentations.

9.2.1 Optical Probes for Neural Recording

It is well known that the intracellular calcium ions play an important role in a variety of means such as muscle contraction, neuronal transmission, apoptosis, cellular movement, cellular growth, and so forth [55]. In presynaptic terminals, it is notable that exocytosis of neurotransmitter is triggered by calcium concentration increase. In postsynaptic regions in dendritic spines, the calcium transient is also involved in synaptic plasticity [56]. In addition, calcium signaling in nucleus is known to regulate the gene transcriptions [57]. More importantly for monitoring neuronal activity, at the resting state, the intracellular calcium concentration is maintained less than 100 nM in most neurons while the concentration rises up to

100 times higher during action potential generation [58]. Accordingly, the recording of intracellular calcium transient signal is conventionally used to monitor the neural activity [59].

In order to detect the neural activity-dependent calcium transient signals, there have been a number of attempts to develop optimized optical probes expressing the intracellular calcium signals as fluorescence changes. The first calcium indicator was proposed based on calcium-binding photoproteins such as aequorin [35]. Later, more sensitive fluorescent calcium indicators became available using the binding chemical reaction between a fluorescent chromophore and calcium-selective chelators like ethylene glycol tetraacetic acid (EGTA) or calcium-specific aminopolycarboxylic acids like BAPTA (1,2-bis(o-aminophenoxy)ethane-N,N,N',N'-tetraacetic acid). Based on this chemical reaction mechanism, to date, several calcium indicators with different excitation spectra and affinities to calcium have been developed including quin-2, fura-2, fluo-3, fluo-4, Oregon Green BAPTA, and so on [60–63]. Since these indicators are easy to use and have relatively large signal to noise ratios, they are currently widely used in neurophysiological studies. In order to load the calcium indicators into the intracellular domain, in early days, the dyes were delivered inside the individual neurons using a sharp metal electrode or glass micropipettes. These techniques are still conventionally used for basic studies of calcium signaling in neurons. To load the indicator dyes into the intracellular domains of neurons, the popular method is the use of membrane-permeable acetoxymethyl (AM) ester forms. The AM calcium dye molecules are trapped inside the cells after the hydrophobic eater residues are removed during the entry into the intracellular cytosol. As the application for in vivo studies, the simple application of AM calcium dyes to the brain tissue results in the loading region with the diameter of several hundred micrometers.

The essential breakthrough of the use of calcium indicator as an optical probe for neural activity was the genetically encoded calcium indicators (GECIs). GECIs enable the measurement of the calcium transient signals from the genetically identified neurons. Among several different methods to make neurons expressing fluorescence GECIs, the most popular one is currently the use of viral transduction using stereotaxic injection into the targeted brain areas. Among various viral vectors, lenti-, adeno-, and adeno-associated viral vectors are commonly used to deliver the GECI constructs into the neurons of interest [64, 65]. The specific expression can be achieved via the use of the specific genetic promoters or the transgenic animals with the Cre recombinase driver [66]. In addition, the production of transgenic mice with stable GECI expression has been also possible recently because it would facilitate the experiment procedure tremendously [67]. Even though there have been a few successes, the generation of GECI transgenic animal is not always possible for the specific types of neurons because the chemical reaction between calcium ions and indicator molecules can interfere the endogenous calcium signaling.

Another popular optical probe is voltage-sensitive dyes. Voltage-sensitive dyes are also fluorescence probes which alter their intensity in response to the exerted electrical voltage changes [68]. Since a number of physiological processes are

accompanied by changes in the membrane potential, the fluorescence measurement can be used as an optical indicator if the dye molecules are properly located near the cellular membrane. In principle, the voltage-sensitive fluorescent protein (VSFP) has a paired structure of hydrocarbon chains as the anchors and a hydrophilic group for alignment of the chromophore. It is believed that the chromophore shows an electrochromic mechanism during change of the electric field. Compared to calcium indicators, voltage-sensitive fluorescence proteins were first introduced more lately and highlighted due to the fast response time and the linear measurement of membrane potential. Recently, it was reported that the genetically encoded voltage-sensitive probes were successfully inserted into voltage-gated potassium or sodium channels [69]. Even though those probes were able to express fluorescence change proportional to the membrane potential, the efficiency of targeting specific membrane was not stable and the signal to noise ratio was not high enough to be detected in vivo environment. By leading researchers, a new generation of VSFPs was developed such as *Ciona intestinalis* voltage-sensitive-containing phosphate (Ci-VSP) based on Förster resonance energy transfer (FRET) sensors, which showed the stable fluorescence dynamics responding membrane potential signaling [70]. However, there are still challenges to overcome in order to use the VSFP as an in vivo optical probe. First of all, low-noise fluorescence signal needs to be maintained for long time since there exist several noise sources such as mechanical noises, hemodynamic noises, and crosstalk noise in in vivo imaging environments [54]. Second, the present in vivo VSFP imaging is limited only to the cortex regions of the brain due to the limitation of the detection system [71].

9.2.2 Optical Systems for Neural Recording

The instrumentation for fluorescence detection requires the light-sensing device and the appropriate light source for the excitation. Unlike in vitro conditions, in vivo fluorescence detection requires the stable and high sensitivity equipment since the fluorescence-emitting target cells are compactly surrounded with other neighboring cells and the scattering effect for exciting light and emitted light is very severe. In addition, the physical movement of brain can deteriorate the signal detections due to the motion artifacts in freely moving animals or clinical applications. If the neurons of interests exist at the superficial cortical layers, the optical signal can be obtained using typical detecting approaches such as photodiode arrays, charged coupled detector (CCD)-based cameras, or complementary metal-oxide-semiconductor (CMOS)-based cameras [72, 73]. Imaging at a deeper region in the brain than the superficial area usually employs confocal or two-photon microscopy. These techniques are also available for long-term experiments in combination with the formation of the chronic window or thinned skull preparations. However, these methods have the limitation due to the shallow imaging depth because the deeper region imaging requires the stronger excitation light and the fluorescence signals also will scatter away before detected from the outside system.

Therefore, in order to avoid this problem, the most realistic way is the implantation of optic fibers for both excitation light delivery and fluorescence detection or the endoscopic approach which requires the insertion of a fiber bundle or gradient refractive index (GRIN) lenses [74, 75]. Since these methods can decrease the distance between the optics and the target fluorescent proteins, the higher SNR can be achieved at the expense of the invasiveness even though tissue damages are not unavoidable.

Other than the methods mentioned above, various optical techniques can be applied for in vivo measurement of fluorescence signaling mostly for reduction of noise and increased sensitivity. Electron multiplying (EM)-CCDs can significantly improve the SNR via on-chip multiplication and higher cooling performance. Another approach for high sensitive detection is to use time-correlated single photon counting (TCSPC) system [75]. The frame rate of the imaging system is also crucial especially for detection of voltage-sensitive calcium signal since the voltage transients-based fluorescence change is as fast as action potentials unlike calcium indicators. In general applications of VSFP, the minimal frame rates are approximately 100–1000 Hz while imaging the calcium transients requires less than 100 Hz due to the slow response time. In addition, there are increasing attempts to develop miniaturized head-mountable imaging devices for measurement in freely moving animals. For example, recently, a miniaturized system included objective, dichroic mirror, and PMT [45].

In this section, the basic principles for fluorescence microscopy are described below. Then the attempts to miniaturize the microscopy systems for implantable neural recording systems are discussed.

9.2.2.1 Miniaturized Microscopy

Microscopic system described above typically includes large optic table and telescopic system that make it impossible to use for in vivo imaging in freely behaving animals. Various forms of miniaturized microscopic system have been explored to construct a small microscope so that an animal can carry the system while behaving freely in the experiment setting. These include the miniaturized design of microendoscopes, confocal microscopes, and two-photon microscopic systems [17–27]. Key technological progress in microscope objectives, laser pulse delivery method, and miniaturized scanning system has been essential to build those small microscopes.

9.2.2.2 Confocal Microscopy

In conventional microscopy, a thin section of tissue is placed on the microscope and entire area of the specimen is illuminated to obtain the image through objective lens. However, using traditional microscopic system, other parts of the tissue can also be illuminated, thereby resulting in a blurred image with background noise.

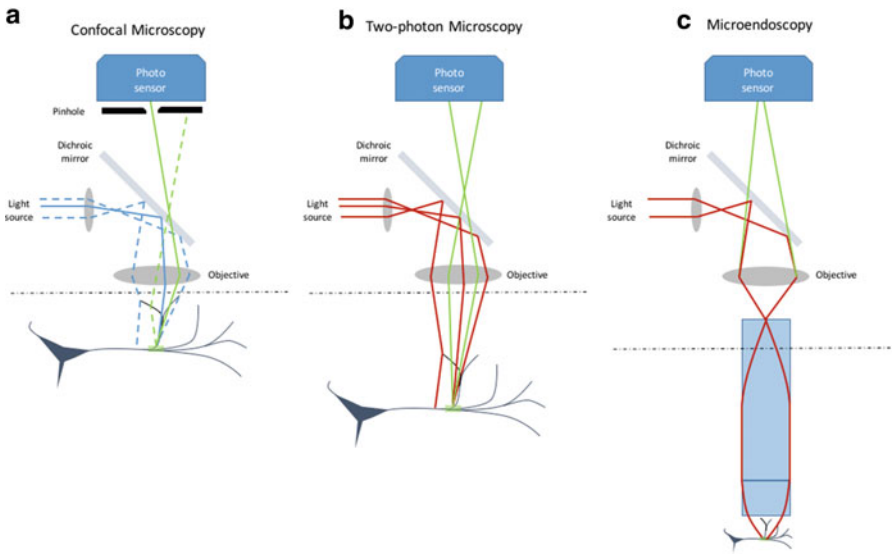


Fig. 9.1 Schematic drawings of different microscopic systems. (**a**) Light scattering pathways in confocal microscopy are shown. Only photons that follow the *solid lines* can pass pinhole and reach the photo sensor. (**b**) In two-photon microscopy, photons near the *focused point* can reach the photo sensor. (**c**) Microendoscopy uses the needle-like lens to get images in much deeper area than two-photon or confocal microscopy

Such a degradation of image quality in resolution and contrast due to light scattering makes even harder to image the three-dimensional biological structures inside the living tissues. Confocal microscopy, developed by Marvin Minsky, had opened up a new possibility of acquiring a high-resolution, high-contrast image in depth [76]. In his original design, to avoid unwanted scattered light from tissue, he illuminated the specimen with focused light through objective and gathered returning light through an another pinhole aperture that may block rays that are incoming from out-of-focus areas.

To understand how the confocal microscopy works, an example of several light rays for imaging is illustrated in scattering medium in Fig. 9.1a. Solid lines represent light rays from the focal point in the specimen whereas dashed lines represent the path of scattered photons. Placing the screen with pinhole at the opposite side of lens system, only photons that follow the solid lines will pass the pinhole while scattered photons are rejected by the screen. It means that only small area of specimen could only be observed at a time in confocal microscopy. Therefore, to create the complete image of the specimen, photodetector builds up the image pixel by pixel mostly while scanning the specimen. Typically, image with 512×512 pixels can be created at a frame rate of 0.1–30 Hz. Since photons from out-of-plane unfocused light are blocked by the screen, confocal microscopes can create a very crisp image of the specimen, which is collection of images from thin planar region of the specimen (Fig. 9.2).

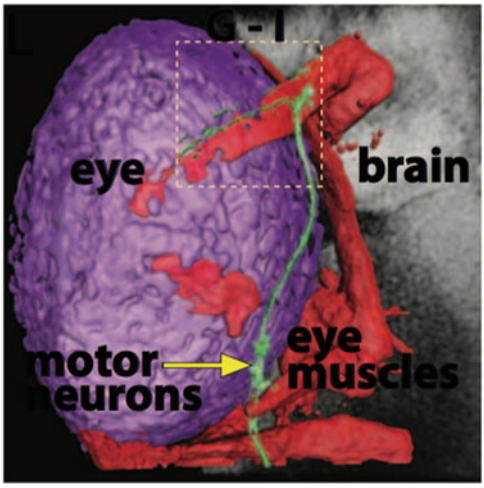


Fig. 9.2 Example image by confocal microscopy (adapted from [5])

The ability of the confocal microscopy for a sharp optical sectioning enables three-dimensional imaging of the specimen. Usually, data from each section along the optical axis are gathered to create the reconstructed 3D images. Confocal system collects only a small fraction of light rays that pass the pinhole aperture so a light source with very high intensity is required. Recently, laser light source based scanning system is widely used for more precise optical sectioning (400–750 nm wavelength) [77].

Even though the confocal microscopy has opened up a new era for fluorescence imaging with high resolution, no studies for miniaturized confocal microscopy for neural activity monitoring have been reported except the development of a new endoscopic system for imaging digestive organ or inspecting human skins [44–46]. The goal of these studies was the early detection and image based therapy of disease in hollow organs, including colon, esophagus, lung, oropharynx, and cervix [47].

9.2.2.3 Two-Photon Microscopy

Invention of confocal microscopy has provided a new window of research opportunity by creating high-resolution image of biological tissue whereas conventional microscopes could not resolve microscopic structures in thick specimens due to out-of-focus image from severe scattering. Despite the superior quality of resulting image, confocal microscopy has been restricted to in vitro preparations, such as cultured neurons or brain slices [77]. The main drawbacks of confocal microscopy are: (1) limited depth of imaging by the poor optical penetration (typically up to 200 μm), (2) photobleaching of fluorescent probes from excessive use of excitation, (3) low frame rate for imaging, and (4) high cost relative to the conventional microscopy.

Some of the above problems have been resolved with the introduction of two-photon microscopy (Fig. 9.1b) [78]. Basic component of two-photon microscopy is similar to that of confocal microscopy but two-photon system requires laser source with longer wavelength and no pinhole structure is necessary. First, optical penetration depth has become around 500 μm by adapting longer excitation wavelength of laser pulse (700–1050 nm) than that of confocal system. For commonly used fluorescent markers, range of photon absorption is in NIR wavelength whereas emission occurs in visible range. Using NIR excitation is useful to imaging in optically thick specimens. NIR light has less absorption in biological tissues than blue–green light but also has less scattering. So two-photon microscopy can create more localized excitation than confocal system does, which leads to high-resolution three-dimensional imaging into much deeper regions of tissue. Some researchers have even reported imaging depth of 1.6 mm from the surface of the mouse brain using 1280 nm two-photon excitation [79]. Finally, two-photon microscopy could have less photodamage. In confocal microscopy, there is severe deviation of light path and some photons from the focused region can also be blocked by the screen. This is why high intensity light source is used for confocal microscopy. But two-photon requires no pinhole configuration, thus less signal loss is expected. Low chance of photodamage in tissue could improve the viability of tissues, thereby increasing the possibility of long-term imaging.

These advantages indeed have made two-photon microscopy suitable for in vivo imaging in intact tissue and in animals. In neuroscience, two-photon microscopy has been applied in many studies regarding neural circuit dynamics from cellular level to population levels [36–40]. With the recent development of optogenetic tools and advanced surgical techniques, its wide application in recording and modulating the activity of neural system could also be possible.

Various forms of miniature two-photon microscopic system have been developed since the first head-fixed one was introduced in early 2000s [41–46]. Several major components are the same as those used for a standard two-photon microscope (see Fig. 9.1b). In both setups, an ultrafast laser source (up to 100 fsec laser pulses, wavelength: 700–1050 nm) is required for fluorescent excitation and telescope for laser beam size adjustment. Specific to the two-photon fiberscope, two parts are needed: (1) optic fiber coupler and (2) miniaturized front piece. Optic fiber coupler connects a single-mode or multimode optical fiber to the existing microscope system so that fluorescent photons can be transferred stably to the photon sensor that is placed on the standard two-photon system.

To make the miniaturized front piece of fiberscope, several components have to be miniaturized or redesigned.

1. A compact laser-scanning mechanism

   Most fiberscopes developed so far contain laser-scanning system at the microscope headpiece [41–43]. Several different approaches have been applied: cantilever fiber-scanners operating at resonant or non-resonant frequencies and microelectromechanical systems (MEMS) scanning mirrors. In Fig. 9.3, four different scanning methods are shown. Three of them employ the

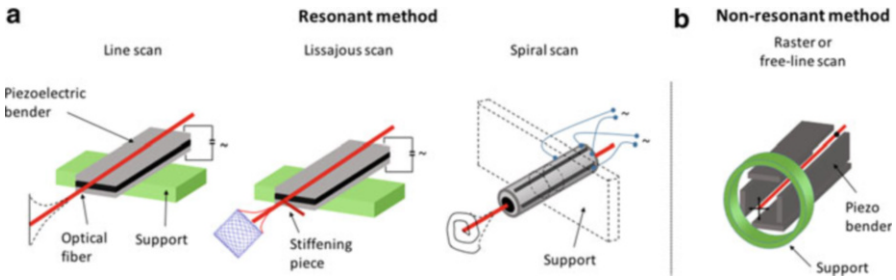


Fig. 9.3 Scanning method used in miniaturized two-photon microscopy. (**a**) Voltage profiles applied on the piezoelectric plate can create different vibration pattern of optical fiber at its resonant frequency. (**b**) Non-resonant method can create a raster or free-line scan of optical fiber

mechanical resonant or non-resonant property of cantilever with the piezoelectric system (Fig. 9.3a).

In resonant setup, the illuminating fiber is attached to the piezoelectric bender so that 10–20 mm of fiber is free to vibrate. By applying modulated voltage signal to the piezoelectric bender, different scanning trajectories can be achieved at the tip of illuminating fiber. However, resonant scanning method does not provide lateral offset or random access scanning. To overcome this problem, non-resonant frequency based scanning method was also proposed (Fig. 9.3b) [43]. In this setup, by adjusting the frequency of the fiber tip, raster-like scanning and random access scanning can be obtained. Recently, MEMS-based scanning mirrors with a modest zoom ability and a flexibility in scan rate adjustment were also applied [38].

2. Miniaturized objective

A water-immersion objective lens is necessary for high-resolution in vivo imaging because water has higher refractive index than air, thereby increasing the numerical aperture (NA) of objective. However, standard water-immersion objective cannot be mounted on the head of small animal due to its size and weight. The first fiberscope utilized this kind of objective but with the advancement in micro-optical components, Gradient-Index (GRIN) lens have been used in two-photon fiberscopes with limited NA (0.48–0.58) [42, 44, 46]. The recent development of GRIN lens with plano-convex lens (NA: 0.85) could also be applied.

3. Efficient transmission of laser pulse

Unlike standard two-photon microscopy setup, animal has to carry the miniaturized front piece that is originally a component of standard two-photon microscope. This means that secure optical path to deliver the laser pulse has to be constructed. With standard optical fiber, for example, step-index single-mode fiber, ultrafast laser pulse gets broaden because of material dispersion and nonlinear effect from high laser powers. Light dispersion could be alleviated through a negative prechirp by passing laser light through the

grating pair, which adjusts the arrival time of different color of light at the surface of specimen. In other words, this method makes red and blue component arrive at the tip of fiber at the same time. However, as the energy of laser pulse increases for deeper imaging, nonlinear effect on pulse broadening still remains. Recently, by applying hollow-core photonic crystal fiber, the problem of pulse broadening is almost resolved. In this fiber, light travels space filled with air and this creates less dispersion and pulse broadening at high laser power. Several fiberscopes utilized this kind of optical fiber and demonstrated successful delivery of laser pulses to the fiberscopes [42, 44, 46].

4. Fluorescence detection system

   There are two different ways to detect fluorescent photons: (1) gather photons using remote photon sensor system on the optical table, and (2) detect fluorescent light through photomultiplier tube (PMT) attached on the side of fiberscopes. In the first design of fiberscope, a small PMT unit was used [41]. Although this setup makes an efficient detection of fluorescent light, entire fiberscopes body becomes heavier (note that the first fiberscope weighs 25 g), which makes it unable for application to small animals like mice to carry the scope. However, most of the fiberscopes that separate the fluorescent detector system from fiberscopes are 3–5 g. In these systems, multimode optical fiber is often used to deliver the fluorescent light to the detector. Although there is loss of photons at the end of fiber, keeping detection unit apart could be useful for multi wavelength imaging of different fluorescent dyes.

Several studies have developed a small size two-photon microscopic system (0.6–25 g) but most of the devices were applied to a deeply anesthetized animal and do not have enough sensitivity to observe fast electrophysiological events. There is only one study that has shown its ability of Ca^{2+} imaging of neurons in visual cortex of freely behaving rats [45]. Weight of the proposed fiberscope was around 5.5 g and the system can monitor 16 different neurons at 11 Hz frame rate. To reduce the weight of the whole system mounted on animal's head, a coherent optical fiber bundle connected with scanning unit of standard two-photon microscopy has also been proposed. Although this approach is advantageous in terms of detection system, however, the resolution of the image is dependent on the quantity of optical fibers in the bundle.

9.2.2.4 Microendoscopy

Two-photon microscopy provides better optical sectioning than confocal microscopy does. However, optical penetration depth by two-photon microscopy is still around 500 μm. Therefore, it can only support imaging near the superficial area of

Table 9.1 Optical specifications of different microendoscopes

Microendoscope type	Diameter (mm)	Length (mm)	Field of view (μm)	lateral resolution (μm)
Doublet [81]	1 (objective)/2 (focusing unit)	1.4–6.2	240–370	2.8–3.9
Doublet [51]	0.5	9.86	200	0.85
Doublet [49]	0.35–1.0	11–35	75–275	0.86
GRIN/plano-convex doublet (LaSFN9 plano-convex lens) [82]	1.0	4.0	75	0.6

the brain. Although it has been shown that multiphoton microscopy can image the mouse brain up to 1.6 mm, the method is still in development phase [80] (Table 9.1).

Optical microendoscopy has been proposed to extend the optical penetration depth into tissue (Fig. 9.1c) [47–50]. The method uses needle-like micro-optical probe (diameter: 350–1000 μm), which acts like an optical relay and is inserted into the target area (up to 1 cm into tissue). Most of the micro-optical probe is made of GRIN lens. Its diameter, length, and numerical aperture values are dependent on the imaging condition [33]. Sometimes, two different GRIN lenses with distinctive NA values (objective and relay lens) could be combined for higher resolution and deeper insertion. Recently developed high-resolution micro lens can provide a spatial resolution close to the conventional water-immersion objective (up to 0.6 μm) [34].

There are two different ways of imaging tissues lying in deeper area from the surface of specimen with microendoscope: (1) micro lens is held by the surrounding tissue after insertion and only microscope objective is free to move for fine focus adjustment, and (2) both micro lens and microscope objective are fixed to separate movement device. They can move either independently or in tandem for focus control. Since the diameter of micro lens can be up to 2800 μm (note that typical electrode for electrophysiology is around 100–200 μm), great care has to be taken when inserting the probe into the tissue. In one study, they inserted GRIN lens (diameter: 350 μm) a few tens of micrometers in each step and paused between insertion steps to minimize the possible tissue damage and compression [48]. Possibility of long-term functional imaging of deep nuclei in the mouse brain (2.5–5.0 mm depth) was also explored for acute imaging with GRIN lens inserted through chronically implanted guide cannula made of thin polyimide film [51]. However, significant damage of the blood vessels along the insertion path cannot be avoided.

Diameter and length of micro lens are determined based on the experimental conditions. Different diameter of probes could change the resolution and magnification values. For example, wider probes at some NA value provide longer working distances and broader field of view. Length of microendoscopes probe depends on the depth of the tissue to be imaged. Typically, length of relay micro lens is adjusted

Table 9.2 Characteristics of miniaturized microendoscopes and two-photon microscope

Specification	Miniaturized microendoscope	Miniaturized two-photon microscope
Mass (g)	<4	>4
Frame rate (Hz)	30–100	1–36
Optical resolution (μm)	1.5	0.9
Image size (pixels)	~640 × 480	~720 × 480
Field of view (μm)	600 × 800	200

for substantial depth of imaging. Detailed design condition can be found in [49] and we have provided optical parameters of different microendoscopes in Table 9.2 (adapted from [33]).

Microendoscope has been applied to imaging through standard fluorescence microscope system (one-or two-photon microscopy) [47–50]. In several studies, miniaturized systems were also developed so that a small animal like rat or mice can carry the whole imaging setup mounted on the skull (see miniaturized microscopy section for the detail) [35–37]. Microendoscopy has shown its possibility in examining brain areas that have never been imaged before in live animals. In these studies, microendoscope is combined with the standard one- or two-photon microscope systems. Typically, animal's head is restrained by strong metal post on the experiment table during imaging session, which could alternate the state of neural activities in the brain. In recent studies, miniaturized version of endoscopes has been developed and successfully applied to small animals like mice and bird [27, 35–37, 41].

Configuration of miniaturized microendoscope is similar to that of miniaturized two-photon microscope except that light source is directly attached to the miniaturized system and fluorescent photons are collected through embedded CMOS system [36, 41, 51]. In Table 9.2, we have compared the key specifications of the two microscopy systems. Although many of studies that developed miniaturized two-photon microscope did not succeed in imaging the neural activities in behaving animal (except one study [41]), miniaturized microendoscope has been successfully applied to freely behaving animals [36, 41, 51]. This is because weight of the whole system is less than 4 g and frame rate is relatively high enough to monitor the Ca^{2+} dynamics of neural population than miniaturized two-photon system (see Table 9.2 for detail). Recently, several companies have introduced commercialized products and neuroscience society has shown their ability of imaging neural populations in behaving animal's brain [37, 42, 43]. Also, in one study, they have also built their own microendoscope and applied it to imaging the neural activities in HVC of song bird [41]. They have also made all the components of microscopes in public and even the software for data acquisition is available at https://github.com/WALIII/FreedomScope.

9.3 Optical Neural Modulation of Nerve Systems

As described above, using genetically encoded activity sensors, optical detection has become a powerful tool for monitoring the neural activity of action potentials. If the genetically encoded activity sensors are replaced with genetically encoded actuator controlling transmembrane ionic flow, then the optogenetic modulation (stimulation or inhibition) can be achieved. Since the optogenetic modulation has very distinct advantages from conventional electrical modulation in terms of cell-type specificity and less invasiveness, it can be an improved alternative for conventional neural prosthetic systems using electrical stimulation. The term optogenetics was also coined from the optogenetic control of neurons expressing light-sensitive channels and pumps. Similar to the optical detection of neural activity-dependent fluorescence, the optogenetic neuromodulation method consists of genetic modifications and optical systems. In this section, the development of opsins for neuromodulation is first overviewed. Then, the optics systems for implantable optogenetic control are reviewed.

9.3.1 Light-Sensitive Proteins

The earliest method that used light to control genetically selected neurons was reported in 2002. They employed rhodopsin photoreceptors of drosophila for controlling neural activity in cultured neurons [83]. In 2003, they also developed a method for light-dependent activation of neurons using ionotropic channels such as TRPV1 and TRPM8 combined with caged ligands responding to light [84]. In 2005, Miesenböck and colleagues reported the first photostimulation of genetically targeted neuronal channels in dopaminergic systems of fruit flies, resulting in controlling of the behavior of an animal [85]. In the same year, Karl Deisseroth and his colleagues firstly demonstrated the optogenetic control system using Channelrhodopsin-2 (ChR2) in cultured mammalian neurons [86, 87].

Channelrhodopsin is a subfamily protein of rhodopsin that functions as light-gated ion channels first founded in green algae [88]. Originally, channelrhodopsin plays a role to control alga's movement in response to light. However, when expressed in cells of other organisms, it enables the optical control of ionic flow across the membrane of cells including neurons. In resting state, the inside of neurons is negatively charged by the action of sodium-potassium pumps and selectively permeable membrane. If the light-sensitive channels or light-sensitive ion pumps induce ionic flux across the cellular membrane of neurons, the membrane potential will change, resulting in either activating or inhibiting the neural activity. Since the complementary DNA sequence of channelrhodopsin in the green alga *Chlamydomonas reinhardtii* is identified, it became possible to artificially make the cells to express the opsin-related proteins in the animals. In the early 2000s, it was demonstrated that various light-sensitive ion transport proteins

including light-driven pumps and light-gated ion channels like channelrhodopsin-1 (ChR1) and ChR2 can be expressed in oocytes of *Xenopus laevis* and mammalian cells. Later, these opsins became available to use in neuronal cells in mammals. One of the most important improvements in optogenetics was the introduction of fast light-activated channels. Even though light-sensitive channels or pumps act as a controller of ionic currents, the gating speed was not fast enough at early attempts. However, millisecond-scale temporal resolution of control was also realized by Deisseroth group in 2005 [86]. In addition to the optogenetic activation of neurons, inhibiting of neural activity became possible due to introduction of chloride ion pump, halorhodopsin (NpHR), and enhanced halorhodopsins (eNpHR2.0 and eNpHR3.0) [89, 90]. Compared to NpHR, eNpHR allows high-level expression in mammalian neurons without toxicity via screening a number of membrane trafficking modulators.

In this manner, various microbial opsins have been identified and modified to adapt to the mammalian nerve systems. The development of opsins has been achieved mostly in terms of high photosensitivity and expression levels, fast response time, and low toxicity. Recently, instead of immediate photoactivation, prolonged photoactivation also became possible via introduction of step-function opsins (SFO) and stabilized step-function opsins (SSFO) [91, 92]. They enabled a sustained photocurrent for longer duration approximately for 30 min. It offered a potential to slowly change the balance between excitation and inhibition for longer timescale, which might be a crucial treatment method for various neurological disorders such as epilepsy, Parkinson's disease, and so forth.

Once the light-sensitive opsins are ready to use, they should be delivered to the targeted neurons for successful neuromodulation. To date, most optogenetic approaches for stimulating or inhibiting neurons have been applied to mice due to the large number of transgenic strains available. The most popular ways to express the opsins at the target cells are Cre recombinase targeting and viral transfection. Cre recombinase is an enzyme derived from a bacteriophage, which is widely used in molecular biology in order to catalyze specific recombination sites located in genes. Cre recombination occurs with loxP recognition sites located near the target gene sequences, resulting in expression of the encoded opsins. Since various transgenic cell type-specific Cre mice are available, the Cre recombinase method became very popular in optogenetic research [93, 94]. Currently, combined with cell type-specific Cre/loxP genetic methods, optogenetics has been routinely used to activate or inhibit specific cell types of neurons in vivo [95].

However, it is inapplicable to clinical application due to the complex procedures. Compared to Cre recombinase method, viral transfection is more promising for opsin delivery. The viral transfection employs viruses as carriers to deliver the opsin genes. The most common viruses for this purpose are adeno-associated viruses (AAV) or lentiviruses. Recent studies showed that the opsin delivery through AAV is feasible for rodents as well as primates [96–98]. In the primate study, it was shown that the optogenetic control of neural activity is stable and no harmful effect on the tissue for more than a month even though the results in action potential appeared more complicated than conventional electrophysiology experiments. Besides, the FDA in USA approved the delivery of AAVs into the brain for clinical researches [99].

9.3.2 Optical Neuromodulation System

When the optogenetic technique was first introduced, the optical system was not customized for activation of the light-sensitive opsins. Rather, it just employed the light source equipped with fluorescence microscope for excitation of fluorophore. Since the power of the built-in light source are around 10 mW/mm^2 after filtering light in unwanted spectra, it was enough to activate the light-sensitive opsins in neuronal cultures or slice preparations [86, 87]. However, in order to apply optogenetic neuromodulation to in vivo studies, it is required to deliver the light with specific wavelength to brain structures located deep within the cortex. Since the light penetration through brain tissue is limited by scattering effect, it is obviously impossible to deliver the enough light energy to the target cells in deep brain region even with highly sensitive opsins [100, 101]. Therefore, it became very common to use penetrating optical waveguide structures also known as an optrode [102–105].

The easiest way to make a penetrating optrode is to use optical fibers coupled with a light source. Since the various types of optical fibers are commercially available, the use of optical fibers coupled with specific wavelength lasers or LEDs has become a popular method. However, for in vivo animal experiments, once the optical fibers are located into the target brain region, the fibers are firmly fixed commonly with biocompatible cement. Then, it is impossible to disconnect the optic fiber from either the light source or animals. Therefore, the use of optical connectors became popular, providing freedom to disconnect the light source while not delivering light. The connection is typically realized by combination of a ceramic sleeve and a patch cord of optic fiber connected with a metal ferrule. The sleeve is fixed on top of the skull through a skin and the patch cord is inserted into the brain while the other end is located in the sleeve with ferrule. This type of connector is called a fiber-optic cannula, which has been used for long time in optics. In addition, the optical commutator (also known as optical rotating joint) has also become available, which allows the animals with implanted optic fibers to rotate freely in a certain experiment space without twisting the fibers.

Another approach for optrodes is to integrate the optical waveguide with a microelectrode as shown in Fig. 9.4. Combining electrode arrays for electrical neural recording and optical waveguide for optogenetic stimulation enabled simultaneous optical stimulation and electrophysiological recording of target neurons. Zhang and his colleagues first demonstrated the dual modality hybrid system for optogenetic stimulation via a tapered optical waveguide and simultaneous multichannel recording via 100 channel intracortical multielectrode arrays [102, 103]. In order to implant the optrode and the electrode into the deeper region with minimal cortical damage, a novel method is proposed to form a coaxial optrode, a tapered, gold-coated optical fiber inserted in an insulation tube [105]. This device enables the electrophysiological recording through the exposed gold coating in addition to the light delivery optic fiber.

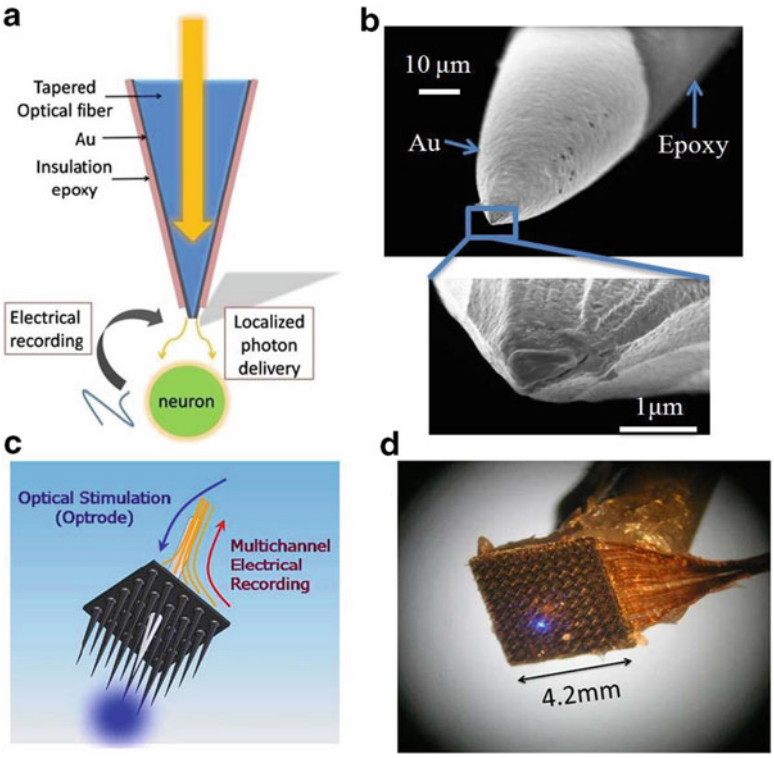


Fig. 9.4 Single optrode and optrode-MEA device. (**a**) Concept schematic of single optrode. (**b**) SEM images of the optrode tip. The exposed metallic part of the tip is approximately 50 μm, appearing brighter in the upper image. (**c**) Schematic of the hybrid device. The optrode is coupled to the MEA through a laser drilled hole. (**d**) An optical microscope image of the device, showing blue laser light emanating from the tip of the optrode [103] (Copyright©2009 IEEE)

As the optogenetic techniques rapidly develop, the optogenetics are not employed only for the neuroscience studies but also for the neuromodulation therapies. Even though there are several issues prior to the clinical use such as stability of opsin expression, miniaturized optical systems, and even ethical issues, there are attempts to use the optogenetics as a more effective neuromodulation tool for neuroprosthetic systems. However, the optical systems described above are implemented mainly for in vitro or in vivo animal experiments. In order to use optogenetics for clinical neuromodulation systems, the optical systems should be far more improved in terms of miniaturization, safety, and removal of external systems. First of all, the long connection between the optic fiber and light source should be minimized or removed. It has been already reported that LED light source can be integrated directly with the implanted optic fibers or waveguides [106]. As a prototype system, a miniaturized LED device for wireless control was also proposed to optogenetically stimulate cortical neurons via transcranial light delivery [107].

9.4 Challenges and Perspectives

In the present chapter, we have discussed several optical interface methods for in vivo or in vitro monitoring neurons and optogenetically modulating neurons. These techniques have enabled researchers to record activity of hundreds of different neurons at the same time or altering the behavior of population of neurons with specific cell types. However, current method can be used in the experimental conditions. In this section, we are going to discuss challenging issues so that optical interface method can be applied for long-term reliable experiments or even application to human patients.

9.4.1 Miniaturization of Optical Hardware

To optically monitor the activities of neurons, light from laser source should be delivered to the target region of the brain. In most cases, light is transferred through optical fiber. This means that animal should be tethered to the cable and this could restrain the behavior of animal. What is more, most miniaturized microscope still use the high-end optical systems such as ultrafast laser source and telescopic systems placed on the optic table during experiment. These devices are not easy to be miniaturized at this point. Unlike optical imaging system, optical modulation system can take an advantage of relatively small LED system as a light source and a wireless optical neural control system has been introduced recently [108]. This system weighs only 2 g and is equipped with wireless powering system as well.

9.4.2 Power Consumption

For standard optogenetics application, LED (Light Emitting Diode) has been widely used because it is cheap, easy to control, and could be easily incorporated with optic systems. However, the major disadvantage is a low light transmission efficiency. While the emitted power of standard LED can be up to 5 W, actual light reaching the surface of the target tissue is only in the order of a few milliwatts, which meets the light sensitivity of the cells expressing the opsin. In general, the power transmission efficiency is much less than 10 % from the light source to the tissue. It means that optical interface system cannot be run on a small battery and the power loss will case heat dissipation. Typical 3 V coin battery can discharge current less than 0.1 Ah and such a low power capacity is not enough to turn the LED on constantly. As suggested in recently developed wireless system [108], wireless power system with supercapacitor may be required. Laser system has much higher efficiency than LED (up to 50 %) but laser system is not easy to be miniaturized.

9.4.3 Heating Problem

Continuous application of light can increase the temperature of tissues in the brain. If the tissue at 300 μm from the surface were continuously receiving light (23.5 mW/mm^2), temperature of the brain will be risen up to 10 °C which is much more than the range of temperature fluctuation suggested by ISO standards (1 °C) [109]. However, in typical conditions for optogenetic modulation, pulsed light stimulation is delivered to the target tissues. For example, 20 Hz pulsed-modulated stimulation increases the temperature by only 0.5 °C. However, the optogenetic monitoring for the activity-dependent fluorescence change or the prolonged optical stimulation still requires the longer period of light exposure. Therefore, it would be important to design the minimal but effective light delivery protocol that may not significantly increase the tissue temperature. Since not only the light delivery but also the heat dissipation from waveguides, electronics, or batteries can induce the temperature increase, the whole implantable system should be carefully designed to decrease the possibility of temperature increase. Even with a slight increase of the tissue temperature, it has been known that homeostatic signaling can modulate the neural systems [110–115]. It was also reported that singing behavior of a song bird can be radically altered by directly cooling and warming the motor area, HVC [116].

9.4.4 Single-Cell Manipulation

In many optogenetics experiments, population of neurons are activated or silenced. Traditional optical interface technology is not enough to support an experiment for controlling individual cells comprising a neural network for specific behavior or perception. Recently, a novel tool was proposed to enable spatial control of light modulation and single-cell targeted experiment. The proposed system does not require any scanning devices under two-photon microscopy system maintaining a high spatial resolution [117, 118]. This technique has great advantage in activating ChR2 in a single cell and monitoring the network-wise changes by two-photon fluorescence microscopy. However, the technology is in the early developmental stage and far from the development of the implantable devices yet.

9.4.5 Clinical Application

To date, optogenetic neural interface methods have only been applied to mice, birds, and nonhuman primate [98, 119, 120]. To utilize the optogenetics technique for human clinical application, several key issues still remain to be resolved. First, efficient way to express the viral product in human brain has to be established.

Currently, fewer number of cell types-specific gene sequences of the human brain are known. Extensive basic research on anatomical and functional grouping of cell types is required to select clinically effective cells in the human brain. Another issue is the longevity of expression level for microbial product in the human brain. To our knowledge, this property has not been characterized yet except the long-term effect on the morphological changes of neurons in rat brain [121]. Finally, a miniaturized fully implantable optical interface system powered by a battery has to be developed so that all the essential optical and electronics can be chronically implanted inside human brain or human body. Considering current technologies for optic systems and batteries, in order to develop a commercialized implantable optical neural interface, the power consumption of current systems should be lowered at least more than 10 times.

9.5 Conclusions

During the last decade, the optogenetics enabled new neuromodulation and monitoring technologies from genetically targeted neurons. It has furthered the fundamental scientific understanding in neuroscience research regarding how specific types of neurons perform the function in the complicated neural networks. Moreover, on the clinical side, optogenetics research has led to insights into several neurological diseases and psychiatric disorders such as Parkinson's disease, autism, schizophrenia, drug abuse, depression, and so forth. Due to these precedent meaningful results, the clinical use of optogenetics is already predicted by pioneers. Considering the possibility of clinical optogenetic technology, in this chapter, extrinsic optical neural interface systems are reviewed in terms of monitoring and modulating the neural activity in vivo. As described in this chapter, there are a number of critical hurdles to overcome such as miniaturization of optical systems, gene delivery for opsin expression, thermal and biological safety, high power consumption, and heat dissipation issues. However, as the optogenetic technology has been rapidly improved in terms of optical systems as well as opsins and optical probes, the remaining issues will be also resolved in near future. Therefore, it is very likely that the optogenetic neural interface can be a futuristic therapeutic tool connecting nerve systems and external systems beyond its scientific impact.

Acknowledgments This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT and Future Planning as the Global Frontier Project (CISS-2012M3A6A6054204) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2014R1A2A2A09052449, 2014R1A1A1A05003770).

References

1. Shepherd GM (2004) The synaptic organization of the brain, 5th edn. Oxford University Press, New York
2. Pinaud R, Terleph TA, Tremere LA, Phan ML, Dagostin AA, Leão RM et al (2008) Inhibitory network interactions shape the auditory processing of natural communication signals in the songbird auditory forebrain. J Neurophysiol 100:441–455
3. Zhou M, Liang F, Xiong XR, Li L, Li H, Xiao Z et al (2014) Scaling down of balanced excitation and inhibition by active behavioral states in auditory cortex. Nat Neurosci 17:841–850
4. Pinto L, Dan Y (2015) Cell-type-specific activity in prefrontal cortex during goal-directed behavior. Neuron 87:437–450
5. Roche JP, Hansen MR (2015) On the horizon: cochlear implant technology. Otolaryngol Clin North Am 48:1097–1116
6. Kilgard MP, Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. Science 279:1714–1718
7. Zrenner E (2002) Will retinal implants restore vision? Science 295:1022–1025
8. Zhang Y, Hakes JJ, Bonfield SP, Yan J (2005) Corticofugal feedback for auditory midbrain plasticity elicited by tones and electrical stimulation of basal forebrain in mice. Eur J Neurosci 22:871–879
9. Lim H, Anderson D (2006) Auditory cortical responses to electrical stimulation of the inferior colliculus: implications for an auditory midbrain implant. J Neurophysiol 96:975–988
10. Thompson AC, Stoddart PR, Jansen ED (2014) Optical stimulation of neurons. Curr Mol Imaging 3:162–177
11. Frostig RD, Lieke EE, Ts'o DY, Grinvald A (1990) Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in vivo high-resolution optical imaging of intrinsic signals. Proc Natl Acad Sci 87:6082–6086
12. Malonek D, Grinvald A (1996) Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. Science 272:551
13. Kim SA, Byun KM, Lee J, Kim J, Kim D-G, Baac H et al (2008) Optical measurement of neural activity using surface plasmon resonance. Opt Lett 33:914–916
14. Kim SA, Kim SJ, Moon H, Jun SB (2012) In vivo optical neural recording using fiber-based surface plasmon resonance. Opt Lett 37:614–616
15. Lazebnik M, Marks DL, Potgieter K, Gillette R, Boppart SA (2003) Functional optical coherence tomography for detecting neural activity through scattering changes. Opt Lett 28:1218–1220
16. Stepnoski R, LaPorta A, Raccuia-Behling F, Blonder G, Slusher R, Kleinfeld D (1991) Noninvasive detection of changes in membrane potential in cultured neurons by light scattering. Proc Natl Acad Sci 88:9382–9386
17. Shevelev IA (1998) Functional imaging of the brain by infrared radiation (thermoencephaloscopy). Prog Neurobiol 56:269–305
18. Wells J, Kao C, Jansen ED, Konrad P, Mahadevan-Jansen A (2005) Application of infrared light for in vivo neural stimulation. J Biomed Opt 10:064003
19. Tan X, Rajguru S, Young H, Xia N, Stock SR, Xiao X et al (2015) Radiant energy required for infrared neural stimulation. Sci Rep 5:13273
20. Wells J, Konrad P, Kao C, Jansen ED, Mahadevan-Jansen A (2007) Pulsed laser versus electrical energy for peripheral nerve stimulation. J Neurosci Methods 163:326–337
21. Izzo AD, Walsh JT Jr, Ralph H, Webb J, Bendett M, Wells J et al (2008) Laser stimulation of auditory neurons: effect of shorter pulse duration and penetration depth. Biophys J 94:3159–3166
22. Huang H, Delikanli S, Zeng H, Ferkey DM, Pralle A (2010) Remote control of ion channels and neurons through magnetic-field heating of nanoparticles. Nat Nanotechnol 5:602–606

23. Richter CP, Matic AI, Wells JD, Jansen ED, Walsh JT Jr (2011) Neural stimulation with optical radiation. Laser Photonics Rev 5:68–80
24. Shapiro MG, Homma K, Villarreal S, Richter CP, Bezanilla F (2012) Infrared light excites cells by changing their electrical capacitance. Nat Commun 3:736
25. Wells J, Kao C, Konrad P, Milner T, Kim J, Mahadevan-Jansen A et al (2007) Biophysical mechanisms of transient optical stimulation of peripheral nerve. Biophys J 93:2567–2580
26. Duke AR, Jenkins MW, Lu H, McManus JM, Chiel HJ, Jansen ED (2013) Transient and selective suppression of neural activity with infrared light. Sci Rep 3:2600
27. Katz EJ, Ilev IK, Krauthamer V, Kim do H, Weinreich D (2010) Excitation of primary afferent neurons by near-infrared light in vitro. Neuroreport 21:662–666
28. Jaque D, Martinez Maestro L, del Rosal B, Haro-Gonzalez P, Benayas A, Plaza JL et al (2014) Nanoparticles for photothermal therapies. Nanoscale 6:9494–9530
29. Carvalho-de-Souza JL, Treger JS, Dang B, Kent SB, Pepperberg DR, Bezanilla F (2015) Photosensitivity of neurons enabled by cell-targeted gold nanoparticles. Neuron 86:207–217
30. Eom K, Kim J, Choi JM, Kang T, Chang JW, Byun KM et al (2014) Enhanced infrared neural stimulation using localized surface plasmon resonance of gold nanorods. Small 10:3853–3857
31. Yong J, Needham K, Brown WG, Nayagam BA, McArthur SL, Yu A et al (2014) Gold-nanorod-assisted near-infrared stimulation of primary auditory neurons. Adv Healthc Mater 3:1862–1868
32. Yoo S, Hong S, Choi Y, Park JH, Nam Y (2014) Photothermal inhibition of neural activity with near-infrared-sensitive nanotransducers. ACS Nano 8:8040–8049
33. Barnard JE, Welch FV (1936) Fluorescence microscopy with high powers. J R Microsc Soc 56:361–364
34. Abramowitz M (1993) Fluorescence microscopy. Olympus America, New York
35. Shimomura O, Johnson FH, Saiga Y (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea. J Cell Comp Physiol 59:223–239
36. Helmchen F, Denk W (2005) Deep tissue two-photon microscopy. Nat Methods 2:932–940
37. Kerr JND, Denk W (2008) Imaging in vivo: watching the brain in action. Nat Rev Neurosci 9:195–205
38. Rothschild G, Nelken I, Mizrahi A (2010) Functional organization and population dynamics in the mouse primary auditory cortex. Nat Neurosci 13:353–360
39. Roberts TF, Tschida KA, Klein ME, Mooney R (2010) Rapid spine stabilization and synaptic enhancement at the onset of behavioural learning. Nature 463:948–952
40. Clancy KB, Koralek AC, Costa RM, Feldman DE, Carmena JM (2014) Volitional modulation of optically recorded calcium signals during neuroprosthetic learning. Nat Neurosci 17:807–809
41. Helmchen F, Fee MS, Tank DW, Denk W (2001) A miniature head-mounted two-photon microscope. Neuron 31:903–912
42. Flusberg BA, Jung JC, Cocker ED, Anderson EP, Schnitzer MJ (2005) In vivo brain imaging using a portable 3.9 gram two-photon fluorescence microendoscope. Opt Lett 30:2272–2274
43. Sawinski J, Denk W (2007) Miniature random-access fiber scanner for in vivo multiphoton imaging. J Appl Phys 102:034701
44. Engelbrecht CJ, Johnston RS, Seibel EJ, Helmchen F (2008) Ultra-compact fiber-optic two-photon microscope for functional fluorescence imaging in vivo. Opt Express 16:5556–5564
45. Sawinski J, Wallace DJ, Greenberg DS, Grossmann S, Denk W, Kerr JND (2009) Visually evoked activity in cortical cells imaged in freely moving animals. Proc Natl Acad Sci U S A 106:19557–19562
46. Piyawattanametha W, Cocker ED, Burns LD, Barretto RPJ, Jung JC, Ra H et al (2009) In vivo brain imaging using a portable 2.9 g two-photon microscope based on a microelectromechanical systems scanning mirror. Opt Lett 34:2309–2311

47. Jung JC, Schnitzer MJ (2003) Multiphoton endoscopy. Opt Lett 28:902–904
48. Levene MJ, Dombeck DA, Kasischke KA, Molloy RP, Webb WW (2004) In vivo multiphoton microscopy of deep brain tissue. J Neurophysiol 91:1908–1912
49. Jung JC, Mehta AD, Aksay E, Stepnoski R, Schnitzer MJ (2004) In vivo mammalian brain imaging using one- and two-photon fluorescence microendoscopy. J Neurophysiol 92:3121–3133
50. Llewellyn ME, Barretto RPJ, Delp SL, Schnitzer MJ (2008) Minimally invasive high-speed imaging of sarcomere contractile dynamics in mice and humans. Nature 454:784–788
51. Bocarsly ME, Jiang W-c, Wang C, Dudman JT, Ji N, Aponte Y (2015) Minimally invasive microendoscopy system for in vivo functional imaging of deep nuclei in the mouse brain. Biomed Opt Express 6:4546–4556
52. Grienberger C, Konnerth A (2012) Imaging calcium in neurons. Neuron 73:862–885
53. Cao G, Platisa J, Pieribone VA, Raccuglia D, Kunst M, Nitabach MN (2013) Genetically targeted optical electrophysiology in intact neural circuits. Cell 154:904–913
54. Mutoh H, Perron A, Akemann W, Iwamoto Y, Knopfel T (2010) Optogenetic monitoring of membrane potentials. Exp Physiol 96:13–18
55. Gonzalez D, Espino J, Bejarano I, Lopez JJ, Rodriguez AB, Pariente JA (2010) Caspase-3 and -9 are activated in human myeloid HL-60 cells by calcium signal. Mol Cell Biochem 333:151–157
56. Zhang H, Liu J, Sun S, Pchitskaya E, Popugaeva E, Bezprozvanny I (2015) Calcium signaling, excitability, and synaptic plasticity defects in a mouse model of Alzheimer's disease. J Alzheimers Dis 45:561–580
57. Lyons MR, West AE (2011) Mechanisms of specificity in neuronal activity-regulated gene transcription. Prog Neurobiol 94:259–295
58. Berridge MJ, Lipp P, Bootman MD (2000) The versatility and universality of calcium signalling. Nat Rev Mol Cell Biol 1:11–21
59. Wachowiak M, Knopfel T (2009) Frontiers in neuroscience optical imaging of brain activity in vivo using genetically encoded probes. In: Frostig RD (ed) In vivo optical imaging of brain function. CRC Press and Taylor & Francis Group, LLC, Boca Raton, FL
60. Paredes RM, Etzler JC, Watts LT, Zheng W, Lechleiter JD (2008) Chemical calcium indicators. Methods 46:143–151
61. Pozzan T, Arslan P, Tsien RY, Rink TJ (1982) Anti-immunoglobulin, cytoplasmic free calcium, and capping in B lymphocytes. J Cell Biol 94:335–340
62. Tsien RY, Pozzan T, Rink TJ (1982) Calcium homeostasis in intact lymphocytes: cytoplasmic free calcium monitored with a new, intracellularly trapped fluorescent indicator. J Cell Biol 94:325–334
63. Neher E (1995) The use of fura-2 for estimating Ca buffers and Ca fluxes. Neuropharmacology 34:1423–1442
64. Dittgen T, Nimmerjahn A, Komai S, Licznerski P, Waters J, Margrie TW et al (2004) Lentivirus-based genetic manipulations of cortical neurons and their optical and electrophysiological monitoring in vivo. Proc Natl Acad Sci U S A 101:18206–18211
65. Monahan PE, Samulski RJ (2000) Adeno-associated virus vectors for gene therapy: more pros than cons? Mol Med Today 6:433–440
66. Wirth D, Gama-Norton L, Riemer P, Sandhu U, Schucht R, Hauser H (2007) Road to precision: recombinase-based targeting technologies for genome engineering. Curr Opin Biotechnol 18:411–419
67. Tsai PS, Friedman B, Ifarraguerri AI, Thompson BD, Lev-Ram V, Schaffer CB et al (2003) All-optical histology using ultrashort laser pulses. Neuron 39:27–41
68. Chemla S, Chavane F (2010) Voltage-sensitive dye imaging: technique review and models. J Physiol Paris 104:40–50
69. Siegel MS, Isacoff EY (1997) A genetically encoded optical probe of membrane voltage. Neuron 19:735–741

70. Akemann W, Mutoh H, Perron A, Rossier J, Knöpfel T (2010) Imaging brain electric signals with genetically targeted voltage-sensitive fluorescent proteins. Nat Methods 7:643–649
71. Akemann W, Sasaki M, Mutoh H, Imamura T, Honkura N, Knöpfel T (2013) Two-photon voltage imaging using a genetically encoded voltage indicator. Sci Rep 3:2231
72. Ross WN, Werman R (1987) Mapping calcium transients in the dendrites of Purkinje cells from the guinea-pig cerebellum in vitro. J Physiol 389:319–336
73. Baker BJ, Kosmidis EK, Vucinic D, Falk CX, Cohen LB, Djurisic M et al (2005) Imaging brain activity with voltage- and calcium-sensitive dyes. Cell Mol Neurobiol 25:245–282
74. Adelsberger H, Garaschuk O, Konnerth A (2005) Cortical calcium waves in resting newborn mice. Nat Neurosci 8:988–990
75. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM et al (2013) Concurrent activation of striatal direct and indirect pathways during action initiation. Nature 494:238–242
76. Minsky M (1988) Memoir on inventing the confocal scanning microscope. Scanning 10:128–138
77. Pawley JB (ed) (2006) Handbook of biological confocal microscopy. Springer, Boston, MA
78. Denk W, Strickler J, Webb WW (1990) Two-photon laser scanning fluorescence microscopy. Science 248:73–76
79. Kobat D, Horton NG, Xu C (2011) In vivo two-photon microscopy to 1.6-mm depth in mouse cortex. J Biomed Opt 16:106014
80. Horton NG, Wang K, Kobat D, Clark CG, Wise FW, Schaffer CB et al (2013) In vivo three-photon microscopy of subcortical structures within an intact mouse brain. Nat Photonics 7:205–209
81. Flusberg BA, Nimmerjahn A, Cocker ED, Mukamel EA, Barretto RPJ, Ko TH et al (2008) High-speed, miniaturized fluorescence microscopy in freely moving mice. Nat Methods 5:935–938
82. Barretto RPJ, Messerschmidt B, Schnitzer MJ (2009) In vivo fluorescence imaging with high-resolution microlenses. Nat Methods 6:511–512
83. Zemelman BV, Lee GA, Ng M, Miesenbock G (2002) Selective photostimulation of genetically chARGed neurons. Neuron 33:15–22
84. Zemelman BV, Nesnas N, Lee GA, Miesenbock G (2003) Photochemical gating of heterologous ion channels: remote control over genetically designated populations of neurons. Proc Natl Acad Sci U S A 100:1352–1357
85. Lima SQ, Miesenbock G (2005) Remote control of behavior through genetically targeted photostimulation of neurons. Cell 121:141–152
86. Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. Nat Neurosci 8:1263–1268
87. Li X, Gutierrez DV, Hanson MG, Han J, Mark MD, Chiel H et al (2005) Fast noninvasive activation and inhibition of neural and network activity by vertebrate rhodopsin and green algae channelrhodopsin. Proc Natl Acad Sci U S A 102:17816–17821
88. Nagel G, Ollig D, Fuhrmann M, Kateriya S, Musti AM, Bamberg E et al (2002) Channelrhodopsin-1: a light-gated proton channel in green algae. Science 296:2395–2398
89. Zhao S, Cunha C, Zhang F, Liu Q, Gloss B, Deisseroth K et al (2008) Improved expression of halorhodopsin for light-induced silencing of neuronal activity. Brain Cell Biol 36:141–154
90. Gradinaru V, Thompson KR, Deisseroth K (2008) eNpHR: a Natronomonas halorhodopsin enhanced for optogenetic applications. Brain Cell Biol 36:129–139
91. Berndt A, Yizhar O, Gunaydin LA, Hegemann P, Deisseroth K (2009) Bi-stable neural state switches. Nat Neurosci 12:229–234
92. Yizhar O, Fenno LE, Prigge M, Schneider F, Davidson TJ, O'Shea DJ et al (2011) Neocortical excitation/inhibition balance in information processing and social dysfunction. Nature 477:171–178
93. Madisen L, Mao T, Koch H, Zhuo JM, Berenyi A, Fujisawa S et al (2012) A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. Nat Neurosci 15:793–802

94. Madisen L, Garner AR, Shimaoka D, Chuong AS, Klapoetke NC, Li L et al (2015) Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. Neuron 85:942–958
95. Tsien JZ, Chen DF, Gerber D, Tom C, Mercer EH, Anderson DJ et al (1996) Subregion- and cell type-restricted gene knockout in mouse brain. Cell 87:1317–1326
96. Bass CE, Grinevich VP, Vance ZB, Sullivan RP, Bonin KD, Budygin EA (2010) Optogenetic control of striatal dopamine release in rats. J Neurochem 114:1344–1352
97. Witten IB, Lin SC, Brodsky M, Prakash R, Diester I, Anikeeva P et al (2010) Cholinergic interneurons control local circuit activity and cocaine conditioning. Science 330:1677–1681
98. Han X (2012) Optogenetics in the nonhuman primate. Prog Brain Res 196:215–233
99. Carter BJ (2005) Adeno-associated virus vectors in clinical trials. Hum Gene Ther 16:541–550
100. Lin JY, Knutsen PM, Muller A, Kleinfeld D, Tsien RY (2013) ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. Nat Neurosci 16:1499–1508
101. Chuong AS, Miri ML, Busskamp V, Matthews GA, Acker LC, Sorensen AT et al (2014) Noninvasive optical inhibition with a red-shifted microbial rhodopsin. Nat Neurosci 17:1123–1129
102. Zhang J, Laiwalla F, Kim JA, Urabe H, Van Wagenen R, Song YK et al (2009) Integrated device for optical stimulation and spatiotemporal electrical recording of neural activity in light-sensitized brain tissue. J Neural Eng 6:055007
103. Zhang J, Laiwalla F, Kim JA, Urabe H, Van Wagenen R, Song YK et al (2009) A microelectrode array incorporating an optical waveguide device for stimulation and spatiotemporal electrical recording of neural activity. Conf Proc IEEE Eng Med Biol Soc 2009:2046–2049
104. Chen S, Pei W, Gui Q, Chen Y, Zhao S, Wang H et al (2013) A fiber-based implantable multioptrode array with contiguous optical and electrical sites. J Neural Eng 10:046020
105. Ozden I, Wang J, Lu Y, May T, Lee J, Goo W et al (2013) A coaxial optrode as multifunction write-read probe for optogenetic studies in non-human primates. J Neurosci Methods 219:142–154
106. Cao H, Gu L, Mohanty SK, Chiao JC (2013) An integrated muLED optrode for optogenetic stimulation and electrical recording. IEEE Trans Biomed Eng 60:225–229
107. Iwai Y, Honda S, Ozeki H, Hashimoto M, Hirase H (2011) A simple head-mountable LED device for chronic stimulation of optogenetic molecules in freely moving mice. Neurosci Res 70:124–127
108. Wentz CT, Bernstein JG, Monahan P, Guerra A, Rodriguez A, Boyden ES (2011) A wirelessly powered and controlled device for optical neural control of freely-behaving animals. J Neural Eng 8:046021
109. Warden MR, Cardin JA, Deisseroth K (2014) Optical neural interfaces. Annu Rev Biomed Eng 16:103–129
110. Davis GW (2006) Homeostatic control of neural activity: from phenomenology to molecular design. Annu Rev Neurosci 29:307–323
111. Davis GW, Goodman CS (1998) Genetic analysis of synaptic development and plasticity: homeostatic regulation of synaptic efficacy. Curr Opin Neurobiol 8:149–156
112. Marder E, Prinz AA (2002) Modeling stability in neuron and network function: the role of activity in homeostasis. Bioessays 24:1145–1154
113. Ginsberg MD, Sternau LL, Globus MY, Dietrich WD, Busto R (1992) Therapeutic modulation of brain temperature: relevance to ischemic brain injury. Cerebrovasc Brain Metab Rev 4:189–225
114. Pérez-Otaño I, Ehlers MD (2005) Homeostatic plasticity and NMDA receptor trafficking. Trends Neurosci 28:229–238
115. Turrigiano GG, Nelson SB (2000) Hebb and homeostasis in neuronal plasticity. Curr Opin Neurobiol 10:358–364

116. Long MA, Fee MS (2008) Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature 456:189–194
117. Andrasfalvy BK, Zemelman BV, Tang J, Vaziri A (2010) Two-photon single-cell optogenetic control of neuronal activity by sculpted light. Proc Natl Acad Sci U S A 107:11981–11986
118. Papagiakoumou E, Anselmi F, Bègue A, de Sars V, Glückstad J, Isacoff EY et al (2010) Scanless two-photon excitation of channelrhodopsin-2. Nat Methods 7:848–854
119. Roberts TF, Gobes SMH, Murugan M, Olveczky BP, Mooney R (2012) Motor circuits are required to encode a sensory model for imitative learning. Nat Neurosci 15:1454–1459
120. Yizhar O, Fenno LE, Davidson TJ, Mogri M, Deisseroth K (2011) Optogenetics in neural systems. Neuron 71:9–34
121. Miyashita T, Shao YR, Chung J, Pourzia O, Feldman DE (2013) Long-term channelrhodopsin-2 (ChR2) expression can induce abnormal axonal morphology and targeting in cerebral cortex. Front Neural Circuits 7:8

Chapter 10
Real-Time Programmable Closed-Loop Stimulation/Recording Platforms for Deep Brain Study

Hung-Chih Chiu and Hsi-Pin Ma

Abstract Biomedical systems have expanded markedly in recent years, spreading into many areas of human life. Rapid advances in biological science have led to the creation of novel electrical circuits and signal processing methods and the development of tools for diagnosing and treating human diseases. Many biomedical engineering researchers have developed novel tools designed to tackle specific medical conditions.

The instruments used represent an interface between biology and electronics. These interfaces enable biological phenomena to be quantified and characterized, thus allowing the biological processes underlying them to be elucidated. A typical interface comprises a sensor or electrode for detecting some biological parameter, the signals from which are then amplified and converted into a digital form. These digital data can be processed by hardware or transferred to a personal computer for closed-loop control, long-term storage, and more precise signal processing. The guidelines for such signal processing algorithms require low complexity, short latency, high sensitivity, and accurate characterization. Microprocessors are used to make the design of an electronic algorithm flexible and adaptable. Depending on the requirements of a specific application, the data can be transferred through wired or wireless links. Communication can be achieved using widely available and clearly defined technical specifications.

This chapter discusses the main hardware and software components used in closed-loop deep brain stimulation systems and describes the evaluation procedures that are used to ensure that the system performs as specified. Even when the system parameters can change with the physiological characteristics, a closed-loop control system can accurately extract the signals of interest.

Keywords Neural recording/stimulation • Closed-loop stimulation platform • Deep brain stimulation • Parkinson's disease • Time-frequency signal processing • Phase analysis • Phase synchrony • EEG • Local field potential

H.-C. Chiu • H.-P. Ma (✉)
Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
e-mail: hp@ee.nthu.edu.tw

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_10

10.1 Introduction

Bioelectronic systems are used to measure and quantify physiological parameters within the human body and to treat certain medical conditions. The interface with the human body makes it possible to investigate the biological world and the function of human body systems. Compared with conventional biomedical electronic systems based on digital signal processing (DSP) units, a real-time electronic system can provide a programmable approach to recording and stimulating the system, low-complexity DSP, and a closed-loop strategy for recording feedback from the stimulated nuclei.

Electrical stimulation systems have been used to study the behavior of neurons in the brain [1–4]. Various therapies use deep brain stimulation (DBS) to alleviate neurological disorders and to treat Parkinson's disease [5], tremors [6], epilepsy [7, 8], depression [9], cluster headaches [10, 11], and other maladies [12, 13]. In order for DBS therapy to be effective, the electrical pulses used must be of high frequency. Many previous studies have quantified the energy used in neural activities. This has been done by calculating the power density in a particular bandwidth of the brain, because neural synchrony exhibits large variability in amplitudes and recurrence. There is increasing evidence to suggest that deep electrical stimulation of brain structures suppresses neuronal synchrony at the basal ganglia (BG) [14]. Suppression of neuronal synchrony in Parkinson's disease usually involves open-loop deep electrical stimulation of brain structures using local field potentials (LFPs) [15]. Open-loop deep electrical stimulation delivers preprogrammed electric signal patterns, but an effective feedback loop for maintaining the neurotransmitters cannot be selected. By contrast, a closed-loop stimulation strategy has a considerably stronger effect while preserving battery life and allowing precise control of the functional electrical stimulation of the brain. However, closed-loop stimulation strategies are not yet well enough understood to allow the selection of the stimulation conditions. Few guidelines are available covering the selection of appropriate closed-loop stimulation strategies, in which the signal characteristic changes in the power spectra of the LFPs [16], or the phase is synchronous in the specific frequency band [17]. Closed-loop stimulation strategies must take account of power consumption and battery life. A long-term objective is to develop a system that can automatically adjust the stimulation strategies to achieve suppression of neuronal synchrony based on the electrical signals from the deep brain. Closed-loop stimulation platforms for deep brain study are discussed in detail in this chapter, which is organized as follows.

The remainder of this chapter is organized as follows: design considerations for closed-loop stimulation strategies are discussed in Sect. 10.2. Standard closed-loop DBS and recording system design are described in Sect. 10.3. Section 10.4 discusses closed-loop control policies and DSP. Section 10.5 presents an integrated electronic and signal processing system, including system architecture, firmware design, mathematical instruments for measuring neural activity, and closed-loop neural phase synchrony detection. Section 10.6 offers some conclusions.

10.2 Considerations for Closed-Loop System Design

In closed-loop DBS, bioelectrical signals are used to extract information from a biological system. These signals exhibit different electrical characteristics, including different levels of complexity, because an organ comprises multiple tissue types and functional units, but is composed of cells that have common features. The electrical signal received will reflect these similar cell features, but intracellular variation will have an effect on the electrical properties. The electrical properties can be captured by devices as simple as a closed-loop device on the skin surface or microelectrodes placed in direct contact with the biological tissue. The characteristics of the bioelectrical signals place further constraints on the design and application of biomedical instrumentation systems. Novel closed-loop design strategies must be considered for the instrumentation systems used in medical facilities and research centers. In the design of a biomedical system, the following factors must be considered:

(1) *Energy source*: The human or animal body can provide an original source for the DBS system. Many biomedical systems allow human health to be monitored by an implanted device. Different body parts produce different signals. Bioelectric signals are produced by muscles and neuron cells, with the cell potential providing the electricity source.

(2) *Sensors*: A sensor is used to convert the physical condition or property into an electrical signal with specific characteristics. Factors including noise and electrode impedance can influence the choice of analog front-end architecture. For example, the electrode impedance is not dependent on the electrical properties of the materials, but on chemical reactions at the interface between the electrode and the electrolyte [18]. In general, commercial electrodes for bioelectrical recording have an impedance of 1 kHz, a range from 10 to 1000 kΩ, and capacitance values of between 100 and 350 pF. Hence, the amplifier must be properly designed to accommodate such large capacitive impedance. In real applications, the electrode impedance includes thermal noise [19], whose background fluctuations generate 5–10 μV over 10 kHz.

(3) *Signal acquisition and processing systems*: Converting the physical condition into an electrical signal can assist the user of the closed-loop DBS system. Signal acquisition mainly involves the use of function blocks such as amplifiers, A/D converters, D/A converters, wireless circuits, and digital control circuits.

After the biosignals have been converted into a digital form, the data must be passed through a closed-loop processing algorithm. Digital recording must be tightly controlled as it affects both the signal traces produced and the performance of the algorithm. Effective control of these parameters sets the system state under which the processing algorithm is characterized to the expected condition. For example, different kinds of bioelectrical sensing systems require different sampling rates, electrodes, and closed-loop control policies. All of them must be designed to a specification. In real clinical applications, motion artifacts from the patient may be

produced while recording the electrical signal, and such artifacts can significantly affect algorithm performance.

(4) *System Latency*: The operation of closed-loop system control can be divided into three steps: data acquisition, processing, and event detection [16]. Each step can be controlled by a digital core or by independent programs that communicate with each other via defined protocols. Each step introduces time and delay latency. If this latency is too long or unpredictable, it may interfere with the real-time closed-loop stimulation.

The time period for data acquisition in a sample block is controlled by the system protocol. The exact latency depends on the type of A/D converter used. As the data stored in the buffers is processed by the algorithms, the processing latency depends on the complexity of processing and the clock speed. The detection latency is determined by a number of factors that successively trigger the stimulator. Therefore, assessment of the timing characteristics is a critical issue in the development of closed-loop systems. Exact latency can be achieved with careful implementation.

(5) *Monitor system*: The monitor system is the bridge between the biomedical system and the host PC. Results can be displayed on a user-friendly graphical user interface. The monitor system can be numerical or graphical, or show features characteristic of the biomedical activity being investigated.

10.3 General Closed-Loop Deep Brain Stimulation and Recording System Design

The efficacy of DBS in treating neurological disorders such as movement disorders, pain, epilepsy, and psychiatric disorders has been demonstrated, and a closed-loop control policy can further improve these treatments by precisely monitoring the neurotransmitters. A typical closed-loop DBS system can be characterized as follows: (1) an adaptive DBS system can be used to measure and analyze a chosen variable reflecting ongoing pathological changes in the patient's clinical condition to derive new stimulation settings [20]. (2) Key elements that can be added to the generalized bi-directional brain-machine interface to facilitate research on chronic conditions include multichannel LFP amplification, accelerometers, spectral analysis, and wireless telemetry for data uploads [21]. (3) Closed-loop DBS systems have further developments in the charge transfer mechanisms at the electrode and tissue interface. This can be used to investigate the symptoms of neurological disorders and any side effects that may occur. Transgenic animals may be used in the testing of systems that improve the energy efficiency of the stimulation [22, 23]. (4) Closed-loop stimulation methods can dissociate changes in BG discharge rates and patterns, providing insights into the pathophysiology of PD [16]. This will have a significantly greater effect on akinesia. (5) While modulating neural activity is an effective treatment for neurological diseases, systems have been demonstrated that can identify a biomarker and the transfer functions of

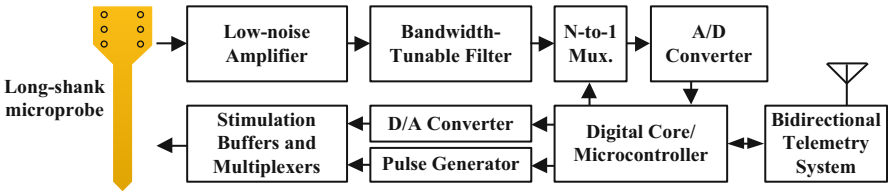


Fig. 10.1 System architecture of a real-time programmable closed-loop stimulation and recording platform comprising a neural recording system, a stimulation system, and a digital core

different stimulation amplitudes in a chronic animal mode [24]. (6) Signal unit recording can be done using a variety of brainwave recording techniques, and these parameters can be used for closed-loop activation of spinal circuitry below the level of injury [25]. A typical model of a closed-loop DBS system is illustrated in Fig. 10.1.

The DBS system mainly comprises a monitor and signal acquisition system, a microprocessor, a stimulator, and peripheral interconnection components. All the subsystems must be properly calibrated. Microprocessors form the kernel of the DBS, and their incorporation into biomedical systems has added computing power and greater control capability. The signal processor is used to acquire data, control the transducer, and provide closed-loop DSP. The performance of the signal processing algorithm is assessed prior to hardware implementation, and an input biosignal is passed through the algorithm. It must be demonstrated through simulation that the signal processing algorithm is rigorous, accurate, and produces reproducible data [23, 26]. The wider availability of microcontrollers has provided functional sub-circuits that offer both manual programming and automatic digital control, which can help reduce the complexity of the circuit interface. Compared with conventional biomedical electronic systems using DSP, the real-time closed-loop electronic system offers programmable recording and stimulation, a low-complexity and short latency DSP, and a closed-loop strategy for measuring feedback within the stimulated nuclei. In biomedical applications where low-power operation is a major concern, energy efficiency when performing a specific task becomes a significant consideration. These issues are discussed in Sect. 10.3.1.

10.3.1 Neural Recording System

The deep brain signals measured by the sensors are amplified to levels suitable for signal processing. The electrical signals are on the order of microvolts to millivolts, for example, 10–300 μV for LFPs, 1 μV for evoked potentials, and 10–20 μV for EEGs [27]. Two different methods are used for brainwave recording. One is called monopolar recording, in which a channel potential is compared to a reference electrode placed at a distant location. Monopolar recording involves amplitudes up to 50 μV, which are more sensitive to global neural activity and motion artifacts.

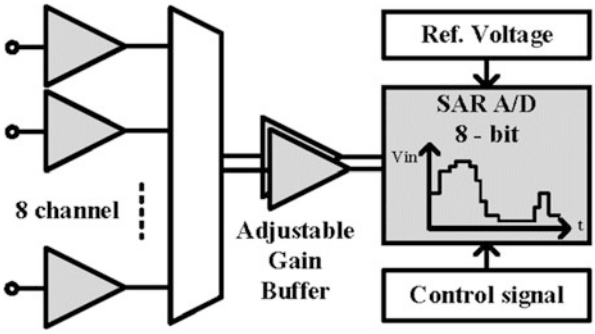


Fig. 10.2 The architecture of the recording system

The other method is called bipolar recording, in which a channel is compared to any one of the other channels, at amplitudes of approximately 5–20 μV, depending on the electrode distance and location. This method is less sensitive to noise and motion artifacts and more sensitive to localized neural activity.

As already noted, in a DBS application, the neural activities are captured by the probes or electrodes, and the bioelectrical signals generated can be fed into a microprocessor. Generally, amplifiers must be used without distortion and suppress system or environmental noise. Figure 10.2 shows a schematic of a recording system, which comprises a low noise amplifier and successive approximation (SAR) A/D conversion. The amplifier has an analog output or is integrated with the SAR A/D unit.

10.3.1.1 Amplifier

In principle, biosignal amplifiers are differential amplifiers which can measure a potential difference and contain high-impedance input circuits with a high common mode rejection ratio (CMRR) of between 60 and 110 dB. High CMRR can minimize noise from the power line and AC inductive power link. Ideally, the amplifier can amplify the difference between the brain signals from two input electrodes and eliminate the signal components that are common to both signals. When electronics are applied to the electrodes before the preamplifier or filtering, the CMRR of the amplifier may be reduced. For example, neural spikes contain high-frequency information (e.g., 300–6 kHz) and have a high-pass cutoff of around 6 kHz [28]. Analog filtering allows these high-frequency signals to pass through unattenuated. The anti-aliasing filter includes a low-pass cutoff, and the sampling rates must be high enough to sense high-frequency signals.

Amplifiers designed for deep brain signal recording must meet specific requirements. First, a high sampling rate over 20 kHz is required by the A/D converter, so an anti-aliasing filter must be incorporated. These filters typically remove high-frequency signals (e.g., higher than 500 Hz), in order to retain the LFPs waveform. Second, the input impedances must be higher than the probe impedance.

A typical neural probe impedance is in the range of 1–1000 kΩ at 1 kHz. The input impedance should be at least 100 times larger than that of the probe. Pre-amplifiers with high impedance are therefore commonly used at the buffering stage in the differential amplifier [29, 30].

10.3.1.2 A/D Converter

Many modern biomedical recording systems provide multichannels with high sampling rates, but the amplitude resolution and dynamic range of the A/D converter has a significant effect on the frequency range of the electrical signals recorded. In a deep brain recording system, an SAR A/D is commonly used to test the accuracy of the substrate resistance network, because it is widely applied to bio-chips and complete analysis methods of the SAR A/D. The advantages of this architecture are a low latency-time, high accuracy, and low-power consumption, with maximum sample rates of 2–5 MHz.

After neural signals are amplified, the digitized signal from each amplifier cycles many times per second, and this is called the sampling rate. The sampling frequency must satisfy the Nyquist criterion and must be at least 2 times larger than the highest frequency occurring in the biosignal. If the system does not satisfy the Nyquist criterion, the information contained in the sequence of samples will be distorted by aliasing [31]. In practice, the sampling rate may be several times higher than the recording bandwidth, since an anti-aliasing filter cannot filter out the higher frequency signals.

As described above, the brain signal is recorded using a multichannel differential amplifier. The signal input from different amplifiers and the entire analog signal must be digitized simultaneously. However, the digitization system contains a single A/D converter and the samples are acquired at different times. Several methods are available to address this problem: (1) a sample-and-hold method can hold the analog signal until the digitizer is ready to read the data or (2) a high sampling rate can reduce the sample time between each channel. A more detailed discussion of sampling and digitizing theory can be found in this reference [32].

10.3.2 Neural Stimulation System

The stimulation system consists of a responsive stimulator, electrode leads, and a function of the programmable parameters [33]. The stimulator is capable of communication with external components such as baseband circuits, a microcontroller, or wireless telemetry. The operator can use the bi-directional telemetry system to rewrite the digital core firmware and adjust the stimulator setting. The sensing data transmitter communicates with the central data management system where data can be stored and reviewed by physicians.

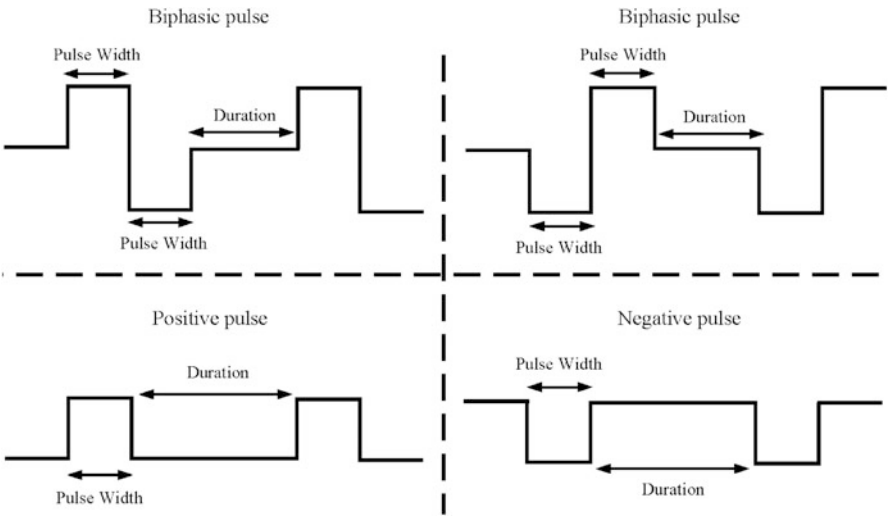


Fig. 10.3 Types of stimulation pulse shape, width, and duration

In principle, the closed-loop stimulator is a multi-functional stimulation device, which performs the following functions: (1) responsive stimulation, (2) digital to analog (D/A) decoding, and (3) battery voltage monitoring. Closed-loop responsive stimulation is controlled by a detection system in the digital core. As shown in Fig. 10.3, a range of stimulation parameters can be adjusted, including the stimulus voltage, stimulation frequency, and the pulse width and duration [34]. As discussed above, stimulation can be monopolar or bipolar. Monopolar stimulation requires the same polarity, while in bipolar stimulation the input is cathodic and the reference potential is anodic.

The D/A decoder is used to classify the digital patterns. The stimulation parameters can identify the pattern across the electrographic events and detection algorithm. After the stimulation parameter is set, the stimulator transmits the stimulation current to the tissue via the neural probe. A closed-loop stimulation control system can automatically evaluate the neural disorder and suppress neuronal synchrony. In addition, closed-loop stimulation incorporating battery voltage monitoring offers significantly extended battery life and reduced power consumption.

10.3.3 Digital Integrated Circuits

The digital circuit is the core of the closed-loop DBS system, performing data acquisition, DSP, sensor interfacing, and event detection. The generic architecture of the digital module is presented in Fig. 10.4.

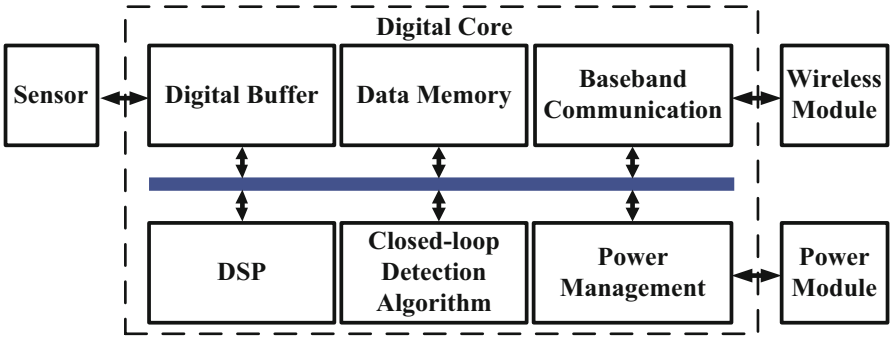


Fig. 10.4 A general architecture of digital core in closed-loop system

The DSP component of a closed-loop system performs two essential functions: feature extraction and control policy. Feature extraction is used to identify signal characteristics from the raw data and to convert the data into the mathematical domain. Data can be transferred to the frequency domain using Fourier transform or by assessment of power spectra, coherence, or cross-correlation coefficients [35, 36]. In closed-loop system design, it is essential that the signal processing is dynamic to allow optimal control policies to be applied to the digital circuits. Detection methods have therefore been adapted to fit the low-complexity constraints on the DSP. It is also important to take account of the system control policy when considering the closed-loop function for different depths of the brain. These functions include turning the stimulator on and off, feedback control of specific features, evaluation of the detection latency, and switching between the various subsystem interfaces.

The digital buffer is the bridge between the analog front-end and the digital circuits and can be used to acquire the digital signals from the subject. Different kinds of bioelectrical signal require different sampling rates. However, when the buffer cannot be merged with the digital module, tunable bus encoder technology can be used to reduce the biosignal activity [37].The digital buffer requires clocks, because different levels and functions of the system require different clock frequencies in the digital domain. As each subsystem can decrease the frequency range for different algorithms or baseband circuits, system power consumption and battery life can be substantially affected. For example, a sleep mode can be used to reduce the system power consumption. This mode is controlled by a low clock manager, and the controller can shut off the system main clock. Synchronous circuit design is very critical in a system which includes two different clock domains. Hence, the design of a state machine is a critical issue.

10.4 Closed-Loop Control Policy and Digital Signal Processing

Biomedical signals carry information on a biological system. Biomedical measurement systems must be secure and perform as intended to allow medical personnel to make precise diagnoses and choose the correct treatments. To satisfy these requirements, the digital core must meet numerous standards and regulations for each biomedical signal processing method used in the system. However, precise detection of the target signal and noise components remains a critical issue.

Most closed-loop stimulation systems use components of brain signals that are clearly characterized as biosignal features. They are detected using data processing algorithms that continuously monitor the brain activity. A data processing algorithm analyzes each incoming sample block from each biosignal to identify these features. The detection system is capable of comparison and specific detection. In general, two detection methods, linear and nonlinear, are used by the detection algorithm. These processing methods can be applied independently or to integrated circuits, and electrical stimulation systems can also be configured for detection. After the signal is processed, feedback is required to improve the detection of the probes at different depths in the brain. As the signals may include ECoG from the cortex or from the subthalamic nucleus in the deep brain, they clearly supply different amounts of information, based on the frequency domain [38]. The detection system must be able to adapt its parameters to improve the accuracy of the closed-loop stimulation system. Possible improvements in signal processing fall into the following categories: (1) extraction of information in more useful forms, (2) predictive data to anticipate the information from a specific signal, (3) data compression, and (4) filtering out of nonessential information, such as power-line noise or motion artifacts. Thus, it is common for a DBS system to incorporate features such as digital filters, linear and nonlinear combinations, or other modifications.

10.4.1 Digital Filtering

Digital filters are central to all signal processing systems. Each sample of the digitized signal is passed through a specific type of filter. Digital filters are used to filter out unwanted noise or artifacts from the biosignal to enhance the quality of the signal and prepare it for closed-loop detection.

If input data x[n] enter the filter sequentially, the output data y[n] is the weighted sum of the current and past values, and is given by:

$$y[n] = \sum_{i=0}^{M} a_i x[n-i] - \sum_{i=1}^{N} b_i y[n-i],$$

where x[n−i] is the past input data, y[n−i] is the past output data, the a_i and b_i parameters are the weights, and M and N are the numerical data lengths. The filtered output data have the same number of samples, in which the first sample of y[n] corresponds to the first sample of x[n]. Filters that depend on current and past data are known as causal filters. In real-time applications, causal filtering is necessary because future input and output data are not yet available. A simple example is the computation of a moving average. When computing a moving average, the signal passes through the filter, and the filter data is equal to the sum of the past input samples (N) each weighted by 1/N. This filter attenuates the high frequencies and preserves the low frequencies. The guidelines for developing closed-loop control demand low complexity and short latency. The sums of the past N consecutive samples have to be considered in the optimal range.

10.4.2 Time Domain Signal Processing Techniques

Linear signal processing can incorporate a number of different procedures. These can be categorized as (1) block processing, (2) peak detection and integration, and (3) wave detection. In a closed-loop system, it is desirable for the signal processing to occur in real time. Before being fed into the processing algorithm, the incoming raw data samples are segmented consecutively, either with or without overlapping. If the signal features are computed more frequently than is necessary, the system will consume additional computational time and power. In efficient real-time implementations, therefore, the data window size and overlap of the blocks should closely reflect the processing algorithm, detection latency, and available processing power.

Peak detection and integration are the most straightforward and simple methods for achieving this. Peak detection determines the maximum or minimum value of the data in a window and uses this value as the feature. Features can be averaged or integrated to provide more detailed information on the time domain. These methods can also be applied to the detection of transient spectrum peaks in the frequency domain.

Wave detection is calculated as the sum of the amplitude changes within a time window. This method is mainly used to measure the complexity of the fractal dimensions of a deep brain signal. The wave difference or the ratio of the average change between a short and long window can be used to determine whether the signal complexity is increasing or decreasing.

10.4.3 Frequency Domain Signal Processing Techniques

Deep brain signals have continuous amplitude and frequency modulated oscillations. A number of researchers are tracking these changes in the frequency domain. Much of frequency domain theory is based on Fourier analysis, which transforms time domain data into a frequency domain representation. Depending on the specific constraints or objectives, techniques such as fast Fourier transform (FFT) or autoregressive (AR) modeling are used. An FFT spectrum tracks the brain signal at a corresponding frequency, and the power spectrum of the FFT can be obtained by squaring the magnitude. This is an efficient implementation method for brain signals in a closed-loop system. AR modeling can use higher spectral resolution for signal window sizes shorter than those used in FFT. The development of small window sizes is necessary for closed-loop systems, because long latencies significantly affect real-time operations.

10.5 A Design Case: A Real-Time Closed-Loop Neurostimulation System Based on Neural Phase Synchrony Detection

10.5.1 Introduction to Closed-Loop Neurostimulation System

Parkinson's disease (PD) is one of the most prevalent diseases in people aged 50–60 years [39]. The introduction of levodopa (L-dopa) in the late 1960s caused a sharp decline in the surgical treatment of PD. Oral administration of L-dopa, which transforms into dopamine in the basal ganglia (BG), is a widely adapted chemical therapy for PD [40, 41]. However, long-term use of L-dopa is associated with motor fluctuations and dyskinesia [42], and DBS of BG nuclei is increasingly considered a highly effective and adjunctive therapy for PD symptoms [43–45], such as tremor and dystonia; moreover, it limits drug-induced side effects.

Numerous studies have quantified the energy in neural activities, which is measured by calculating the power density in a particular bandwidth of the brain because neural synchrony exhibits large amplitude and recurrence variability. Furthermore, increasing evidence indicates that deep electrical stimulation of brain structures suppresses neuronal synchrony at both BG [14]. Suppressed neuronal synchrony in PD usually involves open-loop deep electrical stimulation of brain structures such as LFPs [46]. Open-loop deep electrical stimulation delivers preprogrammed electric signal patterns, but the effective feedback loop for maintaining neurotransmitters cannot be chosen. By contrast, closed-loop stimulation strategy has a considerably stronger effect with preserving battery life, decreased neural synchronization, and an ideal control policy for feedback within the stimulated nuclei. For instance, neurological disorders are ameliorated when stimulation is based on electrical signal feedback and matched to the

frequency of the abnormal synchronization [16]. In addition, characteristic changes in the power spectra of LFPs have been identified after DBS of the BG, and an adaptive model based on recursive identification by the brain has been developed [17]. However, the power spectra do not consider the dynamic phase synchronous properties exhibited by the additional information in closed-loop strategies. The phase synchronous in the specific frequency band is a marker of neurological disorders in PD such as the beta band synchronization of BG is associated with the motor symptoms, such as the hypokinetic symptoms [47] and associated symptoms were reduced by different therapies [48]. Hence, the phase synchronous of neurophysiologic signals is used to estimate the neurological disorders at beta frequency oscillations. With this framework, this study is to develop a programmable closed-loop stimulation control system that automatically evaluates the neural phase synchrony to achieve reduction of motor symptoms.

Synchronous neural detection hardware that precisely controls intracortical microstimulation [49], closed-loop spike detection algorithms for triggering electrical stimulation, [50], and electroencephalograph (EEG) seizure detection [51, 52] have been implemented. Advanced algorithms include complexity analysis procedures, such as approximate entropy (ApEn) calculation and ameliorate neurological disorders [53]. In our previous studies, we developed a multichannel open-loop stimulation system-on-chip (SoC) that is based on an open-source 8051 microcontroller for real-time data collection [54, 55]. Because no commercial embedded systems are available for neural phase synchrony stimulation and recording, we developed a closed-loop DSP platform using flexible FFT and a fully programmable stimulus control strategy, which considers parameters such as stimulation amplitude, frequency, and pulse width. This platform is flexible and adaptable to the electronic algorithm design, which is substantially advantageous in the first step of the investigation.

In the present study, an RISC processor [56] is the core of the closed-loop phase detection algorithm; it provides a flexible architecture and a uniform length instruction set that affords system implementation at various processing performance levels. At the system architecture level, a combination of programmable and digital circuits provides implementation feasibility to different algorithms. Hence, the major closed-loop phase detection algorithm and signal processing are set by the user, thus precisely controlling the stimulator. An analysis of the performance requirements in neural networks shows that the proposed microprocessor is adequate for signal processing. At the microarchitecture level, the processor performs feature extraction using diverse FFT resolutions for obtaining neural information. Additionally, a low-complexity algorithm for addition and multiplication was proposed for variable-FFT-length (16–1024 points) implementations that involve a tradeoff between memory bandwidth and run time. Using the proposed algorithm, continuous LFP signals were collected from freely moving PD rats. We focus on the phase interaction of the signal amplitude to achieve highly dynamic modulation of neural activity by using the phase of low-frequency rhythms. Hence, the proposed processor provides a platform for combined statistical testing in a closed-loop microstimulation, which enables onset pattern detection and provides a precise stimulus for adequate treatment of PD symptoms.

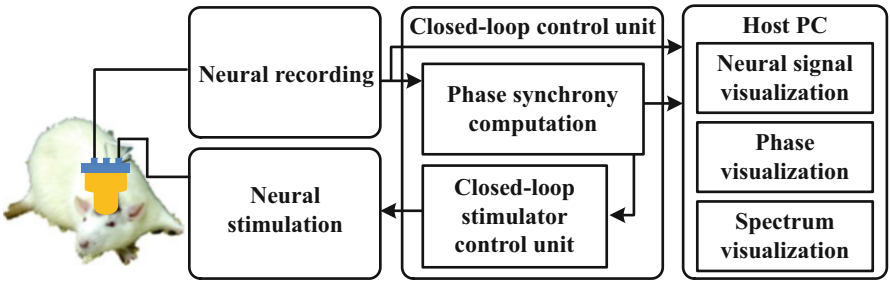


Fig. 10.5 Functional flow diagram of the closed-loop phase detection system. The digital signal processor is responsible for FFT, the control policy, and phase computation. In addition, the host PC can be used to rewrite the digital signal processor firmware, stimulation threshold, and to create a GUI for monitoring and storing data

10.5.2 System Architecture

The closed-loop phase detection system mainly consists of a recording analog front-end (AFE), a RISC processor, a stimulator, and a peripheral component interconnects. All external devices and algorithms are interconnected using a shared 32-bit data Wishbone bus, and the DSP, AFE, and analog-to-digital converter (ADC) interfaces can be connected in parallel with the crossbar switch. In addition, the host PC can rewrite processor firmware and monitor neural signal. The complete platform for closed-loop phase detection and the functional flow diagram is illustrated in Fig. 10.5.

10.5.2.1 Recording and Functional Electrical Stimulation

Adult (2-month-old) male Sprague Dawley rats weighing 250–350 g were used in this experiment. All animal handling, surgical, and behavior testing procedures were carried out in accordance with the guidelines on animal ethics. In each experiment, the rats were first anesthetized with chloral hydrate (400 g/kg) and urethane (0.5 g/kg). Microelectrodes were implanted into layer V of the M1 primary motor cortical region located using a stereotaxic apparatus with an average impedance of 50 kΩ at 130 Hz, after which the variable waveforms were recorded. Experimental experience has shown that a critical constraint in investigating neurophysiologic signals is that simultaneous recording of brain signals during DBS induces electrical artifacts several orders of amplitude larger than the brain waves. To overcome this constraint, investigators commonly adopt two distinct approaches. One approach involves recording the brain waves at the BG projection sites, where stimulation artifacts are less intense. The second approach, which we adopted in this study, involves avoiding stimulation artifacts by immediately recording brain waves after DBS. A 2-min M1 layer V time series was used for signal processing and statistical analysis on the host PC.

Data collection and the electrical stimulation architecture were based on our previous study. Deep brain signals were band-pass filtered between 0.1 and 700 Hz and amplified 3000-fold. A voltage-controlled stimulation (VCS) for a grounded load transmits the stimulation current to the tissues via the neural electrode. Through a programmable VCS interface implemented in the FPGA evaluation board, the user selects the RISC processor, sets the stimulation parameters, and stimulates the neurons.

10.5.2.2 Phase Analysis

Excessive synchronized oscillation in the beta frequency is one of the most common PD biomarker signals [14, 46] and a characteristic feature of the neural network activity in LFPs. However, the neuronal oscillations are quite inconsistent, and frequency spectra do not contain phase variant properties. On the basis of this knowledge, time-series fragmentation techniques are employed for analyzing intermittent synchronized oscillations to investigate the dynamic phase of synchronized signals, and the phase space of the brain signals is used to estimate the neural rhythms. The localized phase synchronizations are analyzed separately using frequency correlation. Each offline signal processing procedure is detailed for PC. All PC-based analyses were performed using MATLAB (The MathWorks, Natick, MA, USA). Before spectral analysis, raw data were band-pass filtered between 10 and 50 Hz to remove power-line noise and low-frequency oscillations. The time-varying power of neural changes in the motor state was estimated using short-time Fourier transform. Power changes were estimated using FFT with windows of 1024 samples, a Hanning window with a width of 0.5 s, and a 50 % overlap, until all signals were analyzed. In this time-varying spectral analysis, the high beta frequencies (20–35 Hz) of different relative power information were measured.

We detected localized phase synchronizations with respect to both time and frequency. Phase synchronizations were analyzed separately using a coherence technique. This technique involves obtaining the cross-spectrum of two signals by using the power spectral density. First, FFT is applied to both auto-power (P_{xx}) and cross-power spectra (P_{xy}).

$$P_{xx}(w) = E\left[|X(w)|^2\right], \tag{10.1}$$

$$P_{xy}(w) = E\left[X(w)Y(w)^*\right], \tag{10.2}$$

where $X(w)$ and $Y(w)$ are obtained using FFT of the time domain signals x(t) and y(t) (Fig. 10.6a). Next, phase relations are computed as follows:

$$\text{Phase}(w) = -\arctan\left(T_{xy}(w)\right) \tag{10.3}$$

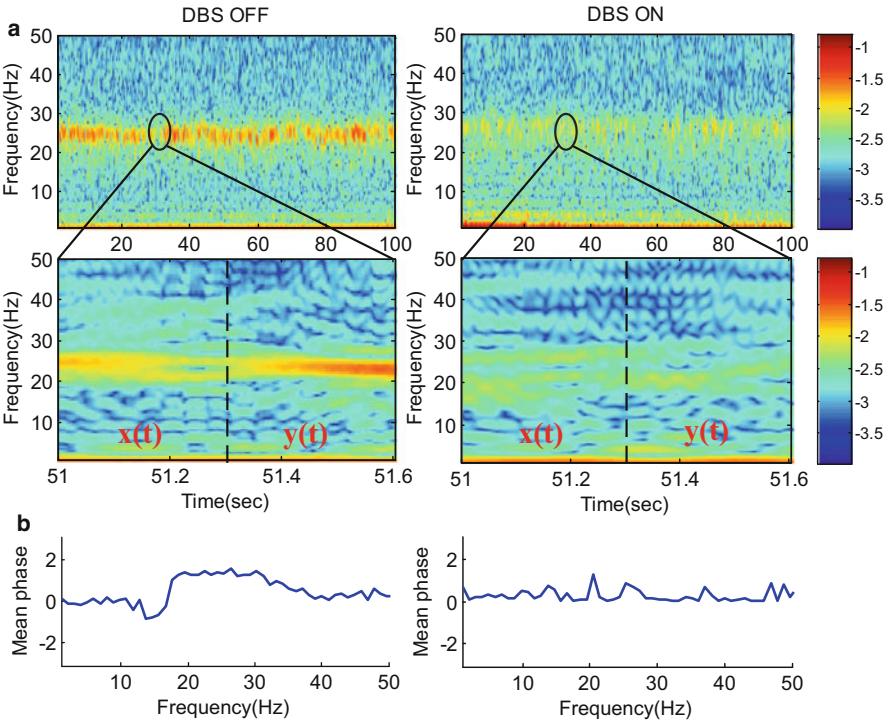


Fig. 10.6 LFP betas in the DBS ON and OFF states exhibit different characteristics. (**a**) Beta is observed in the 13–35 Hz range, where the power range is reduced and DBS ON and OFF is intermittently synchronous. The power ranges shown are plotted logarithmically. Time domain signals x(t) and y(t) are used to calculate the cross-spectrum and phase synchrony. (**b**) The mean phase (n = 37) is calculated for DBS ON and OFF, and the mean phase in DBS ON is more stable than that in DBS OFF

where $T_{xy}(w)$ is a transfer function based on the cross-spectrum of a signal pair; it is defined as

$$T_{xy}(w) = \frac{E\left[X(w)Y(w)^*\right]}{E\left[|X(w)|^2\right]} \tag{10.4}$$

All mean phases (n = 37) in the DBS ON and OFF states are presented in Fig. 10.6b. Phase relations refer to the periodicity labeling of the neural activity signals at a point between the coordinates (−3.141 to 3.141). On analyzing the beta frequency power time series, the phase series was found not to be a constant in the intermittent synchrony of power over time. The mean phase of DBS ON is more stable than that of DBS OFF. Therefore, adding phase time-series detection capability to a VCS system facilitates addressing neurobiological concerns.

10.5.2.3 Closed-Loop Stimulator Control Unit

As mentioned, neural phase dynamics can be calculated using the Fourier theory, and the phase equation of (10.1) enables the processor to activate the stimulator by using a precise threshold. The threshold condition of phase synchronization (PS) is defined as

$$\mathrm{PS} = \begin{cases} 1, & -\beta \times \pi \geq Phase(w) \geq \beta \times \pi \\ 0, & otherwise \end{cases} \tag{10.5}$$

where β is a constant determined using the statistical analysis presented in the subsequent paragraph. $\mathrm{PS} = 1$ activates the stimulator. In this study, precisely control of stimulator in short latency is a critical factor in maintaining the closed-loop system. An operational flowchart is depicted in Fig. 10.7.

A preliminary statistical analysis was performed to evaluate the start threshold β. Before determining the stimulation threshold, the continuous phase parameters were analyzed using paired t tests to evaluate the mean difference between stimulator ON and OFF states that maps β to an optimal stimulator start condition. Such analysis was performed on neural signals exhibiting PD symptoms. First, the phase characteristics of stimulator ON and OFF were independently tested for normality using the Shapiro–Wilk test. When the statistic is significant, the hypothesis that the distribution is normal is rejected. Second, another nonparametric statistical analysis, the paired t test, was performed to evaluate the differences in each pair. Significant p values were obtained in all cases.

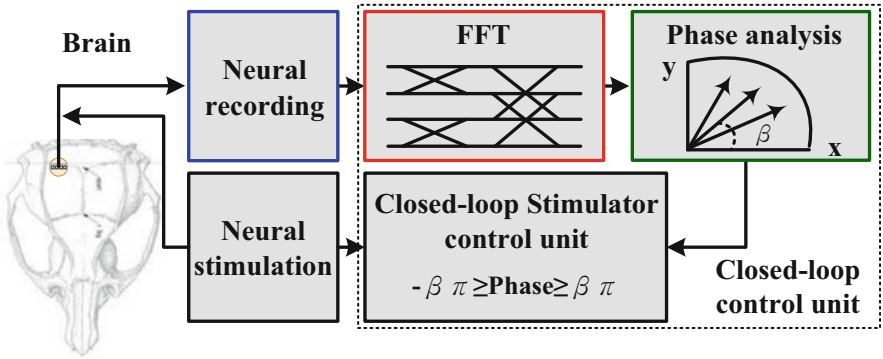


Fig. 10.7 System architecture of the closed-loop phase synchrony detection, consisting of a neural recording, a neural stimulation, and a closed-loop control unit. Closed-loop control unit covert the neural signal into phase domain and trigger stimulator when that phase synchrony exceeds a threshold

10.5.3 Real-Time Digital Processing Platform

Although our previous studies have successfully implemented 8051 microcontrollers in freely moving rats [54, 55], the architecture requires numerous instructions for a single operation. A high number of instructions can negatively affect the memory size, processing bandwidth, and real-time performance. Therefore, this study employed a 32-bit RISC processor, OpenRISC 1200 [56], for arithmetic operations, data storage, system control, and neural variant property evaluation. The 32-bit processor provides a computing speed of 2.1 DMIPS per MHz under the dhrystone benchmark and DSP MAC 32×32 operations per MHz. By contrast, the 8051 microcontroller in our previous studies computed at 1.67 DMIPS per MHz. In addition, virtual memory support, a five-stage pipeline, and basic capabilities are included in OpenRISC 1200. Processor efficiency and flexibility in specific tasks are crucial for precise and complex biomedical algorithms that sense neural activity.

The neural phase analysis of the system is classified into two components: First, the vector mode of the coordinate rotation digital computer (CORDIC) algorithm [57] is proposed, fulfilling the arctangent function, because it requires shift and add operations for each vector rotation. Because of the 16-bit fixed point computation in the processor, the word length of CORDIC computations must be considered for numerical accuracy. To achieve a 10-bit resolution of the fraction part [48], the word length should be at least $(n+\log_2 n+2)$, and $n+1$ rotations must be performed. Second, the variable point radix-2^4 FFT is proposed [58]. In the proposed RISC processor, this architecture achieves the lowest computing complexity for mathematical operations. In addition, the radix-2^4 algorithm provides a variable FFT length ranging from 16 to 1024 points. To increase FFT computation efficiency, the twiddle factor is precomputed, stored in an array in the processor memory, and accessed through table lookup.

Real-time signal processing for the closed-loop detection of localized phase synchronizations is implemented using a digital processor. The proposed algorithm and control policy can be verified on the implantable device firmware by using a wireless module. The firmware is segmented and updated independently and allows a series of instructions into the program memory from the boot ROM. The processor is programmed to deliver information as follows: (1) retrieve data from the SRAM and input it into the asynchronous FIFO; (2) set the stimulation parameters and the threshold for specific brain regions; and (3) start the N-point radix-2^4 FFT and CORDIC for phase computation.

After the phase property is calculated using the proposed algorithm, a hybrid approach is used where Fourier-based segmentation with phase synchrony is applied to maintain the neural activity reflected in the deep brain. Subsequently, the phase threshold condition β, which starts the stimulator, is trained statistically. When phase synchrony occurs, the processor generates a stimulation pulse. The initial programming setting generates a pulse width of 60 μs at a frequency of 125 Hz and amplitude of 5 V and supplies a constant current of 100 μA. These

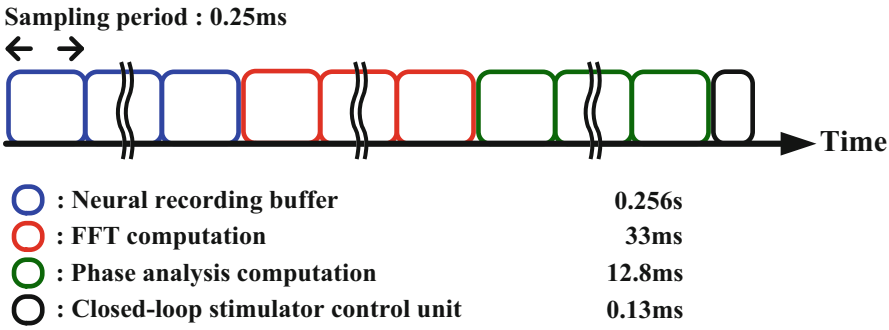


Fig. 10.8 Timing diagram of the phase detection firmware

parameters are effective in managing PD symptoms [59]; the user can flexibly modify the stimulation parameters by modifying the instruction set.

Detection latency is a critical factor in maintaining a real-time closed-loop system in which features of phase change are used to trigger stimulation. To evaluate the detection latency, an EEG dataset that generated output signals serving as AFE triggers was used. When one of the phase characteristics can trigger stimulation, it triggers the stimulator output and sets the identity of the detection unit, which is used in the feedback control scheme. The latency between a phase discriminator and a digital sample is then calculated using the same sample clock. A timing diagram of the experimental protocol is illustrated in Fig. 10.8. Closed-loop stimulation latencies were calculated using numerous individual subsystems, such as the digital sample, control policy, and phase detection algorithm. Using a 1024-point digital sample, the average run time for the asynchronous FIFO was 0.256 s. Furthermore, the phase detection latency calculated using the phase frequency and corresponding change in the Wishbone bus was determined for each firmware option. The difference in reaction latency for the RISC processor is largely attributable to the FFT and CORDIC; the mean latencies for the 1024-point FFT computation and CORDIC were 33 and 12.8 ms, respectively. Table 10.1 illustrates a comparison of the detection latency of the proposed system with those of recently published closed-loop systems. The shortest latency is reported in [53], where the sampling frequency is lower. Nevertheless, the 1024-point FFT and sampling frequency of our system is higher, and the neurological detect state is not the same. For the same neurological detect state, the shortest latency is in [60], where the test platform uses a software and hardware coworker but does not consider the sampling time.

10.5.4 Implementation Results

An ALTERA DE2-115 FPGA evaluation board, capable of reloading instructions and implanted with dedicated RISC processors, was used to verify the functionality and basic control policy. The processor instructions were stored in a 2 MB SRAM; a

Table 10.1 Comparison of the proposed system with recent systems

	[60]	[53]	[51]	[23]	This work
DSP operation	Empirically derived linear threshold on the amplitude	Programmable, 64-point FFT (radix-4), and on-line seizure detection	256-point FFT (radix-2)	LFP amplitude threshold for triggering stimulation	**Flexible N-point FFT (N = 16 through 1024) (Radix-2^4)**
Processor	PC	OR1200	Cortex-M3	Embedded processor (Spike 2 software)	**OR1200**
Latency	0.600 s	0.500–0.800 s	>0.500 s	>0.030–0.040 s[a] + x	**0.302 s (N = 1024) 0.550 s (N = 256, 5 MHz) 0.114 s (N = 64, 13.6 MHz)**
Sampling frequency	422 Hz	200 Hz	256 Hz	1 kHz	**4 kHz**
Operating frequency	N/A	13.6 MHz	7 kHz (0.5 V) 5 MHz (1 V)	N/A	**25 MHz**
Neurological state	LFP, EMG, ECoG	Epilepsy	Epilepsy	LFPs	**LFPs, EMG, sEEG**

[a]*x* without consider 0.400 s moving average filter, digital sample, and software and hardware interface delay time
N/A Not Available

waveform generator was used to store the EEG dataset and regenerate the brain signal for the digital processors. The EEG datasets were collected from our previous study. Figure 10.9 showed the top view of the recording system and a rat implanted with the neuron-probes. In addition, the EEG dataset was used to demonstrate the functionality of the system: data collection, phase characteristic identification, and compatibility with the closed-loop algorithm and the evaluation board.

After the system is initiated, the user can rewrite the firmware codes and start the RISC processors, which can access digital samples directly from the ADCs (waveform generators) and continuously perform phase detection. To facilitate validating the dynamic phase and the stimulator control policy states, the experimental results are streamed in real time to the host PC through a wireless module so that the user can monitor the algorithm operation.

A specific example is the closed-loop DBS for PD, which entails (1) capturing neural signals by controlling the ADC sampling clock; (2) statistically determining the stimulation threshold; and (3) evaluating the phase synchronization through DSP.

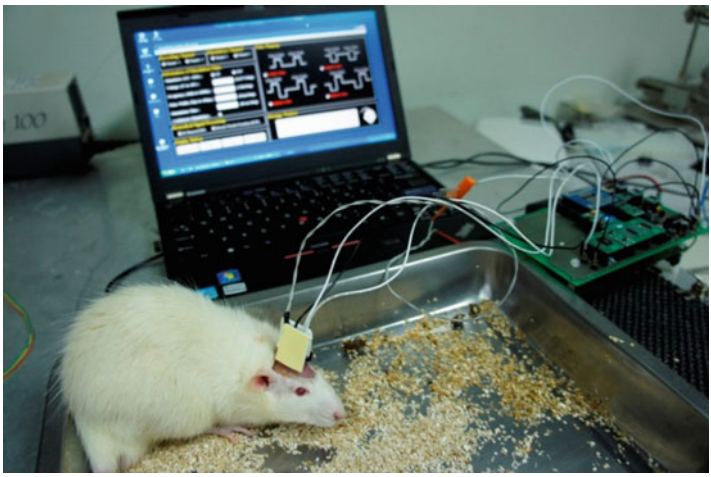


Fig. 10.9 Data collection and electrical stimulation

10.5.4.1 Data Collection

The digital samples are loaded into a waveform generator and output to a 12-bit ADC. The sampling frequency for the neural signal is 4 kHz, and system operation at 25 MHz is adequate for real-time signal processing. Subsequently, the processor has a buffer for retrieving 1024 digital samples through a parallel interface. FFT, CORDIC, and phase are computed for the next 1024 digital samples and phase synchronizations are subsequently detected. The experiment was completed in approximately 2 min. After using the timing diagram to design the firmware, the default instructions are written to the processor; handshakes with the host PC ensure that the data are passed correctly, subsequently; the user sets the stimulation parameters and controls the ADC.

10.5.4.2 Statistical Results

Statistical analyses were performed using the R 3.1.1 software (R Foundation for Statistical Computing, Vienna, Austria). A two-sided p value less than 0.05 was considered statistically significant. Data were expressed as the mean $\pm$ standard deviation. The continuous data of DBS ON and OFF were assessed using the paired t test. Forty data records for DBS ON and OFF, each measured 40 times, were included. Three data signals were excluded because of poor data recording and incomplete recording of brain signals.

All data ($n = 37$) are presented in Table 10.2. The mean phases of DBS ON and OFF show statistically significant (1.084 ± 0.748 vs. 1.600 ± 0.864; $p = 0.002$) that using the paired t test. Normality for DBS ON and OFF was statistically significant

Table 10.2 Effect of electrical stimulation on phase response

	ON	OFF	
Phase	1.084 ± 0.748	1.600 ± 0.864	$p = 0.002$
Shapiro–Wilk test: DBS ON: $p = 0.180$, DBS OFF: $p = 0.113$			

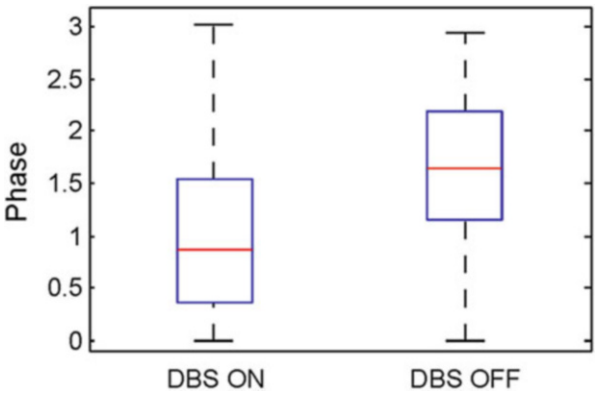


Fig. 10.10 Box–Whisker plot of phase suppression of the beta band. Suppression of phase synchrony of LFPs occurs between 20 and 35 Hz in the DBS ON and OFF states

($p = 0.180$, $p = 0.113$). Figure 10.10 showed the Box–Whisker plot of phase suppression. According to this statistical result, the stimulation threshold based on DBS OFF was sets 1.6°.

10.5.4.3 Phase Synchronization Detection in the EEG Dataset

For detecting phase suppression, the transfer function that measures the signal of synchronous change over short-time intervals is calculated. Every 0.256 s, a 1024-point FFT is performed for the input data with a 50 % overlap in time, followed by the execution of the CORDIC algorithm. Phase synchronization is determined using two 1024-sample cycles along with DSP; phase response detection takes approximately 0.302 s.

Neural recording of M1 layer V stimulation artifacts during electrical stimulation at 125 Hz and pulse width of 60 μs is presented in Fig. 10.11a. To evaluate the effect of electrical stimulation on phase detection, neural signals after DBS (Fig. 10.11b) are used as a dataset for DSP. Figure 10.11c, d shows a 0.302 s segment of the 1024-point FFT and phase irregular stimulation pattern; the upper panels depict power and phase detection events for triggering stimulation. These signals were recorded to determine whether the LFP power spectra and phase were consistent across each segment of the time series despite interventions such as DBS ON or OFF states. In each segment, the signal that differed indicates that the power and phase in the beta range triggered stimulation.

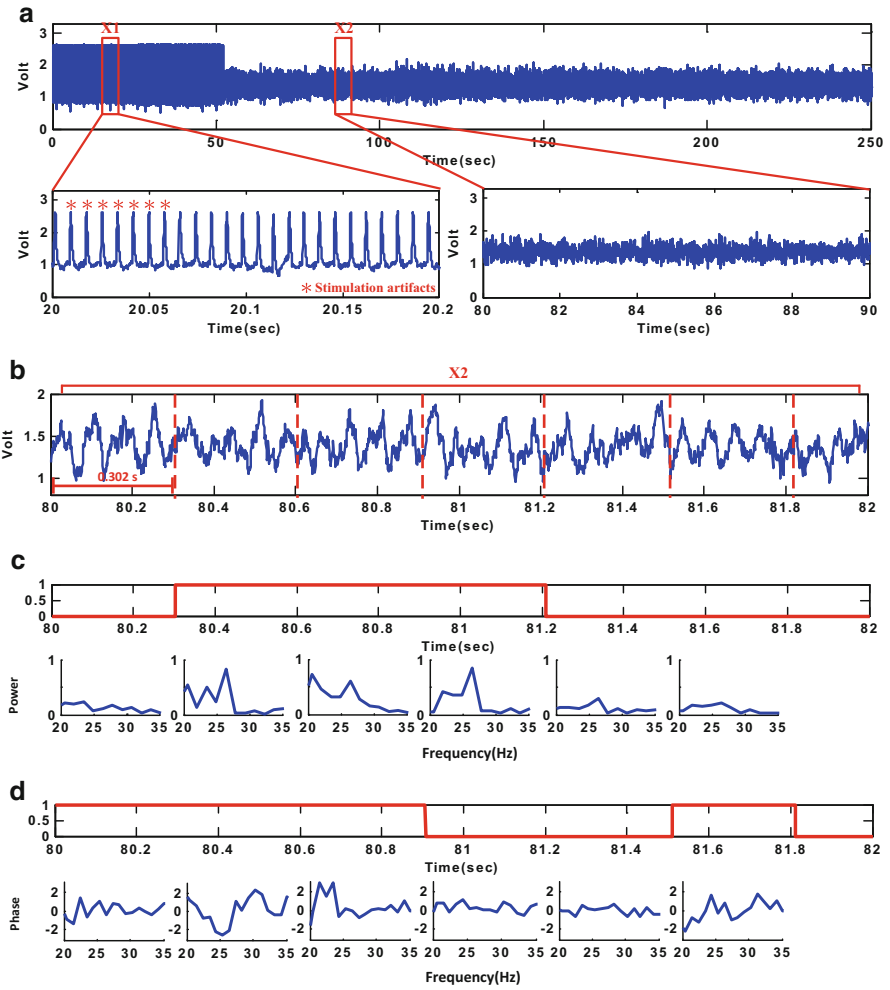


Fig. 10.11 Two-second series analysis in detecting the fine temporal structure of intermittent power spectrum and phase synchrony. (**a**) An example of a 250 s LFP neural signal S1 before showing stimulation artifacts and S2 after applying the stimulus paradigm. (**b**) Neural recording for 2 s after electrical stimulation. (**c**, **d**) 0.302 s trace of the stimulation pattern; 1024-point FFT and phase irregular stimulation pattern

10.5.4.4 Comparison with Other Studies

The foundation of the closed-loop system is related to choose the neural data as a biomarker. All techniques have its advantage, depending on the different algorithm and practical system constraints. Table 10.3 lists the results of the phase detection algorithm compared with monitor power spectrum when that power exceeds a stimulation threshold in the specific frequency band [38] and LFPs oscillations

Table 10.3 Summary of power spectrum, LFP AMPLITUDE, and phase synchrony detections

	[38]	[23]	This work
Stimulation times	371 ± 14	358 ± 31	302 ± 21

was used to control triggered stimulator via a defined threshold [23]. In addition, the EEG dataset ($n = 37$) was used to demonstrate the functionality of the closed-loop system and the total duration for all experiments was nearly 2 min.

To evaluate the closed-loop effect of electrical stimulation on power spectrum, LFPs oscillations and phase synchrony detection, the numbers of triggered stimulator are used as an indicator in this experiment. In an open-loop system, the total detection number is 397, without considering the stimulation threshold. And then closed-loop spectrum detection and LFPs oscillation detection need 371 and 358 times stimulation. Our proposed method requires 302 times stimulation. Hence, the selection of features for detection has important implications for power consumption.

As a result, the spectrum, LFPs oscillations, and phase synchrony detection numbers represent distinct phenomena (Fig. 10.12) and Fig. 10.12d showing that our proposed method in the beta frequency band has a lower percentage of triggered stimulator over time. These three methods were individually significant in the EEG dataset (LFPs amplitude, phase synchrony, and power spectrum; $p < 0.001$) and exhibit a positive relationship over time.

In this study, a real-time closed-loop neurostimulation system should be able to provide localized phase detection and precise control of electrical stimulation for in-vivo experiments; accordingly, such in-vivo experiments have to be considered for the next step of experimentation.

10.6 Conclusions

With the development of bioelectronics, DSP, and effective feedback loops for maintaining neurotransmitters, the closed-loop DBS system is playing a growing role in the treatment of certain medical conditions. Precise control of electrical stimulation requires an advanced signal processing algorithm that maintains therapeutic efficacy at optimal levels. Through detailed analysis of typical closed-loop DBS systems, this chapter has presented an overview of neural recording systems, stimulation systems, and digital integrated circuits. The specifications of each function have been presented.

A real-time closed-loop neurostimulation system based on a neural phase synchrony detection design has also been discussed. We developed a closed-loop digital signal processor platform using the radix-2^4 algorithm, providing a variable FFT length ranging from 16 to 1024 points with a short latency response (0.302 s) and offering fully programmable stimulation. This study focused on the use of transfer-function-based LFP processing for calculating phase synchrony, which

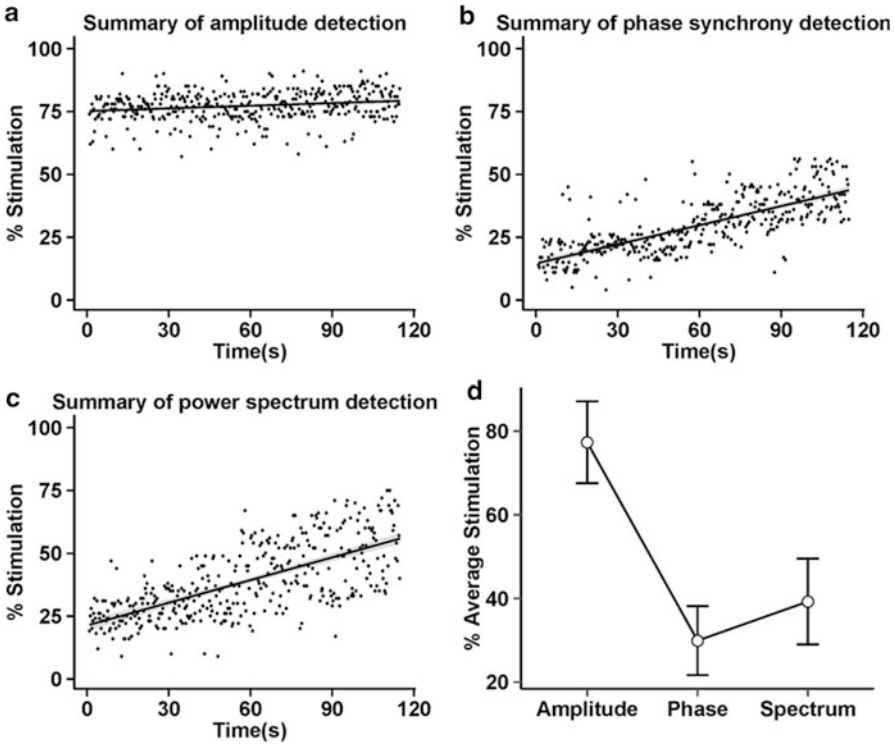


Fig. 10.12 The percentage of triggered stimulator over time (% per 0.302 s block). The solid line is result of linear regression. (**a**) The percentage of LFPs amplitude triggered stimulator. Multiple R-squared is 0.537, $p < 0.001$. (**b**) The percentage of phase synchrony triggered stimulator. Multiple R-squared is 0.482, $p < 0.001$. (**c**) The percentage of phase synchrony triggered stimulator. Multiple R-squared is 0.512, $p < 0.001$. (**d**) Mean and standard error of the percentage of triggered stimulator with different stimulation conditions

was then used to trigger the neural stimulator. A preliminary statistical analysis provided additional evidence that the feedback threshold parameters can be optimized and directly transferred to fit the threshold conditions. Advanced statistical methods can then be used to address neurobiological problems. Finally, the proposed method can be extended to other biomedical signal applications such as electromyography and surface EEG. Commercial platforms can be employed to conduct short-term experiments and functional verification, whereas ICs are more practical for specific and long-term applications.

References

1. Harrison RR (2007) A versatile integrated circuit for the acquisition of biopotentials. Custom integrated circuits conference, p 115–122
2. Kultas-Ilinsky K, Ilinsky IA (2001) Basal Ganglia and Thalamus in health and movement disorders. Springer, New York
3. Lozano AM, Gildenberg PL, Taske RR (2009) Textbook of stereotactic and functional neurosurgery. Springer, Berlin and Heidelberg
4. Perlin GE, Sodagar AM, Wise KD (2006) Neural recording front-end designs for fully implantable neuroscience applications and neural prosthetic microsystems. Engineering in Medicine and Biology Society, p 2982–2985
5. Lopez-Azcarate J, Tainta M, Rodriguez-Oroz MC, Valencia M, Gonzalez R, Guridi J et al (2010) Coupling between beta and high-frequency activity in the human subthalamic nucleus may be a pathophysiological mechanism in Parkinson's disease. J Neurosci 30:6667–6677
6. Schuurman PR, Bosch DA, Bossuyt PMM, Bonsel GJ, van Someren EJW, de Bie RMA et al (2000) A comparison of continuous thalamic stimulation and thalamotomy for suppression of severe tremor. N Engl J Med 342:461–468
7. Fisher R, Salanova V, Witt T, Worth R, Henry T, Gross R et al (2010) Electrical stimulation of the anterior nucleus of thalamus for treatment of refractory epilepsy. Epilepsia 51:899–908
8. Halpern CH, Samadani U, Litt B, Jaggi JL, Baltuch GH (2008) Deep brain stimulation for epilepsy. Neurotherapeutics 5:59–67
9. Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C et al (2005) Deep brain stimulation for treatment-resistant depression. Neuron 45:651–660
10. Schoenen J, Di Clemente L, Vandenheede M, Fumal A, De Pasqua V, Mouchamps M et al (2005) Hypothalamic stimulation in chronic cluster headache: a pilot study of efficacy and mode of action. Brain 128:940–947
11. Leone M, Franzini A, Bussone G (2001) Stereotactic stimulation of posterior hypothalamic gray matter in a patient with intractable cluster headache. N Engl J Med 345:1428–1429
12. Loukas C, Brown P (2012) A PC-based system for predicting movement from deep brain signals in Parkinson's disease. Comput Methods Programs Biomed 107:36–44
13. de Haas R, Struikmans R, van der Plasse G, van Kerkhof L, Brakkee JH, Kas MJ et al (2012) Wireless implantable micro-stimulation device for high frequency bilateral deep brain stimulation in freely moving mice. J Neurosci Methods 209:113–119
14. Meissner W, Leblois A, Hansel D, Bioulac B, Gross CE, Benazzouz A et al (2005) Subthalamic high frequency stimulation resets subthalamic firing and reduces abnormal oscillations. Brain 128:2372–2382
15. McCracken CB, Kiss ZHT (2014) Time and frequency-dependent modulation of local field potential synchronization by deep brain stimulation. PLoS One 9, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102576
16. Rosin B, Slovik M, Mitelman R, Rivlin-Etzion M, Haber SN, Israel Z et al (2011) Closed-loop deep brain stimulation is superior in ameliorating Parkinsonism. Neuron 72:370–384
17. Santaniello S, Fiengo G, Glielmo L, Grill WM (2011) Closed-loop control of deep brain stimulation: a simulation study. IEEE Trans Neural Syst Rehabil Eng 19:15–24
18. Jochum T, Denison T, Wolf P (2009) Integrated circuit amplifiers for multi-electrode intracortical recording. J Neural Eng 6, http://www.ncbi.nlm.nih.gov/pubmed/19139560
19. Gesteland RC, Howland B, Lettvin JY, Pitts WH (1959) Comments on microelectrodes. Proc IRE 47:1856–1862
20. Priori A, Foffani G, Rossi L (2005) Apparatus for treating neurological disorders by means of adaptive electro-stimulation retroacted by biopotentials. European Patent Office
21. Rouse AG, Stanslaski SR, Cong P, Jensen RM, Afshar P, Ullestad D, et al (2011) A chronic generalized bi-directional brain-machine interface. J Neural Eng 8, http://www.ncbi.nlm.nih.gov/pubmed/21543839

22. Nowak K, Mix E, Gimsa J, Strauss U, Sriperumbudur KK, Benecke R, et al (2011) Optimizing a rodent model of Parkinson's disease for exploring the effects and mechanisms of deep brain stimulation. Park Dis 2011, http://www.ncbi.nlm.nih.gov/pubmed/21603182
23. Little S, Pogosyan A, Neal S, Zavala B, Zrinzo L, Hariz M et al (2013) Adaptive deep brain stimulation in advanced Parkinson disease. Ann Neurol 74:449–457
24. Afshar P, Khambhati A, Stanslaski S, Carlson D, Jensen R, Linde D et al (2012) A translational platform for prototyping closed-loop neuromodulation systems. Front Neural Circ 6:117
25. Lobel DA, Lee KH (2014) Brain machine interface and limb reanimation technologies: restoring function after spinal cord injury through development of a bypass system. Mayo Clin Proc 89:708–714
26. Avestruz AT, Santa W, Carlson D, Jensen R, Stanslaski S, Helfenstine A et al (2008) A 5 uW/channel spectral analysis IC for chronic bidirectional brain-machine interfaces. IEEE J Solid State Circuits 43:3006–3024
27. Nunez P (1981) Electric fields of the brain: the neurophysics of EEG. Oxford University Press, New York
28. Zhang K, Ginzburg I, McNaughton BL, Sejnowski TJ (1998) Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J Neurophysiol 79:1017–1044
29. Burke MJ, Gleeson DT (2000) A micropower dry-electrode ECG preamplifier. IEEE Trans Biomed Eng 47:155–162
30. Spinelli EM, Martinez N, Mayosky MA, Pallas-Areny R (2004) A novel fully differential biopotential amplifier with DC suppression. IEEE Trans Biomed Eng 51:1444–1448
31. Oppenheim AV, Schafer RW (2009) Discrete-time signal processing. Prentice Hall
32. Proakis JG, Manolakis DG (1996) Digital signal processing: principles, algorithms, and applications, 3rd ed. Prentice-Hall
33. Psatta DM (1983) Control of chronic experimental focal epilepsy by feedback caudatum stimulations. Epilepsia 24:444–454
34. Hamani C, Diwan M, Isabella S, Lozano AM, Nobrega JN (2010) Effects of different stimulation parameters on the antidepressant-like response of medial prefrontal cortex deep brain stimulation in rats. J Psychiatr Res 44:683–687
35. McCracken CB, Grace AA (2009) Nucleus accumbens deep brain stimulation produces region-specific alterations in local field potential oscillations and evoked responses in vivo. J Neurosci 29:5354–5363
36. McIntyre CC, Chaturvedi A, Shamir RR, Lempka SF (2015) Engineering the next generation of clinical deep brain stimulation technology. Brain Stimul 8:21–26
37. Suresh DC, Agrawal B, Yang J, Najjar W (2005) A tunable bus encoder for off-chip data buses. Presented at the proceedings of the 2005 international symposium on Low power electronics and design
38. Priori A, Foffani G, Rossi L, Marceglia S (2013) Adaptive deep brain stimulation (aDBS) controlled by local field potential oscillations. Exp Neurol 245:77–86
39. Romrell J, Fernandez HH, Okun MS (2003) Rationale for current therapies in Parkinson's disease. Expert Opin Pharmacother 4:1747–1761
40. Deep-brain stimulation of the subthalamic nucleus or the pars interna of the globus pallidus in Parkinson's disease (2001). N Engl J Med 345:956–963, http://www.ncbi.nlm.nih.gov/pubmed/11575287
41. Kleiner-Fisman G, Fisman DN, Sime E, Saint-Cyr JA, Lozano AM, Lang AE (2003) Long-term follow up of bilateral deep brain stimulation of the subthalamic nucleus in patients with advanced Parkinson disease. J Neurosurg 99:489–495
42. Fishman PS (2008) Paradoxical aspects of Parkinsonian tremor. Mov Disord 23:168–173
43. Dostrovsky JO, Lozano AM (2002) Mechanisms of deep brain stimulation. Mov Disord 17 (Suppl 3):S63–S68
44. Bronte-Stewart H, Taira T, Valldeoriola F, Merello M, Marks WJ Jr, Albanese A et al (2011) Inclusion and exclusion criteria for DBS in dystonia. Mov Disord 26(Suppl 1):S5–S16

45. Eusebio A, Thevathasan W, Gaynor LD, Pogosyan A, Bye E, Foltynie T et al (2011) Deep brain stimulation can suppress pathological synchronisation in Parkinsonian patients. J Neurol Neurosurg Psychiatry 82:569–573
46. McCracken CB, Kiss ZH (2014) Time and frequency-dependent modulation of local field potential synchronization by deep brain stimulation. PLoS One 9, http://www.ncbi.nlm.nih.gov/pubmed/16123144
47. Park C, Worth RM, Rubchinsky LL (2010) Fine temporal structure of beta oscillations synchronization in subthalamic nucleus in Parkinson's disease. J Neurophysiol 103:2707–2716
48. Silberstein P, Pogosyan A, Kuhn AA, Hotton G, Tisch S, Kupsch A et al (2005) Cortico-cortical coupling in Parkinson's disease and its modulation by therapy. Brain 128:1277–1291
49. Venkatraman S, Elkabany K, Long JD, Yao Y, Carmena JM (2009) A system for neural recording and closed-loop intracortical microstimulation in awake rodents. IEEE Trans Biomed Eng 56:15–22
50. Angotzi GN, Boi F, Zordan S, Bonfanti A, Vato A (2014) A programmable closed-loop recording and stimulating wireless system for behaving small laboratory animals. Sci Rep 4: 15–22
51. Sridhara SR, DiRenzo M, Lingam S, Seok-Jun L, Blazquez R, Maxey J et al (2011) Microwatt embedded processor platform for medical system-on-chip applications. IEEE J Solid State Circuits 46:721–730
52. Verma N, Shoeb A, Bohorquez J, Dawson J, Guttag J, Chandrakasan AP (2010) A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system. IEEE J Solid State Circuits 45:804–816
53. Chen T-J, Chiueh H, Liang S-F, Chang S-T, Jeng C, Hsu Y-C (2011) The implementation of a low-power biomedical signal processor for real-time epileptic seizure detection on absence animal models. IEEE J Emerging Sel Top Circ Syst 1:613–621
54. Lin YP, Yeh CY, Huang PY, Wang ZY, Cheng HH, Li YT, et al (2015) A battery-less, implantable neuro-electronic interface for studying the mechanisms of deep brain stimulation in rat models. IEEE Trans Biomed Circ Syst, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6081950&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6081950
55. Chun-Yi Y, Hung-Chih C, Hsi-Pin M (2013) An information hub for implantable wireless brain machine interface. 2013 international symposium on VLSI design, automation, and test (VLSI-DAT), p 1–4
56. Damjan Lampret JB, OpenRISC 1200 IP core specification (preliminary draft) [Online]. http://opencores.org/websvn,filedetails?repname=openrisc&path=%2Fopenrisc%2Ftrunk%2For1200%2Fdoc%2Fopenrisc1200_spec.pdf&rev=645
57. Andraka R (1998) A survey of CORDIC algorithms for FPGA based computers. Presented at the proceedings of the 1998 ACM/SIGDA sixth international symposium on field programmable gate arrays, Monterey, CA
58. Ayinala M, Parhi KK (2013) FFT architectures for real-valued signals based on radix-2(3) and radix-2(4) algorithms. IEEE Trans Circ Syst I-Reg Papers 60:2422–2430
59. Wojtecki L, Vesper J, Schnitzler A (2011) Interleaving programming of subthalamic deep brain stimulation to reduce side effects with good motor outcome in a patient with Parkinson's disease. Parkinsonism Relat Disord 17:293–294
60. Isaacson B, Dani S, Desai SA, Denison T, Afshar P (2013) A rapid algorithm prototyping tool for bi-directional neural interfaces. 2013 6th international IEEE/EMBS conference on neural engineering (NER), p 633–636

Chapter 11
Internet of Medical Things: The Next PC (Personal Care) Era

Liang-Gee Chen, Yi-Lwun Ho, Tsung-Te Liu, and Shey-Shi Lu

Abstract Advances in wireless and semiconductor technology have enabled the health care revolution, leading to the emergence of the next personal care (PC) era with the Internet of Medical Things. This chapter will present a comprehensive telecare platform for the next-generation PC era, as well as various innovative biomedical system-on-a-chip (SoC) enablers. In section "Telehealth for the Next Personal Care (PC) Era" of this chapter, we will introduce a fourth-generation synchronous telecare platform that provides 24-h real-time patient monitoring, emergency reminders, and abnormal condition alarms implemented at the National Taiwan University Hospital (NTUH). We will demonstrate the efficacy and efficiency of the telecare platform by applying it to three case studies. The first case study shows that a nursing-led transitional care combining telehealth care and discharge planning can significantly help family caregivers successfully transition from hospital to home. The second and third case studies together show that a fourth-generation synchronous telecare program can further provide both cost-saving and clinical benefits. In section "Biomedical SoC Solutions for the Next Personal Care (PC) Era" of this chapter, we will demonstrate a variety of CMOS biomedical SoC solutions enabling the next PC era, including miniature implementations operating outside and through the body, and implantable solutions inside the body. A CMOS assay SoC for rapid blood screening and a portable gas-chromatography microsystem for volatile compound detection operating outside the body will first be introduced. After that, we will present a 0.5-V biomedical SoC for an intra-body communication system. Three implantable wireless CMOS

L.-G. Chen (✉)
EE2-344, Department of Electrical Engineering, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan
e-mail: lgchen@ntu.edu.tw

Y.-L. Ho
Telehealth Center, No. 7, Chung-Shan South Road, Taipei 100, Taiwan

T.-T. Liu
MD-518, Department of Electrical Engineering, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

S.-S. Lu
EE2-217, Department of Electrical Engineering, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_11

biomedical prototypes, a release-on-demand drug delivery SoC, a pain-control-on-demand batteryless SoC, and a batteryless remotely controlled locomotive SoC, will then be introduced. Finally, we will present several SoC solutions that perform energy-efficient physiological signal processing for real-time biomedical applications in the next PC era.

Keywords Personal care • Telehealth • Telenursing • Discharge planning • Family caregiver • Cardiovascular diseases • Age factors • Cost-benefit analysis • Biomedical • System-on-chip (SoC) • Biosensor • CMOS • Lab-on-a-chip • Rapid blood test • Micro gas chromatography (μGC) • Volatile organic compound (VOC) • Intra-body communication (IBC) • Drug delivery • Implantable • Batteryless • Dorsal root ganglion • Pain control • Electrolysis • Locomotive • Inductive coupling • Wireless powering • ECG • EEG • ECoG • Spike sorting • Processor • Asynchronous circuit • Energy efficiency

Healthcare spending in most developed countries has already reached more than 8 % of their gross domestic product (GDP) (Fig. 11.1) [1]. Among all countries, healthcare spending in the USA is the highest and is expected to reach 19.5 % GDP by 2017 [3, 4]. As a result, the Patient Protection and Affordable Care Act (PPACA), or commonly referred to as "Obamacare," was enacted in 2010 for healthcare system reformation to reduce spending and to improve the quality and efficiency of healthcare. This essentially drives the healthcare system from an expensive "centralized" healthcare routine around hospitals toward a low-cost

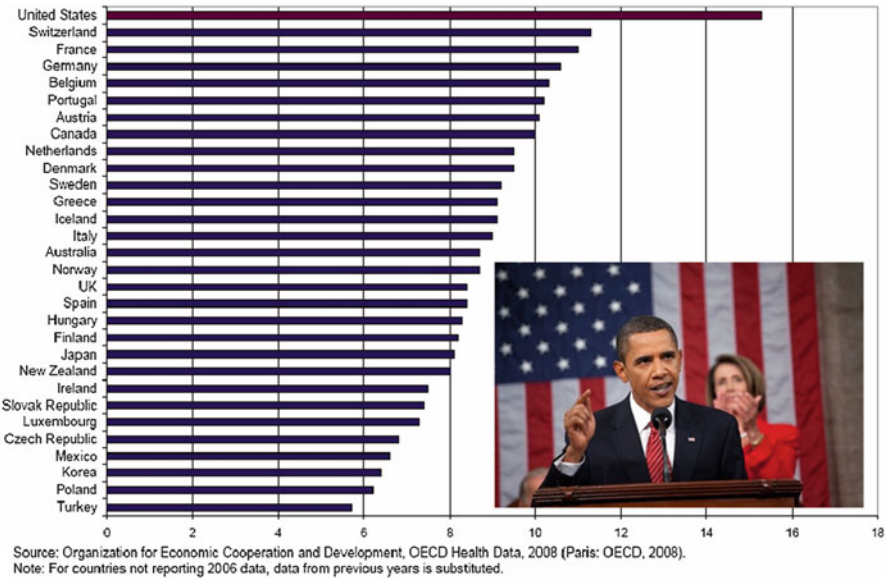


Fig. 11.1 Healthcare spending as a percentage of gross domestic product (GDP) [1, 2]

Fig. 11.2 Remote monitoring of the biometric and clinical data of patients

"distributed" healthcare routine tailored for personal care. In other words, "preventive care and home monitoring will play an important role in increasing the quality of healthcare and decreasing the costs," as stated by Roeen Roashan, an IHS analyst of consumer medical devices and digital health [5].

At the same time, transforming from a centralized healthcare system to a distributed healthcare system demands innovative medical devices and equipment that provide both better performance and affordability. As shown in Fig. 11.2, with the help of wireless medical devices and mobile gateways such as tablets and cellular phones, the biometric parameters of patients can be monitored remotely and closely by professional physicians and nurses. This makes early detection of acute episodes of deterioration and timely intervention possible, which results in real-time closed-loop telehealth systems that can realize distributed personal healthcare from home to hospital.

This chapter therefore introduces two essential components that enable an efficient and affordable next-generation personal healthcare system. We first present a telehealth platform currently deployed at the National Taiwan University Hospital (NTUH). This telehealth platform provides the highest level of patient care with remote monitoring and real-time response. Next, we present a variety of wireless biomedical SoC solutions that support the telehealth platform and enable the next personal care (PC) era with the Internet of Medical Things.

11.1 Telehealth for the Next Personal Care (PC) Era

Long-term healthcare for chronic conditions is one of the greatest challenges to worldwide healthcare systems [6]. This issue is especially important in Taiwan, since elderly people in Taiwan prefer to stay at home close to their family members, rather than in institutions and unfamiliar environments [7]. A telehealth program that incorporates telemonitoring and disease management can therefore provide home-based monitoring and support for patients with chronic diseases.

A telehealth program can be classified into four generations: (1) non-reactive data collection programs, (2) programs with non-immediate analytical structure, (3) remote patient management programs, and (4) fully integrated remote management programs [8]. The fourth-generation synchronous telehealth program provides the highest level of patient care with round-the-clock presence of physicians and nursing staff to analyze and respond in real time to the patient data obtained remotely. The telehealth program at NTUH is a fourth-generation synchronous telehealth program [9, 10] and an Internet-based system. The telehealth platform at NTUH, as shown in Fig. 11.3, is a synchronized, structured, and integrated remote management platform for patients with chronic diseases. Users can access required information on the platform, such as patients' biometric data, electronic medical records, and monthly statistical reports [11]. As shown in Fig. 11.4, the telecare platform horizontally connects the relevant information of electronic health records, patient visits, and future medical treatment planning. Moreover, the

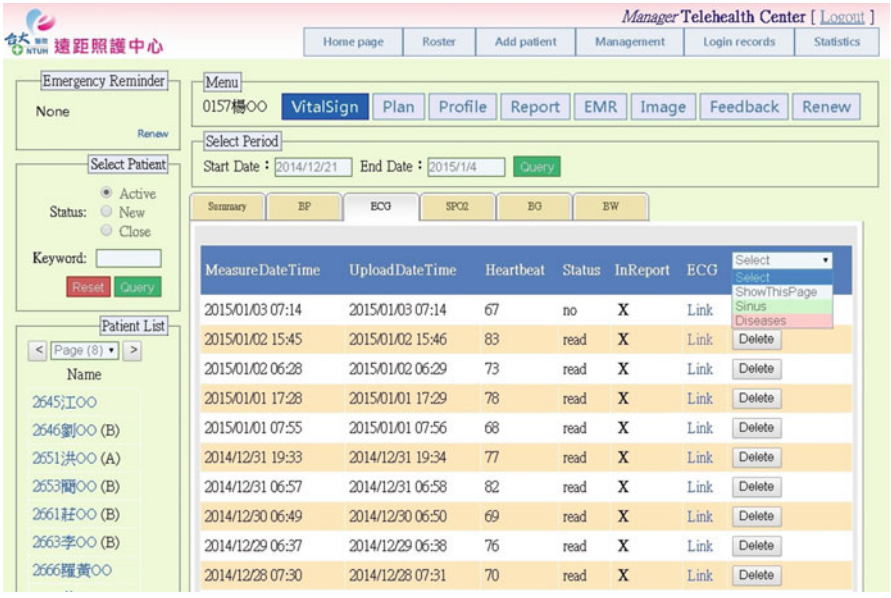


Fig. 11.3 Screenshot of telehealth platform at NTUH [11]

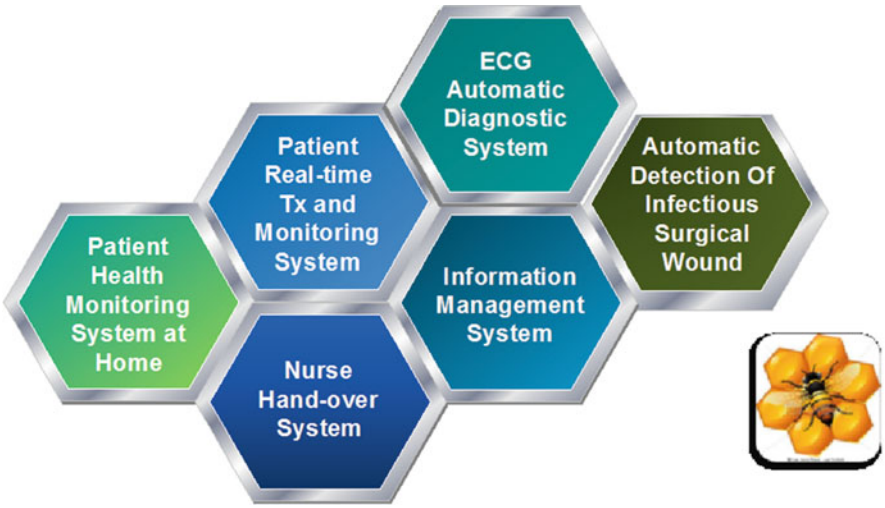


Fig. 11.4 Operating principles and characteristics of telehealth platform at NTUH

telecare platform vertically connects to the providers of medical devices and equipment, exchanging information using the standard HL7 format. It also provides clinical decision support and automatic electrocardiography (ECG) diagnosis.

Several studies have shown that telehealth programs improve the results of long-term care in patients with chronic diseases, including heart failure, chronic respiratory disease, and diabetes [12–16]. In Sect. 11.1 of this chapter, we will further demonstrate the efficacy and efficiency of telehealth programs for the next PC era by presenting three case studies performed at NTUH. The first case study shows that a nursing-led transitional care combining telehealth care and discharge planning could help family caregivers successfully transition from hospital to home by reducing family caregiver burden, enhancing stress mastery, and improving family function [9]. The second and third case studies together show that a fourth-generation synchronous telehealth program could both save costs and provide clinical benefits [10, 17].

11.1.1 The Effectiveness of Telehealth Care on Caregiver Burden, Mastery of Stress, and Family Function Among Family Caregivers

Patients with heart failure must have advanced disease management and appropriate nursing care to help them and their family caregivers to successfully transition from hospital to home [18, 19]. In this section, we evaluate the effectiveness of nursing-led transitional care combining telehealth care and discharge planning on family caregiver burden, stress mastery, and family function [9]. Family caregivers

of heart failure patients receiving telehealth services were evaluated and compared with those receiving only traditional discharge planning during the transition from hospital discharge to home.

11.1.1.1 Experimental Method

The study subjects were sixty patients who suffered from heart failure during May to October 2010. Thirty families in the experimental group participated in telehealth care after discharge from the hospital, while the comparison group comprised thirty families who received only standard discharge planning. After being discharged from the hospital, the experimental group patients were each provided with a telehealth device that connected them to a central platform in the hospital, and their family caregivers were trained to measure patients' physiological parameters at home and to upload the data to the hospital. These data were monitored and analyzed constantly by telenursing specialists, who also provided necessary medical intervention, 24-h health education counseling, and telephonic medical referral services. Data on caregiver burden, stress mastery, and family function were collected from family caregivers twice: once during the discharge planning before the patients' discharge from the hospital and once during the patients' 1-month follow-up visit at the cardiac clinic.

11.1.1.2 Experimental Results

Table 11.1 shows the experimental results at the pretest (discharge) and posttest (1-month follow-up) periods. Caregiver burden was measured by the Caregiver Burden Inventory (CBI) score; the higher the score, the lower the level of well-being. While family caregivers in both groups showed a decrease in CBI at the 1-month follow-up, the degree of improvement was significantly greater in the experimental group, as shown in Fig. 11.5.

The Mastery of Stress Scale (MSS) score was used to measure the mastery of stress related to a caregiver role; the higher the score, the higher the level of mastery of stress. Although family caregivers in both groups improved their mastery of stress, as shown in Fig. 11.6, those in the experimental group showed greater improvement, except in the Acceptance domain. Since this study followed up with the family caregivers for only 1 month, initiating acceptance of the critical situation of the family members in this short-term intervention could be difficult. We believe that family members need a long time to adapt to patients' life-threatening conditions.

Finally, the Feetham Family Functioning Scale (FSSS) score measured the family functioning of the caregivers; a higher score indicates better family functioning. Similar to the results of CBI and MSS, family function in both groups improved at posttest, while the experimental group demonstrated greater improvement, as shown in Fig. 11.7. However, telehealth care had no significant impact on

Table 11.1 Mixed model: repeated measures of caregiver burden, mastery of stress, and family functioning by group [9]

Outcome measure	Pretest[a] Mean ± SD	Posttest[b] Mean ± SD	Between groups, F_b (p)[c]	Within-times, F_w (p)[d]	Interaction, F_{in} (p)[e]
CBI[f] *score*					
Total					
Experimental	43.93 ± 12.39	23.27 ± 10.91	−2.433 (0.382)	−20.667 (<0.001**)	11.433 (<0.001**)
Comparison	41.50 ± 10.12	32.37 ± 9.15			
Time burden					
Experimental	7.60 ± 2.33	4.10 ± 1.88	−0.233 (0.636)	−3.500 (0.004*)	1.533 (<0.001**)
Comparison	7.37 ± 1.85	5.40 ± 1.45			
Development burden					
Experimental	7.53 ± 1.83	3.83 ± 1.93	−1.633 (0.001**)	−3.700 (<0.001**)	2.267 (<0.001**)
Comparison	5.90 ± 1.93	4.47 ± 1.61			
Physiological burden					
Experimental	7.50 ± 2.49	4.13 ± 2.36	0.267 (0.627)	−3.367 (<0.001**)	2.000 (0.002*)
Comparison	7.77 ± 1.89	6.40 ± 1.63			
Emotional burden					
Experimental	7.60 ± 2.54	4.17 ± 2.17	0.167 (0.769)	−3.433 (<0.001**)	1.833 (<0.001**)
Comparison	7.77 ± 2.00	6.17 ± 2.04			
Social burden					
Experimental	6.33 ± 2.41	3.10 ± 2.16	−0.533 (0.301)	−3.233 (<0.001**)	1.967 (<0.001**)
Comparison	5.80 ± 1.65	4.53 ± 1.61			
Cost of care					
Experimental	7.20 ± 2.66	3.97 ± 1.97	−0.533 (0.357)	−3.233 (<0.001**)	1.767 (<0.008*)
Comparison	6.67 ± 2.20	5.20 ± 2.02			

(continued)

Table 11.1 (continued)

Outcome measure	Pretest[a] Mean ± SD	Posttest[b] Mean ± SD	Between groups, F_b (p)[c]	Within-times, F_w (p)[d]	Interaction, F_{in} (p)[e]
MSS[g] *score*					
Total					
Experimental	336.57 ± 19.66	378.53 ± 23.53	2.233 (0.704)	42.933 (<0.001**)	−22.733 (<0.001**)
Comparison	338.17 ± 25.25	358.63 ± 22.16			
Certainty					
Experimental	56.63 ± 3.86	64.73 ± 5.77	3.000 (0.030)	8.100 (<0.001**)	−5.600 (<0.001**)
Comparison	59.63 ± 5.14	62.13 ± 6.13			
Change					
Experimental	54.67 ± 3.70	61.37 ± 5.39	0.567 (0.654)	6.700 (<0.001**)	−4.267 (0.004**)
Comparison	55.23 ± 5.77	57.67 ± 4.41			
Acceptance					
Experimental	55.73 ± 5.21	62.03 ± 6.71	0.767 (0.635)	6.300 (<0.001**)	−0.700 (0.698)
Comparison	56.50 ± 6.11	62.10 ± 6.82			
Growth					
Experimental	58.43 ± 5.06	65.33 ± 6.48	0.933 (0.544)	6.900 (<0.001**)	−4.067 (0.007*)
Comparison	59.37 ± 7.00	62.20 ± 4.90			
Stress					
Experimental	111.00 ± 12.70	125.27 ± 10.94	−3.367 (0.278)	14.600 (<0.001**)	−7.767 (0.018*)
Comparison	107.47 ± 11.51	114.30 ± 12.78			

FFFS[h] *score*					
Total					
Experimental	83.90 ± 13.62	91.47 ± 12.65	−3.267 (0.333)	7.400 (<0.001**)	−5.767 (<0.001**)
Comparison	80.80 ± 13.04	82.40 ± 12.78			
Relationship between family and family members					
Experimental	30.20 ± 4.51	32.20 ± 4.43	−1.900 (0.155)	2.000 (<0.001**)	−0.533 (0.295)
Comparison	28.30 ± 5.61	29.77 ± 5.75			
Relationship between family and subsystems					
Experimental	27.20 ± 4.22	28.60 ± 4.35	−0.533 (0.649)	1.400 (<0.001**)	−1.500 (0.007*)
Comparison	26.67 ± 4.61	26.57 ± 4.83			
Relationship between family and society					
Experimental	26.67 ± 5.82	30.67 ± 5.21	−0.833 (0.557)	4.000 (<0.001**)	−3.733 (<0.001**)
Comparison	25.83 ± 5.53	26.10 ± 5.29			

*p < 0.05; **p < 0.001
[a]Measured at hospital discharge
[b]Measured 30 days after return home
[c]F_b: the F value of between groups comparison
[d]F_w: the F value of within pre- and posttest
[e]F_{in}: the F value of the interaction of between groups and within pre- and posttest
[f]Caregiver burden inventory
[g]Mastery of stress scale
[h]Feetham family functioning scale

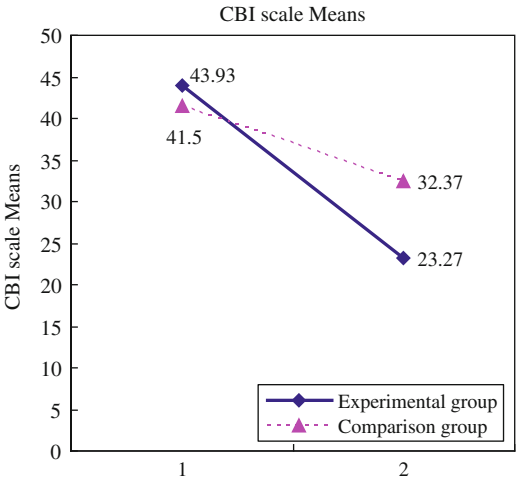


Fig. 11.5 CBI scores of the experimental and the comparison groups [9]

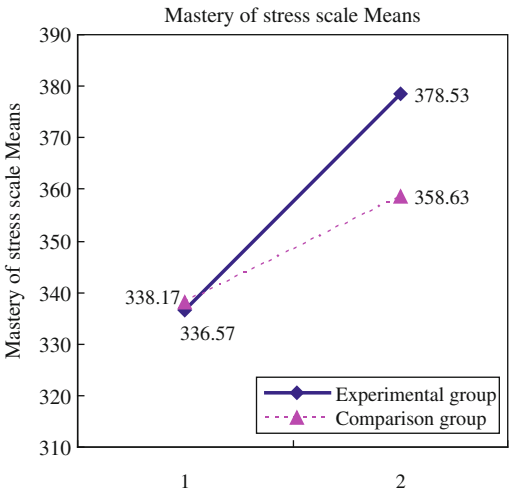


Fig. 11.6 MSS scores of the experimental and the comparison groups [9]

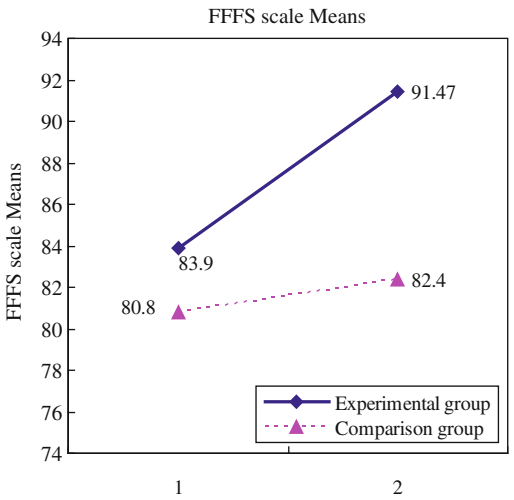


Fig. 11.7 FFFS scores of the experimental and the comparison groups [9]

the subscale score for "relationships between the family and family members." One possible explanation is that telehealth care offered only one outside agent the opportunity to interact with family caregivers through daily communication, and hence, their contact within social networks and community improved. Therefore, the system could improve social interactions but could not ensure enhancement of the relationship inside the family.

In summary, these experimental results show that a nursing-led transitional care combining telehealth care and discharge planning could help family caregivers successfully transition from hospital to home by reducing caregiver burden, increasing stress mastery, and improving family function during the first 30 days at home after patients with heart failure are discharged from the hospital. This demonstrates that in the next PC era, the advanced technology of telehealth care could not only monitor the physical condition of patients during the critical stage of transition from discharge from the hospital to the home but also help, support, and empower family caregivers to achieve a successful transition.

11.1.2 Assessment of the Clinical Outcomes and Cost-Effectiveness of a Fourth-Generation Synchronous Telehealth Program

Although a fourth-generation synchronous telehealth program provides the highest level of patient care, the long-term clinical effect, and the cost-effectiveness of the fourth-generation synchronous telehealth service for patients with chronic cardiovascular diseases have not yet been studied. In this section, we present two studies to evaluate the impacts of a fourth-generation synchronous telehealth program on patients with chronic cardiovascular diseases [10, 17]. The first one is a quasi-experimental study that evaluates whether patients with chronic cardiovascular diseases have better clinical outcomes and cost-effectiveness 6 months before and after the initiation of the fourth-generation telehealth program [17]. The second one is a retrospective cohort study that evaluates whether patients with chronic cardiovascular diseases receiving the fourth-generation telehealth program have better clinical outcomes and cost-effectiveness compared to those receiving only standard healthcare in a longer follow-up period [10].

11.1.2.1 First Study: Experimental Method

The study subjects were 141 patients with cardiovascular diseases who received telehealth services at NTUH from November 2009 to May 2010, including a senior group of 93 patients older than 65 years and a non-senior group of 48 patients. The telehealth services for cardiovascular diseases included (1) real-time transmission of biometrics such as blood pressure, pulse rate, ECG, oximetry, and glucometry

from the patients to the healthcare team for analysis, (2) telephonic communication between the healthcare team and the patients to promote health, and (3) continuous analytical and decision-making support from full-time (24 h) case managers and cardiologists in charge of care. The clinical impact and the cost-effectiveness of the telehealth service on cardiovascular patients were assessed on the basis of the data on hospital visits and health expenditures collected 6 months before and after the initiation of the telehealth service.

11.1.2.2 First Study: Experimental Results

Table 11.2 shows the admission rates and the duration of hospital stays 6 months before (pre) and after (post) the initiation of the telehealth service for patients with cardiovascular diseases in the non-senior and senior groups. After receiving the telehealth services, patients of both groups had a significantly decreased rate of all-cause admission per month, a significantly decreased average day per month duration of all-cause hospital stay, and a significantly increased rate of all-cause outpatient visit per month. The patients in this study did not have any all-cause emergency department visits during the study period. These experimental results demonstrate that synchronous telehealth intervention may reduce both the all-cause admission rate and duration of all-cause hospital stay for patients with cardiovascular diseases regardless of age.

Table 11.3 shows the monthly average cost per patient (in US$) for the 6-month periods before (pre) and after (post) the initiation of the telehealth service. In both groups, the expenditure during outpatient care per month increased, while the expenditure during inpatient care per month decreased. The expenditure during emergency department care increased in the senior group, but decreased in the non-senior group. The total cost of all-cause healthcare, nonetheless, decreased significantly in both groups after initiation of the telehealth service.

In summary, these experimental results show that the fourth-generation telehealth service can decrease the all-cause admission rate, reduce the duration of all-cause hospital stay, and lower the total cost of all-cause healthcare for patients with cardiovascular disease regardless of age. This is particularly important for patients older than 65 years whose application of the telehealth service was in doubt, they also benefitted from reduced healthcare expenditures and better clinical outcomes when their illnesses were managed by the telehealth service. This demonstrates that the advanced technology of the fourth-generation synchronous telehealth program could both save costs and provide clinical benefits in the next PC era.

11.1.2.3 Second Study: Experimental Method

The study subjects were 1754 patients with chronic cardiovascular diseases from December 2009 to April 2013. The case group consisted of 576 patients aged

Table 11.2 Admission rates and duration of hospital stays for patients with cardiovascular diseases, stratified according to age [17]

	Age group, median (IQR)								
	Non-seniors (<65 years)			Seniors (<65 years)			Total, median (IQR)		
	n = 48			n = 93			N = 141		
Final measure	Pre	Post	*P* value	Pre	Post	*P* value	Pre	Post	*P* value
All-cause admission rate[a]	0.09 (0–0.14)	0 (0–0)	0.002	0.10 (0–0.18)	0 (0–0)	<0.001	0.10 (0–0.17)	0 (0–0)	<0.001
Duration of all-cause hospital stay[b]	0.70 (0–1.96)	0 (0–0)	<0.001	0.59 (0–2.24)	0 (0–0)	<0.001	0.60 (0–2.2)	0 (0–0)	<0.001
All-cause outpatient visits[a]	0.77 (0.20–1.64)	1.60 (1.06–2.57)	0.002	1.40 (0.53–2.63)	1.76 (1.12–2.75)	0.02	1.17 (0.36–2.27)	1.70 (1.15–2.72)	<0.001
All-cause emergency department visits[a]	0 (0–0.10)	0 (0–0.20)	0.10	0 (0–0.14)	0 (0–0)	0.24	0 (0–0.13)	0 (0–0)	0.06

[a]Visits per month per person
[b]Day(s) per month per person

Table 11.3 Monthly average cost per patient with cardiovascular diseases, stratified according to age [17]

	Non-senior, mean (SD)			Senior, mean (SD)			Total participants, mean (SD)		
	n = 48			n = 93			N = 141		
Final measure	Pre	Post	*P* value	Pre	Post	*P* value	Pre	Post	*P* value
Outpatient cost	127.08 (309.34)	263.51 (569.44)	0.04	137.57 (253.47)	153.21 (215.45)	0.08	134.00 (272.71)	190.76 (376.98)	0.007
Inpatient cost	814.93 (1000.40)	217.39 (771.01)	0.001	768.27 (1148.20)	301.14 (926.92)	<0.001	784.15 (1096.74)	272.63 (875.08)	<0.001
Emergency department cost	12.76 (26.89)	4.16 (12.76)	0.01	22.3 (35.18)	40.51 (120.93)	0.11	19.09 (32.82)	28.14 (99.82)	0.007
Total cost of all-cause healthcare	954.78 (998.70)	485.06 (952.47)	<0.001	928.20 (1194.11)	494.87 (1047.08)	<0.001	937.25 (1127.84)	491.52 (1012.45)	<0.001

20 years or above receiving the telehealth program at NTUH, while the control groups included 1178 patients who visited our cardiovascular center at NTUH and received only conventional healthcare, but did not participate in the telehealth program. The two groups were matched for age, gender, and the Charlson comorbidity index. Because of the different follow-up times for the two groups, the costs and events were adjusted by the follow-up time (month) in the subsequent analysis. The primary clinical outcomes included the rate of hospitalizations, length of hospitalization, and emergency department visits adjusted by the follow-up period. The total costs for the telehealth group were defined as the sum of medical costs and intervention costs, while the total costs for the control group were only the medical costs. The intervention costs of the telehealth program included direct costs (in-house staff costs, contract costs, and fees to other organizations) and indirect costs (marketing, business development, and administrative costs). The cost-effectiveness was evaluated by the cost saved for each hospitalization that was averted and the cost per hospitalization stay that was averted.

11.1.2.4 Second Study: Experimental Results

Table 11.4 shows the clinical events adjusted by follow-up months for two study groups. There were significantly fewer emergency department visits, hospitalizations, hospitalization days, and intensive care unit admissions per month in the telehealth group compared with the control group. The outpatient department visits between two groups, nevertheless, exhibited similar results. Table 11.5 shows the multivariate Cox regression analysis for event-free survival. Repeated events Cox regression analysis results showed significantly longer hospitalization-free survival for the telehealth group, but the emergency department visit-free survival of the telehealth group was not significantly longer. While age, the Charlson comorbidity index, and telehealth were independent predictors for re-hospitalization and repeated hospitalization, only age and the Charlson comorbidity index were independent predictors for an emergency department visit. These experimental results show that in general a fourth-generation telehealth program can reduce the rate of hospitalizations and the length of hospital stay for patients with chronic cardiovascular diseases.

Table 11.4 Clinical events, adjusted by follow-up months [10]

Events	Cases (n = 576) mean (SD)	Controls (n = 1178)	*P* value
Follow-up months	20.4 (11.4)	25.8 (14.5)	<0.001
ED visits	0.06 (0.13)	0.09 (0.23)	<0.001
Hospitalizations	0.05 (0.12)	0.11 (0.21)	
Hospitalization days	0.77 (2.78)	1.4 (3.6)	
ICU admissions	0.01 (0.07)	0.04 (0.14)	
OPD visits	1.57 (1.12)	1.66 (1.78)	0.75

ED emergency department, *ICU* intensive care unit, *OPD* outpatient department

Table 11.5 Multivariate Cox regression analysis for event-free survival [10]

	Time to first hospitalization		Time to first emergency department visit		Hospitalization, multiple event	
	Hazard ratio	*P*	Hazard ratio	*P*	Hazard ratio	*P*
Age	1.01 (1.01–1.02)	<0.001	1.01 (1.0–1.01)	<0.001	1.01 (1.0–1.02)	0.013
Sex	1.11 (0.97–1.29)	0.13	1.01 (0.86–1.19)	0.9	0.94 (0.69–1.29)	0.71
Telehealth	0.76 (0.65–0.89)	0.001	1.11 (0.94–1.35)	0.19	0.5 (0.34–0.74)	0.001
Charlson comorbidity index	1.23 (1.19–1.28)	<0.001	1.3 (1.25–1.35)	<0.001	1.41 (1.32–1.52)	<0.001

Table 11.6 Medical cost (US$ per patient/month) [10]

Medical costs		Case mean (SD) ($)	Control mean (SD) ($)	*P* value
By clinical setting	Emergency department costs	20.90 (66.60)	37.30 (126.20)	<0.001
	Hospitalization costs	386.30 (1424.30)	878.20 (2697.20)	<0.001
	Outpatient clinic visit costs	180.40 (278.60)	248.20 (984.60)	0.06
	Total medical costs	587.60 (1497.80)	1163.60 (3036.60)	<0.001
	Total healthcare costs	812.40 (1497.80)	1163.00 (3036.60)	<0.001
By items	Laboratory examinations	66.10 (171.10)	120.2 (270.90)	<0.001
	Imaging	20.00 (56.20)	56.40 (150.10)	<0.001
	Medication	130.00 (304.00)	226.60 (864.50)	0.009
	Other treatment and management	56.10 (286.60)	81.30 (315.00)	0.11
	Physician visit	16.10 (65.20)	26.40 (69.40)	0.003
	Nursing	42.60 (224.30)	69.40 (244.60)	0.03
	General ward	51.90 (240.00)	59.70 (212.40)	0.49
	ICU[a] ward	19.20 (135.70)	30.30 (146.10)	0.13

[a]*ICU* intensive care unit

Table 11.6 shows the medical and healthcare costs for the two different study groups. The average total healthcare costs per month in the telehealth group were US$821.4, including a medical cost of US$587.6 and an intervention cost of US $224.8. This was lower than the total costs in the control group, which comprised medical costs only. The generalized linear model (GLM) analysis shown in Table 11.7 revealed that telehealth, heart failure, and cancer were significantly associated with the total cost. The emergency department costs, hospitalization costs, outpatient clinic visit costs, and the composite medical costs in the control group were higher compared with those in the telehealth group. Note that the hospitalization costs accounted for the largest portion of the total costs and were significantly higher in the control group. The GLM analysis also showed that only telehealth was significantly associated with the hospitalization costs. These experimental results show that a fourth-generation telehealth program can reduce the medical costs for patients with chronic cardiovascular diseases, even with additional intervention costs.

Figure 11.8 shows 5000 bootstrapped replicates of incremental costs with respect to the rate of hospitalization, hospitalization days averted, and emergency department visits averted on a cost-effectiveness plane. Since 99.9 % of the 5000 bootstrapped replicates were in the cost-saving quadrant for all three analyses, we can conclude that the telehealth program was a dominant strategy.

In summary, these experimental results show that the fourth-generation telehealth program can reduce the rate of hospitalization, the length of hospital stay, and the accompanying medical costs for patients with chronic cardiovascular diseases. The additional intervention costs associated with this new generation of

Table 11.7 Generalized linear models for costs [10]

	Outpatient department		Emergency department		Hospitalization		Total (medical + intervention)	
Costs/factors	OR	*P*	OR	*P*	OR	*P*	OR	*P*
Age	1.01 (1.0–1.01)	0.01	1.02 (1.01–1.03)	<0.001	1.00 (0.99–1.01)	0.33	1.0 (0.99–1.01)	0.7
Telehealth	0.72 (0.57–0.9)	0.05	0.72 (0.51–1.01)	0.06	0.67 (0.46–0.95)	0.009	0.4 (0.32–0.55)	<0.001
Heart failure	1.61 (1.2–2.17)	0.002	1.47 (1.01–2.12)	0.05	–	–	1.56 (1.11–2.19)	0.009
Diabetes	1.32 (1.02–1.7)	0.03	–	–	–	–		
Liver cirrhosis	4.0 (2.0–8.1)	<0.001	–	–	–	–	–	–
Cancer	–	–	2.3 (1.48–3.6)	<0.001	1.48 (0.97–2.28)	0.07	1.86 (1.23–2.8)	0.003

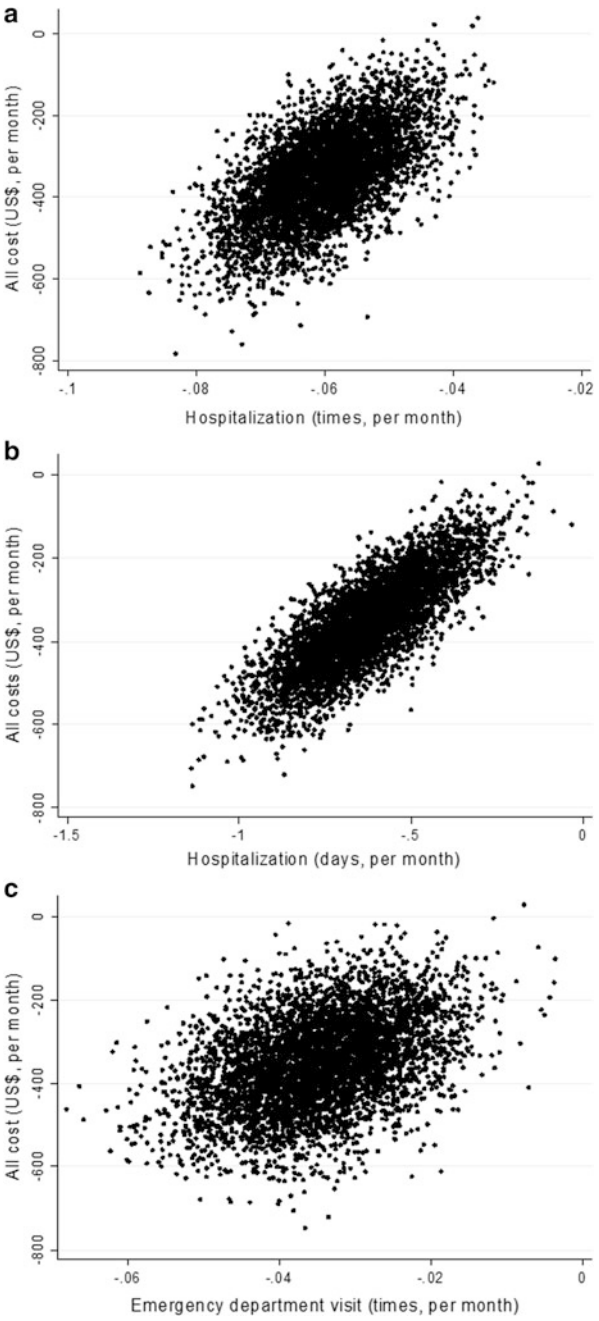


Fig. 11.8 Cost-effectiveness planes for (**a**) hospitalization times, (**b**) hospitalization days, and (**c**) emergency department visits averted [10]

telehealth program did not increase the total costs for patient care. This demonstrates that the use of a fourth-generation synchronous telehealth program can realize both better clinical outcomes and cost-effectiveness in the next PC era.

11.2 Biomedical SoC Solutions for the Next Personal Care (PC) Era

In Sect. 11.2 of this chapter, we demonstrate a variety of CMOS biomedical system-on-a-chip (SoC) solutions enabling the next PC era. Innovative medical electronics not only support the telehealth platform as mentioned in Sect. 11.1 but also can potentially change future medicine practices in a revolutionary way. As shown in Fig. 11.9, several novel biomedical SoC solutions such as DNA and glucose sensing SoCs have been developed at National Taiwan University (NTU) to support future telehealth systems [66]. In this section, we also further introduce a variety of miniature biomedical SoCs for the next PC era: (1) SoC operating outside the body—a CMOS assay SoC for rapid blood screening and a portable gas-chromatography microsystem for volatile compound detection; (2) SoC communicating through the body—a 0.5-V biomedical SoC for an intra-body communication system; (3) SoC operating inside the body—a release-on-demand drug delivery SoC, a pain-control-on-demand batteryless SoC, and a batteryless remotely controlled locomotive SoC; and (4) SoCs performing energy-efficient biomedical signal processing.

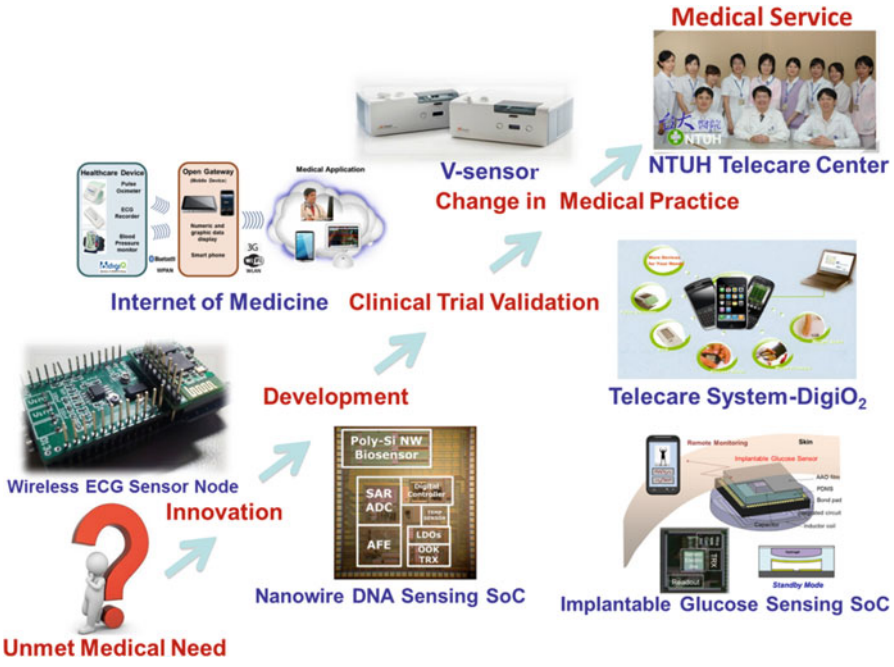


Fig. 11.9 Innovative medical electronics roadmap

11.2.1 Biomedical SoC Outside the Body: CMOS Assay SoC for Rapid Blood Screening

Blood test is an essential procedure for disease assessment and physical condition evaluation. The conventional blood test procedure using enzyme-linked immunosorbent assay (ELISA) is complex and time-consuming, preventing it from becoming a practical solution for self-tested point-of-care applications. As a result, a more efficient blood test tool that offers a simpler procedure and a faster response with a lower cost is essential for the next PC era. In this section, a highly integrated CMOS lab-on-a-chip SoC prototype is therefore presented to realize a rapid and cheap disease-screening solution with only a single drop of blood sample [20].

11.2.1.1 System Architecture

Figure 11.10 shows a CMOS assay SoC for blood-screening immunoassay [20]. The proposed design can provide a rapid blood-screening test of risk prediction by determining whether the concentration of the target biomolecule in a human blood sample, such as tumor necrosis factor-alpha (TNF-alpha) and N-terminal pro-brain natriuretic peptide (NT-proBNP), exceeds the referenced warning threshold. The CMOS assay SoC employs a micro-controller unit (MCU) to orchestrate the whole test procedure that involves five steps: (1) blood filtration, (2) biomolecular conjugation, (3) electrolytic pumping, (4) magnetic flushing, and (5) biomolecular detection. The proposed sandwiched assay detection protocol can avoid the pre-purified antigen process while realizing higher sensitivity and specificity compared with the approach based on direct detection [21]. Figure 11.11 shows the schematic of the proposed CMOS assay protocol flow and mechanism. After receiving a power-on-reset signal (Rst), the blood test is started and the five steps are processed as follows:

1. Blood filtration: A nanoporous aluminum oxide membrane with a pore size of 200 nm diffuses the biomolecules from the blood sample into the mixing reservoir underneath, while preventing the blood cells from leaking into the mixing reservoir. The blood filtration time is controlled by the signal V_{AAO}.
2. Biomolecular conjugation: A simple MEMS-based mixing reservoir prefilled with phosphate-buffered saline (PBS) containing magnetic beads with linked detection antibody is integrated in the CMOS SoC, instead of using a traditional complex microfluidic channel [22]. The filtered biomolecules then conjugate with the detection antibodies attached to the magnetic beads.
3. Electrolytic pumping: After conjugation, an electrolytic voltage $V_{Electrode}$ is applied to the electrodes to generate electrolytic bubbles. The gas force induced by the electrolytic bubbles then pumps the conjugated biomolecular sample into the sensing reservoir [23, 24].
4. Magnetic flushing: The sensing reservoir consists of an 8×8 Hall sensor array and magnetic coils. The pre-immobilized antibodies on the surface of the Hall sensor array capture the conjugated biomolecules on magnetic beads [25], while

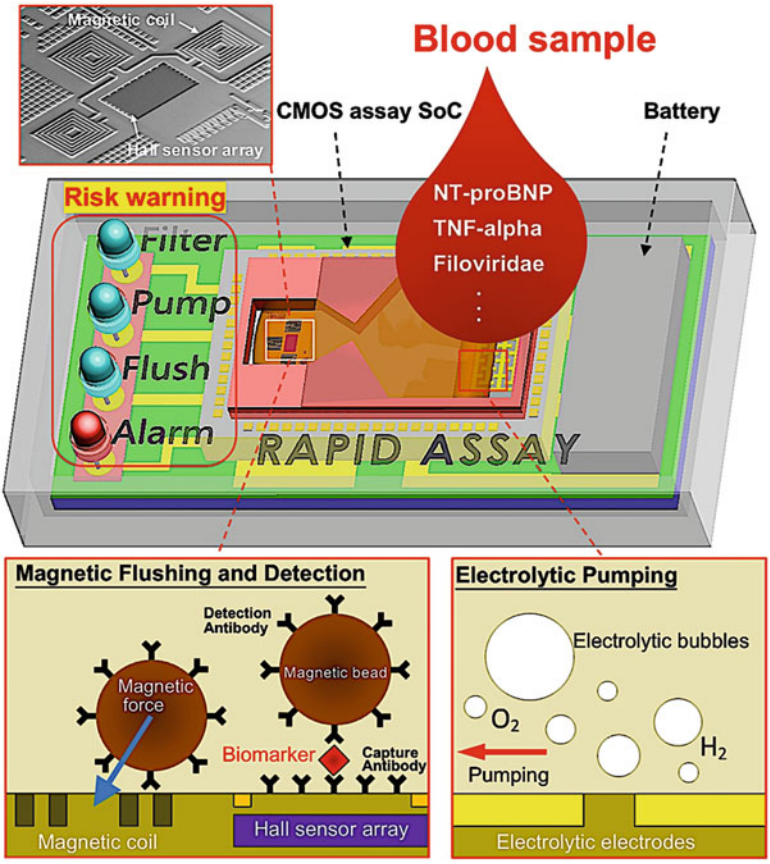


Fig. 11.10 Schematic of the CMOS assay SoC [20]

magnetic coils, driven by a current I_{coil}, generate the magnetic force required to flush out the unbounded magnetic beads from the Hall sensor array.

5. Biomolecule detection: The biomolecules on magnetic beads bounded on the surface of the Hall sensor array are now ready for detection. The MCU issues 64 detection signals (V_{Detect}) scanning the entire Hall sensor array, and sends an alarm signal (V_{Alarm}) to notify the user when the concentration of biomolecules exceeds the preset statistical value.

11.2.1.2 Circuit Implementation

The circuit architecture of the CMOS assay SoC is shown in Fig. 11.12. The SoC consists of a sensor analog front-end, an MCU, and digital control peripheral circuits, which are explained in detail as follows:

1. Sensor analog front-end: The analog front-end is composed of a chopped spinning current circuit, a chopper-stabilized differential difference amplifier

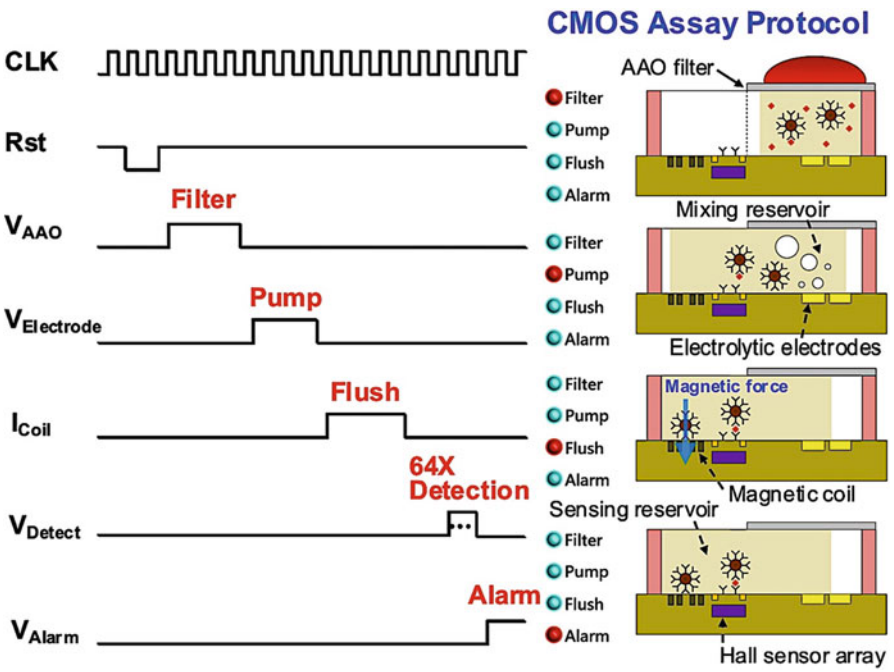


Fig. 11.11 Schematic of the CMOS assay protocol flow [20]

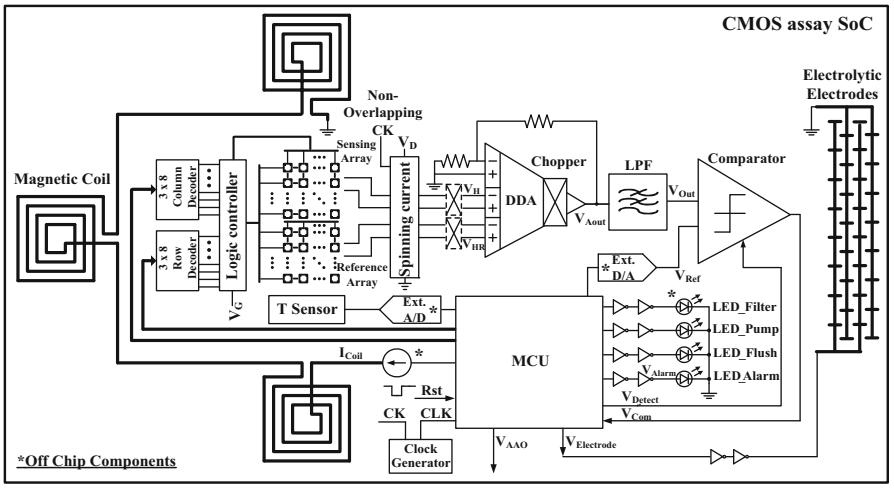


Fig. 11.12 Circuit architecture of the CMOS assay SoC [20]

(DDA), a low-pass filter (LPF), a comparator, and a temperature sensor (T sensor). A chopped spinning current circuit is employed to reduce the dynamic offset resulting from the mismatches between Hall sensors. A chopper-stabilized DDA is then used to amplify and detect the Hall voltage

difference between a sensing Hall sensor voltage V_H and the reference Hall sensor voltage V_{HR}. The DDA circuit consists of a sensing input stage, a reference input stage, a feedback input stage, a transimpedance amplifier (TIA), and a Class-AB output stage. After that, a third-order Bessel-type LPF is employed to filter out the up-converted offset signal. Figure 11.13 shows the measured noise spectra of the LPF output V_{Out} without and with noise reduction techniques using chopped spinning current and chopper DDA circuitry. A reduced input referred noise density of 25 nV/√Hz at a gain of 65 dB demonstrates the effectiveness of the noise reduction technique. Finally, the comparator compares V_{Out} with the preset V_{Ref} determined by the curve plotted for biomolecule concentration against Hall voltage, and then feedbacks the result V_{Com} to the MCU. A T sensor based on the bandgap reference is employed to

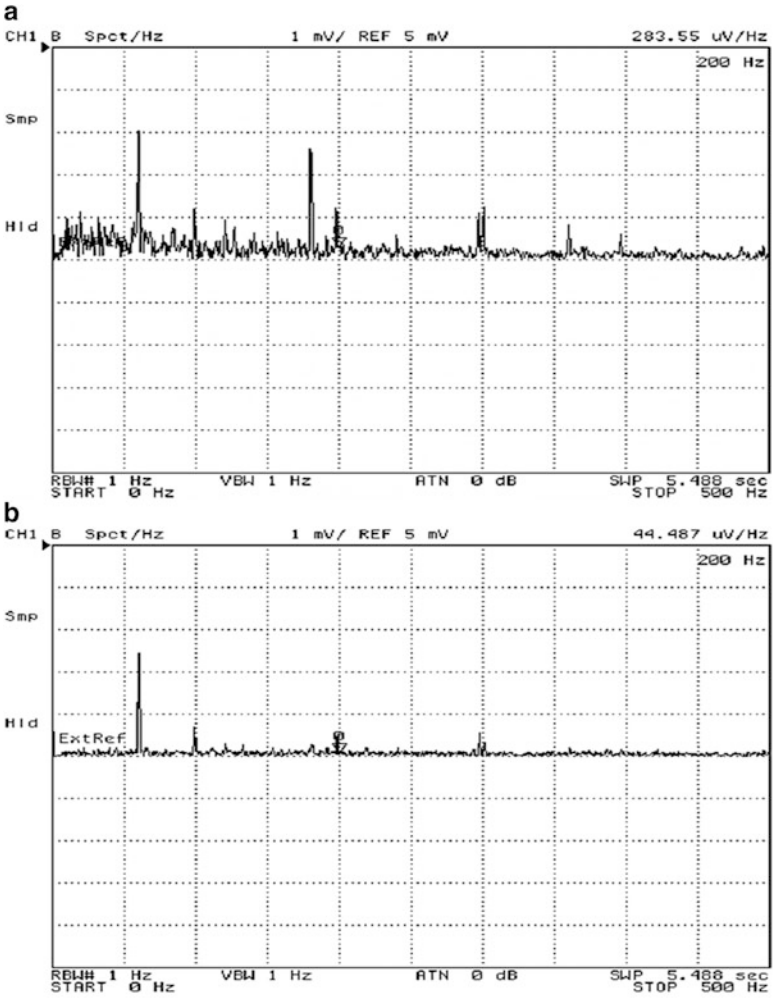


Fig. 11.13 Measured noise spectra of the LPF output (**a**) without and (**b**) with noise reduction techniques [20]

calibrate thermal-induced offset voltage, thereby minimizing the temperature sensitivity of the baseline voltage V_H.
2. MCU and digital control peripheral circuits: The MCU orchestrates the entire five steps of the blood test procedure, and coordinates with digital control peripheral circuits that include a logic controller and row and column decoders. LEDs controlled by the MCU illuminate to indicate different test stages of blood screening, as shown in Fig. 11.11.

11.2.1.3 Experimental Results

Figure 11.14 shows the die photo of the CMOS assay SoC and the image of the post-CMOS MEMS. The first prototype of a CMOS assay SoC is shown in

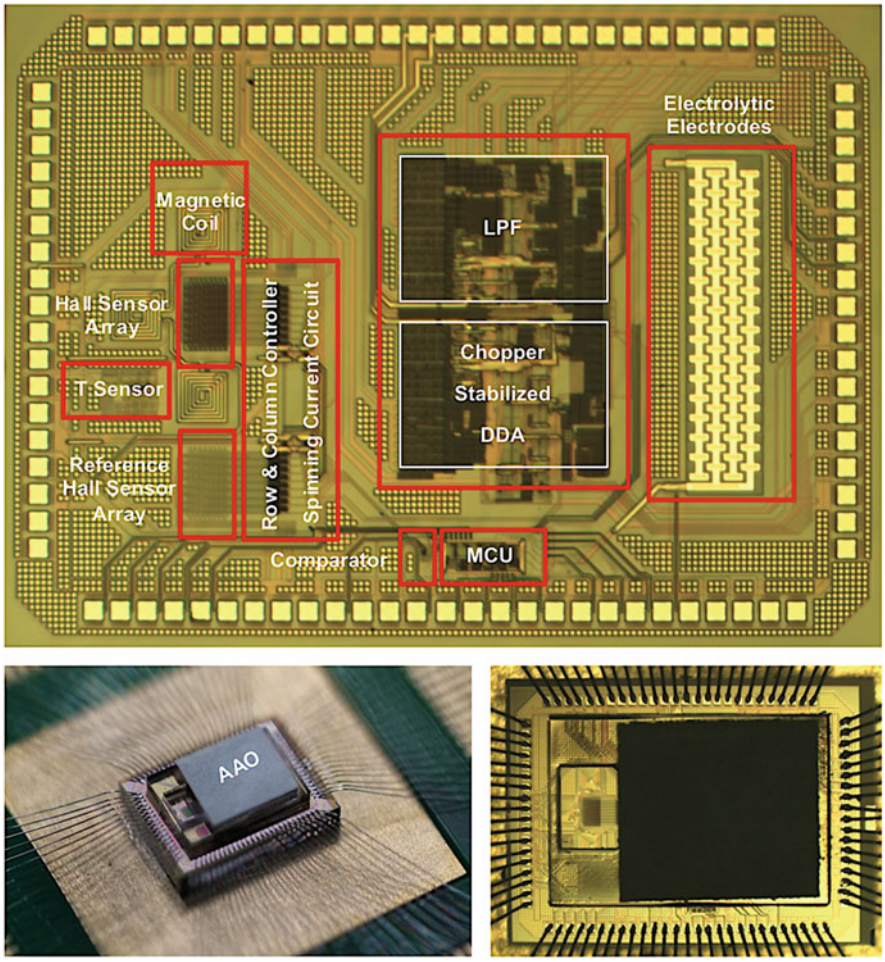


Fig. 11.14 Die photo of the CMOS assay SoC and image of the post-CMOS MEMS [20]

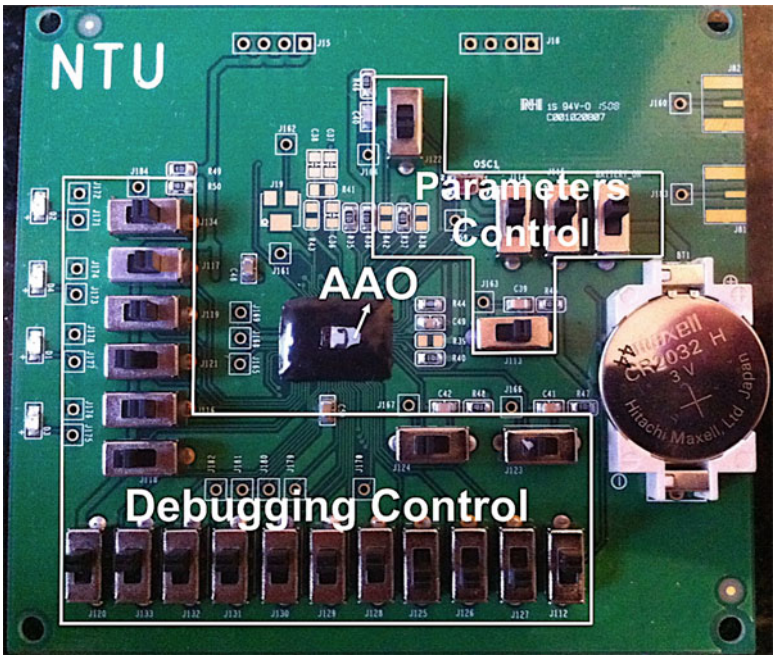


Fig. 11.15 CMOS assay SoC prototype [20]

Fig. 11.15; it is verified by the screening of whole blood samples of normal and abnormal human subjects. Figure 11.16 shows the measured concentration of NT-proBNP through aluminum oxide (AAO) filtration time. As expected, the concentration of NT-proBNP increases with filtration time in both samples, and the proposed SoC can clearly distinguish between the blood samples of normal and abnormal human subjects. Furthermore, in vitro TNF-alpha and NT-proBNP tests were performed to assess the sensitivity of the proposed CMOS assay SoC, as shown in Fig. 11.17. The increase in TNF-alpha concentration is an important sign of the anti-tumor response, while the increase in NT-proBNP concentration indicates possible heart failure response. Based on the clinical data of NTUH, the referenced warning thresholds of TNF-alpha and NT-proBNP concentrations are set at 8.1 and 125 (age <75 years) pg/ml, respectively. The measurement results demonstrate that the CMOS assay SoC can serve as a fast and inexpensive disease-screening device with high sensitivity and specificity for the next PC era.

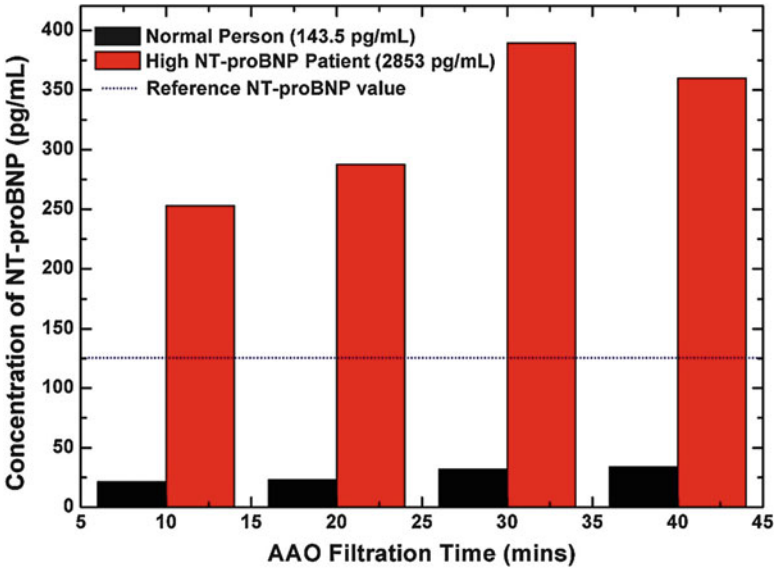


Fig. 11.16 Experimental results of blood filtration as a function of filtration time [20]

11.2.2 Biomedical SoC Outside the Body: Portable Gas-Chromatography Microsystem for Volatile Compound Detection

According to a recent survey in the USA [26, 27], the 5-year survival rate for lung cancer (17.8 %) is much lower than that for many other prevalent cancers, such as breast cancer (90.5 %) and prostate cancer (99.6 %). While the 5-year survival rate for lung cancer cases, which are detected when the disease is still inside the lung and localized, can be as high as 54 %, it is only 4 % when the disease has spread to other organs. However, only 15 % of patients with lung cancer are diagnosed at an early state, since a tumor size smaller than 0.5 cm at an early state is difficult to detect by conventional non-invasive lung cancer diagnostic equipment, such as magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT). Moreover, it takes a long time to complete a scan, and the examinee is exposed to radiation with this equipment. A recent study suggests that volatile organic compounds (VOCs) from human breath gas can be used as biomarkers to detect lung cancer [28]. As a result, using gas-chromatography mass spectrometry (GC-MS) to measure the exhaled air of the examinee can serve as an alternative non-invasive diagnostic method for lung cancer. However, traditional GC-MS equipment is very bulky and expensive and requires specialists to operate. In this section, a portable gas-chromatography microsystem (μGC) that can detect lung cancer-associated VOCs is presented for early diagnosis in the next PC era [29].

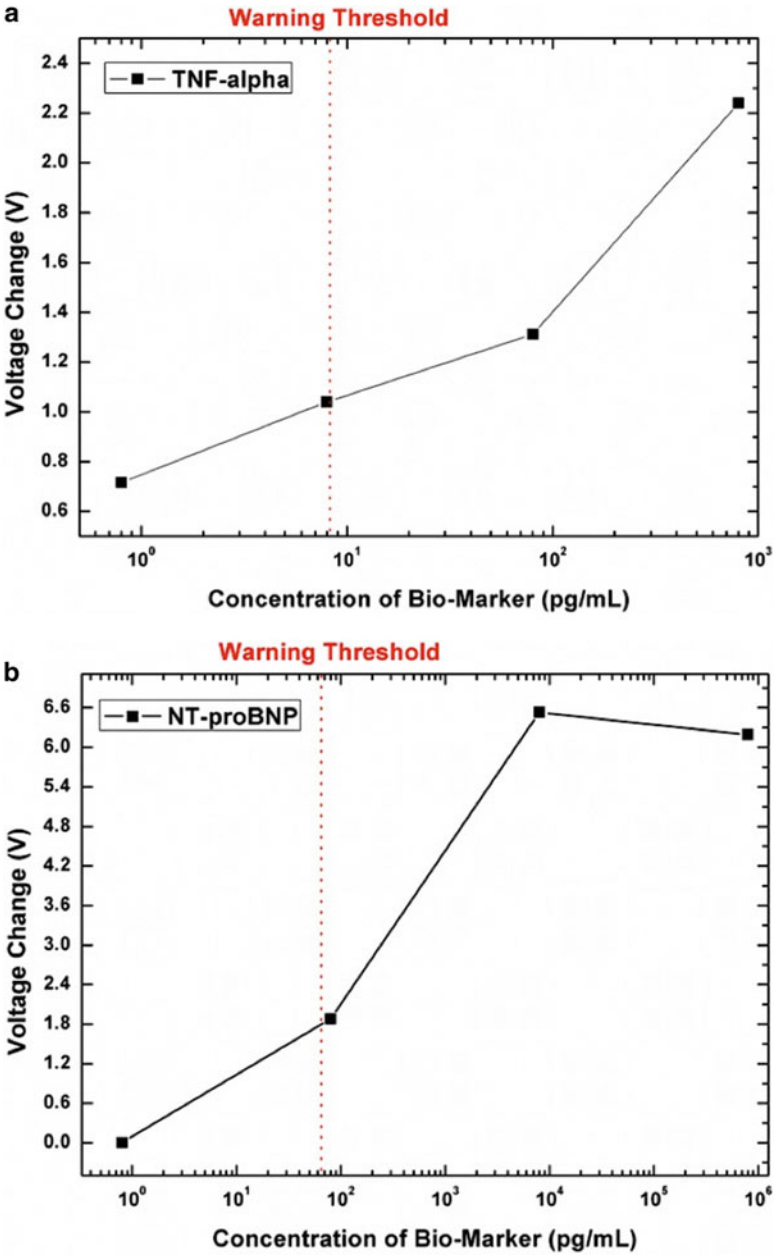


Fig. 11.17 Measured voltage change vs. concentration of (**a**) TNF-alpha and (**b**) NT-proBNP [20]

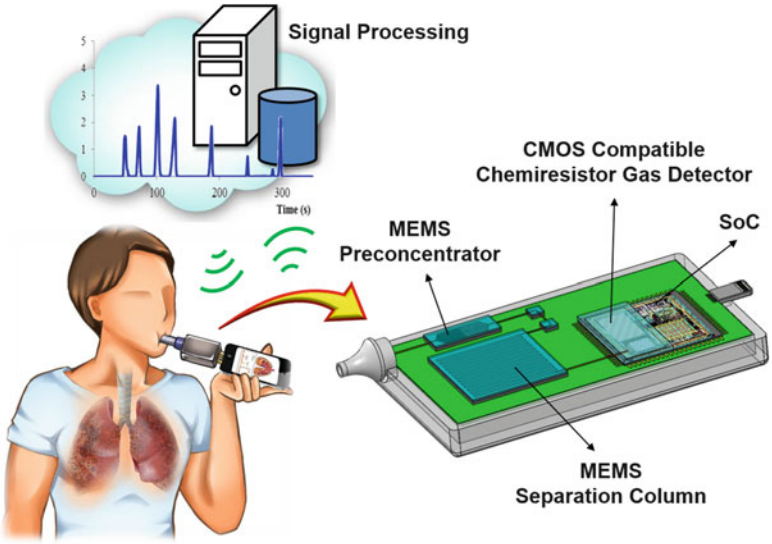


Fig. 11.18 Portable gas-chromatography microsystem (μGC) and its application scenario [29]

11.2.2.1 System Architecture

Figure 11.18 shows a μGC and its application scenario in the next PC era [29]. An examinee can easily use a portable μGC to detect and recode VOCs of the exhaled air anytime and anywhere, transfer the detected VOC data to a smartphone, and finally transmit the data to a cloud server for further signal processing and analysis. The portable μGC consists of an MEMS preconcentrator, an MEMS separation column, a CMOS compatible chemiresist gas detector, and a CMOS SoC. Figure 11.19 shows the operation of the μGC that involves three steps:

1. Target analyte condensation: The MEMS preconcentrator of the μGC, shown in Fig. 11.20a, condenses the target analyte concentration in two successive modes: the sampling mode and the analysis mode. In the sampling mode, the system switches the solenoid valves to allow the gas to flow through the preconcentrator, where the adsorbent traps the target analytes, as shown in Fig. 11.21. Poly (vinylidene chloride-co-vinyl chloride) is used as the adsorbent, and a silver film is deposited on the preconcentrator channel surface as the micro-heater. The preconcentrator factor of sampling 1 ppm toluene is 2170, as shown in Fig. 11.22. After the sampling completes, the system switches to the analysis mode, in which the preconcentrator is heated to desorb the trapped analytes and a miniature gas pump is turned on, as shown in Fig. 11.21. The carrier gas purified by the scrubber then pumps the concentrated analyte samples into the separation column.
2. Analyte separation: The concentrated analytes interact with the stationary phase and form the fragments by a controlled temperature gradient in the separation

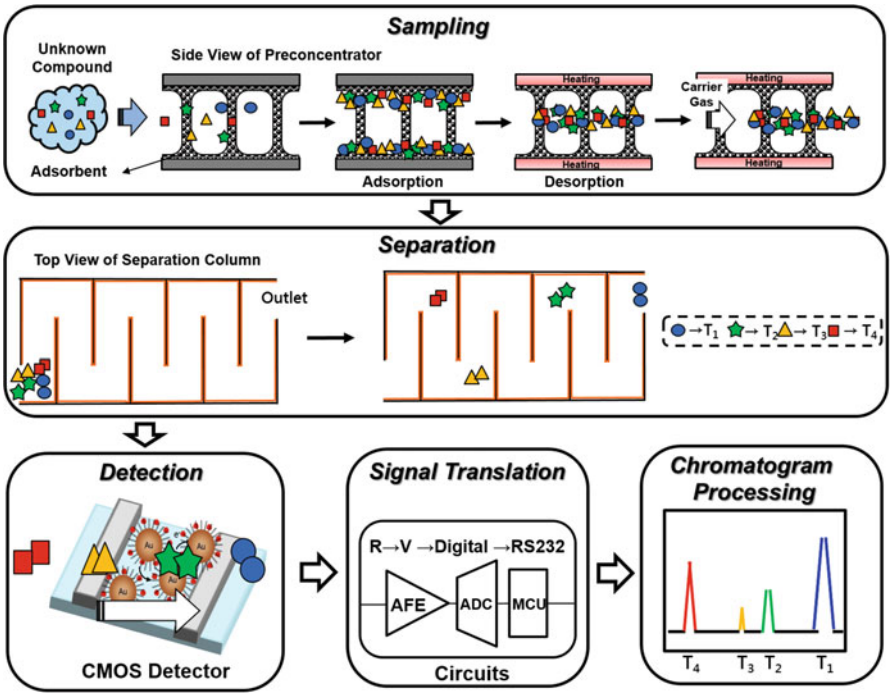


Fig. 11.19 Operation flow of the μGC [29]

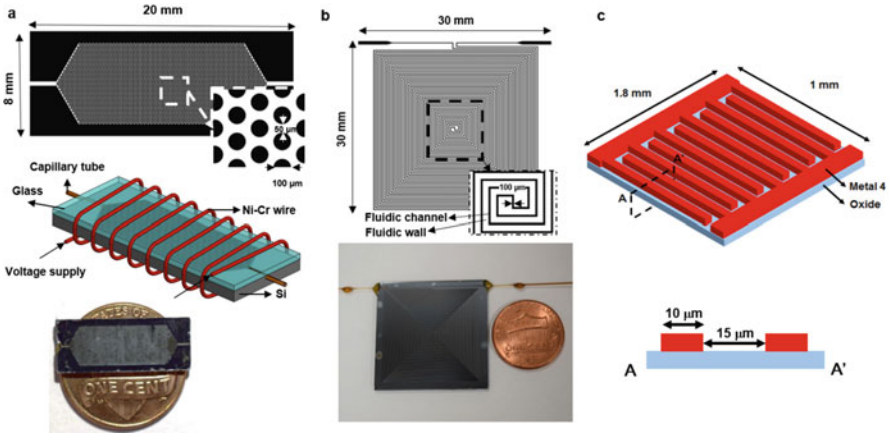


Fig. 11.20 (**a**) MEMS preconcentrator, (**b**) MEMS separation column, and (**c**) universal chemiresist gas detector [29]

column with a long microfluidic channel, as shown in Fig. 11.20b. The microfluidic channel is 3 m long, 200 μm wide, and 0.35 μm deep. The separation column is designed to separate seven lung cancer-associated VOCs:

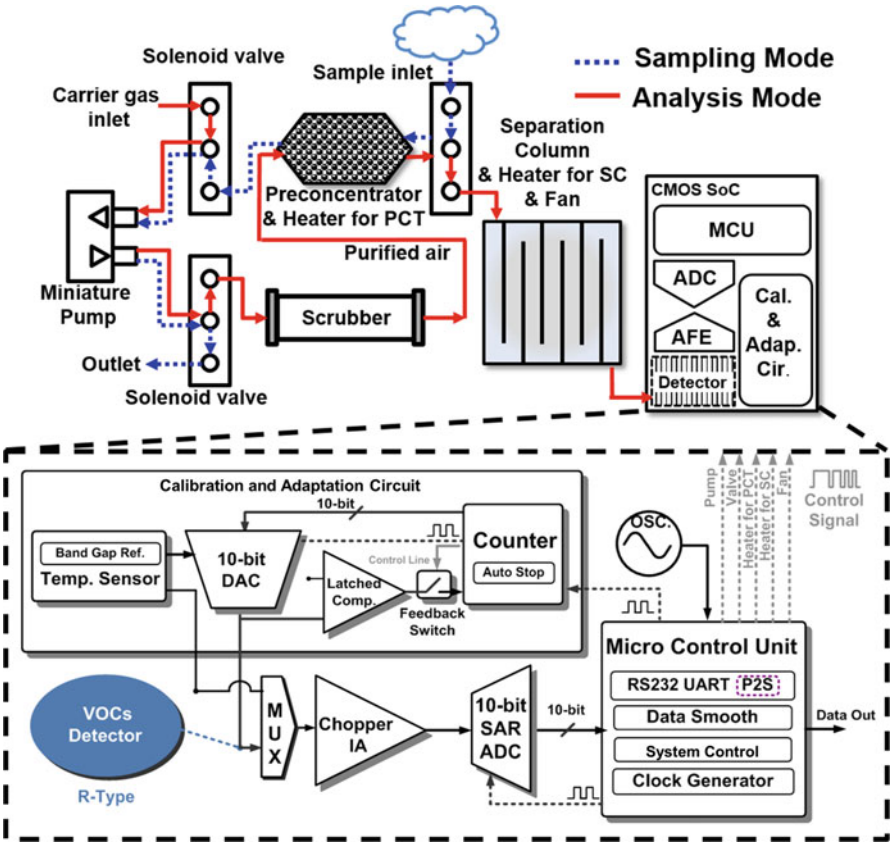


Fig. 11.21 Operating modes and block diagrams of the μGC [29]

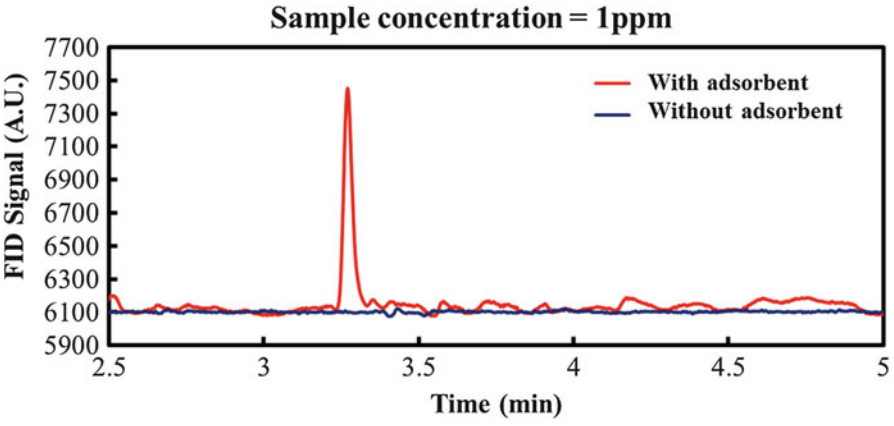


Fig. 11.22 Sample performance of the MEMS preconcentrator [29]

acetone, 2-butanone, benzene, heptane, toluene, m-xylene, and 1,3,5-trimethylbenzene. The analyte fragments finally elute from the separation column to the universal chemiresist gas detector.

3. Analyte detection: Each analyte reacts with the universal chemiresist gas detector based on the CMOS technology, where both interdigitated electrodes and stacked grid electrodes are coated with monolayer-protected gold nanoclusters (MPCs) [30], as shown in Fig. 11.20c. Different kinds of analytes and concentration levels cause various resistance changes, which are detected by the CMOS SoC.

11.2.2.2 Circuit Implementation

The architecture of CMOS SoC consists of an interface calibration circuit, a low-noise analog front-end, a 10-bit SAR ADC, and an MCU, as shown in Fig. 11.21. Since the post-CMOS processes of MPC coating causes the initial resistance of the universal gas detector to vary from 1 to 10 MΩ, a wide range calibration circuit is necessary to calibrate the resistance variation. The complete calibration circuitry is composed of a comparator, a counter, a bandgap reference, and a 10-bit current DAC. The negative feedback loop of the calibration circuitry will convert the resistance change of the universal gas detector into voltage output and calibrate it to the reference voltage. Figure 11.23 shows the measured circuit performance after calibration. To offer higher resolution for low-concentration analytes, a low-noise analog front-end comprising a chopper instrumentation amplifier, and a second-order Sallen–Key LPF is employed. Figures 11.24 and 11.25 show the measured input referred noise density and detection sensitivity. With the low-noise analog front-end, the input referred noise density is reduced to 70 nV/√HZ and the detection sensitivity is up to 15 ppb for 1,3,5-trimethybenzen.

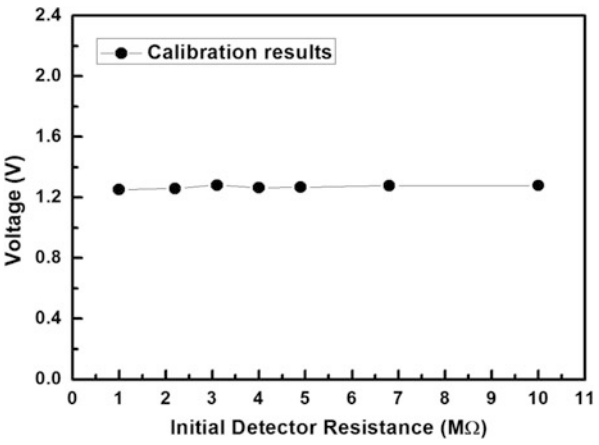


Fig. 11.23 Performance of the calibration circuit [29]

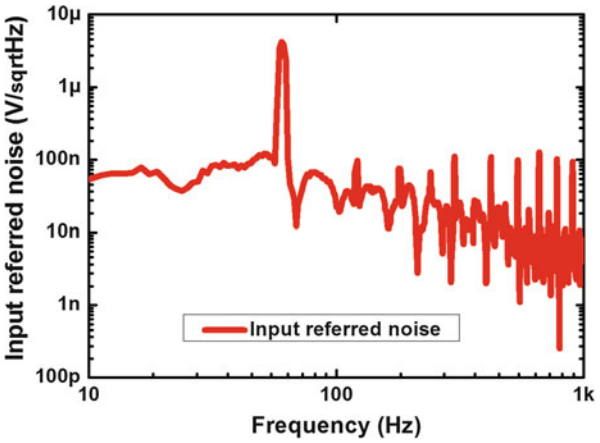


Fig. 11.24 Measured input referred noise density [29]

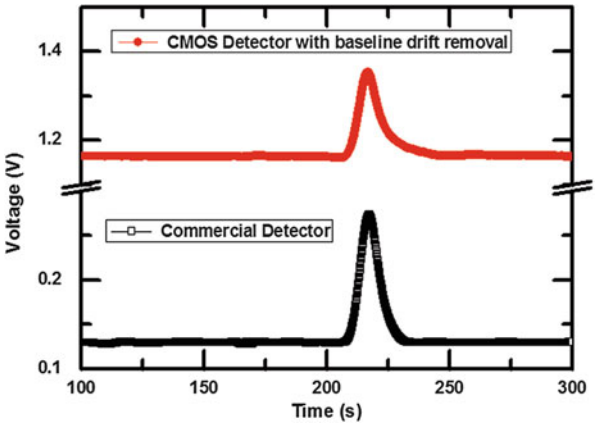


Fig. 11.25 Measured detection sensitivity [29]

11.2.2.3 Experimental Results

Figure 11.26 shows the image of the µGC prototype and the die photo of CMOS SoC. The measured gas chromatographic data are further processed offline through a peak detection and quantification procedure to identify and quantify analytes in a gas chromatogram. The peak detection and quantification procedure involves four steps:

1. Chromatogram smoothing: First, the Savitzky–Golay filter is employed in this study to smooth the measured gas chromatographic data.
2. Baseline correction: A distribution-based classification method is used to estimate the baseline noise of the gas chromatogram [31]. The measured gas chromatogram after smoothing and baseline correction is shown in Fig. 11.27a.
3. Peak detection: The continuous Mexican hat wavelet transform is employed to detect the positions of the gas chromatographic peaks. The coefficient scale of

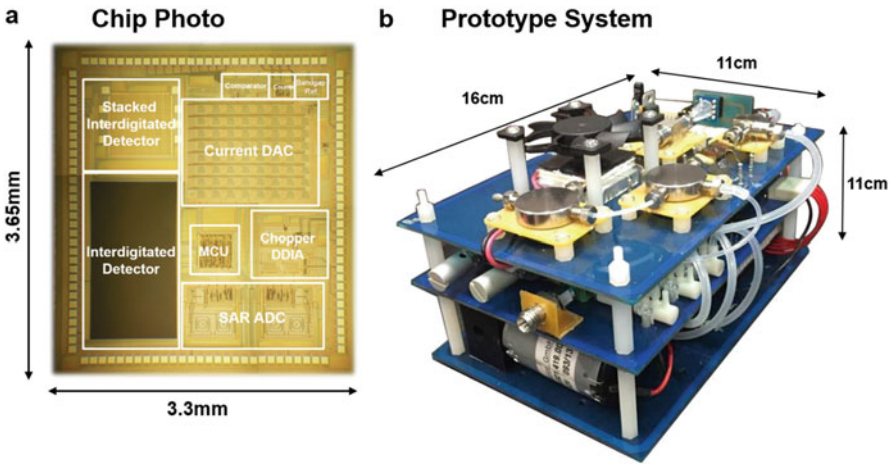


Fig. 11.26 (**a**) CMOS SoC die photo and (**b**) image of the μGC prototype [29]

the continuous wavelet transform (CWT) of each detected peak is then converted to the full width at the half maximum (FWHM) of each detected peak, as shown in Fig. 11.27b.

4. Deconvolution: The exponentially modified Gaussian (EMG) model is selected as the peak shape model. Using the position and FWHM of each detected peak as the initial parameters, the overlapping chromatographic peaks can be deconvoluted and quantified, as shown in Fig. 11.27c.

Table 11.8 summarizes the final detection results of seven lung cancer-associated VOCs after the peak detection and quantification processes. The positions of detected chromatographic peaks are aligned to our in-house VOC retention time (RT) library. The results demonstrate that the portable μGC together with the proposed signal processing method is both efficient and effective to detect lung cancer-associated VOCs for early diagnosis in the next PC era.

11.2.3 Communication Through the Body: Biomedical SoC for Intra-Body Communication System

Intra-body communications (IBCs) employ a human body as the transmission medium, thereby avoiding the use of an antenna, which is subject to the classic tradeoff between the form factor and power dissipation in traditional wireless communications. By choosing a considerably lower carrier frequency than those of the traditional local-area wireless sensor networks, IBC can realize lower power dissipation and a smaller form factor. Moreover, the human-body channel is usually more power efficient than the air over these frequencies of interest. As a result, IBC

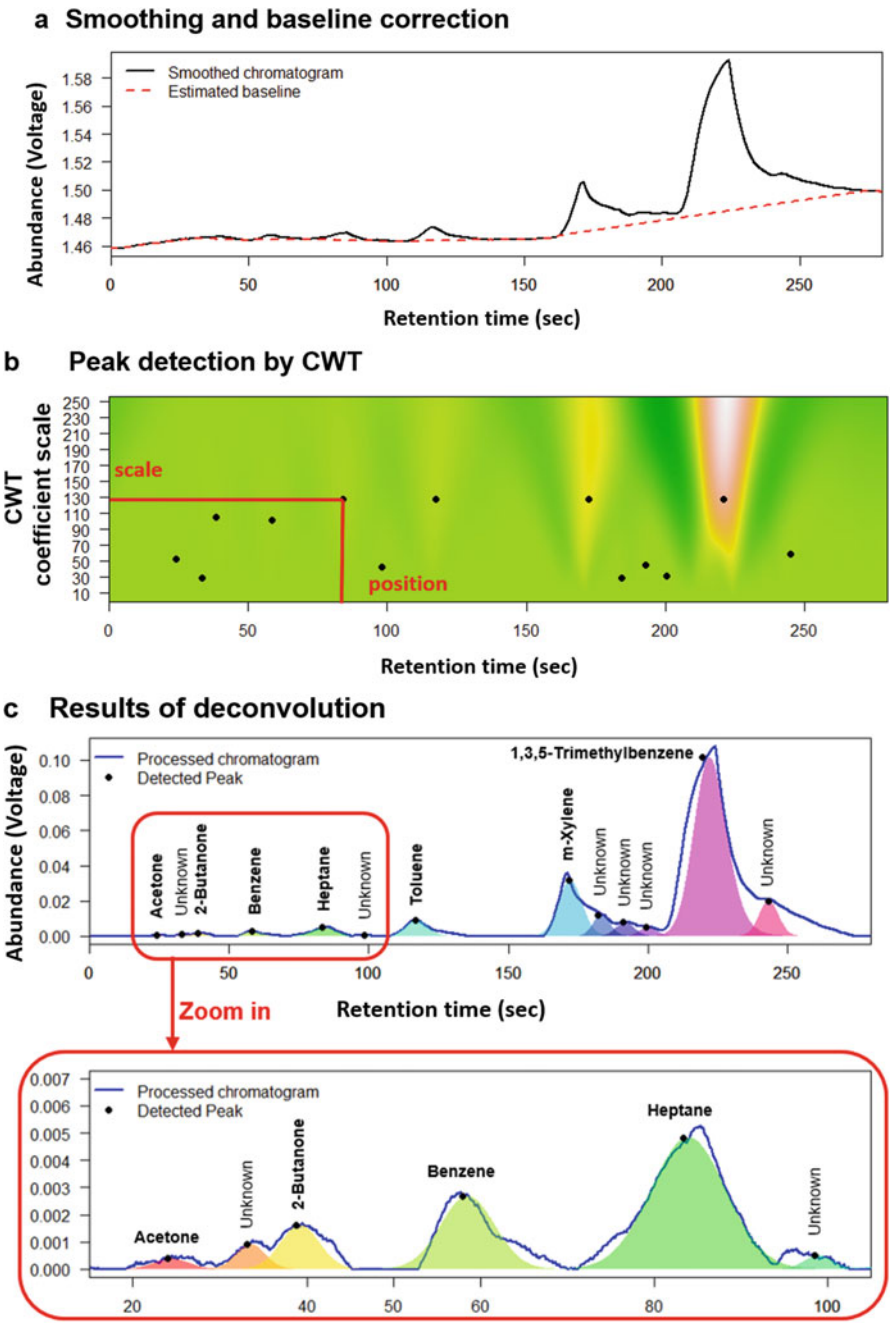


Fig. 11.27 Measured and processed results of seven lung cancer-associated VOCs after (**a**) smoothing and baseline correction; (**b**) after peak detection by CWT; and (**c**) after deconvolution [29]

Table 11.8 Information of seven detected VOCs [29]

VOC name	Abundance (ΔV[a])	RT (s[b])	Area (ΔV s)	SNR	FWHM (s)
Acetone	0.0004	24.2	0.0027	4	6.3
2-Butanone	0.0016	38.5	0.0112	5	6.3
Benzene	0.0027	58.5	0.0217	53	7.4
Heptane	0.0049	84.2	0.0538	113	10.3
Toluene	0.0093	117.3	0.0932	205	9.2
m-Xylene	0.0317	172.3	0.3366	645	9.7
1,3,5-Trimethylbenzene	0.1014	221.4	1.6876	2167	13.7

[a]V is voltage
[b]s is second

is a promising candidate for biomedical applications demanding both a long-time operation and a small form factor in the next PC era. In this section, a fully integrated CMOS IBC SoC is presented for the next PC era [32].

11.2.3.1 System Architecture

Figure 11.28 shows the conceptual diagram of a human-body communication network using IBC [32]. Ag/AgCl electrodes are used to connect an RF gateway and a biomedical SoC to the human body. The RF gateway, attached to the wrist in this application example, comprises a transceiver and an MCU. It serves as a relay station between a remote wire/wireless device and the biomedical SoC attached to the human body. The biomedical SoC consisting of a transceiver, an MCU, and an ADC, wakes up after receiving commands from the RF gateway through the human body, and then transmits the measured biomedical data back to the RF gateway through the human body. The RF gateway finally transmits the collected biomedical data to a cellular phone or a remote computer.

To maximize the communication efficiency of the human-body network, the channel characteristics of the human body have been measured and shown in Fig. 11.29. The human-body channel exhibits characteristics of an LPF and the measurement results show a channel bandwidth of 300 MHz with a transmission loss of 7 dB. It is also suggested that [33] the communication frequency range for IBC should be from 200 to 600 MHz to minimize the path loss. A communication frequency of 200 MHz is therefore chosen for the proposed human-body communication network.

11.2.3.2 Circuit Implementation

The transceiver of CMOS IBC SoC must consume both a low power and small silicon area. Figure 11.30 shows a low power (2.9 mW), miniaturized (520 μm × 220 μm), on–off-keying (OOK) receiver architecture consisting of a

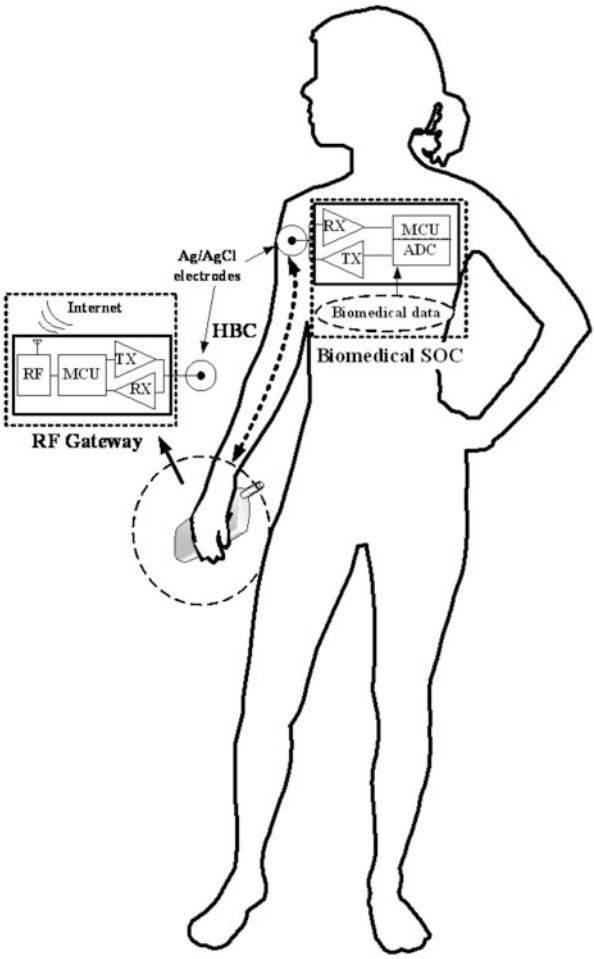


Fig. 11.28 Conceptual diagram of the proposed human-body communication network [32]

low-voltage amplifier (LVA), cascaded gain amplifiers, single-ended to differential (STD) amplifiers, a low-voltage multiplier (LVM), an LPF, a comparator, and buffer circuits. The receiver employs self-mixing methodology without the need to use additional oscillators and demodulators, as shown in Fig. 11.30. The received RF signal with frequency F_C is translated to a DC signal and a higher frequency signal with $2F_C$ because of the self-mixing mechanism. The receiver signal will be finally demodulated and converted to a rail-to-rail signal after the buffer stage.

The LVA uses a common-source amplifier with high input impedance to accommodate the electrode that contacts human skin. The LVM employs the parallel multiplier structure in [34] for high dynamic range and low-power operation. An additional voltage conversion circuit is used to convert the output signal level of the receiver (0.5 V) to the logic level of the MCU (1.8 V). The OOK transmitter is composed of a ring oscillator, a source-follower buffer, and a class-C power

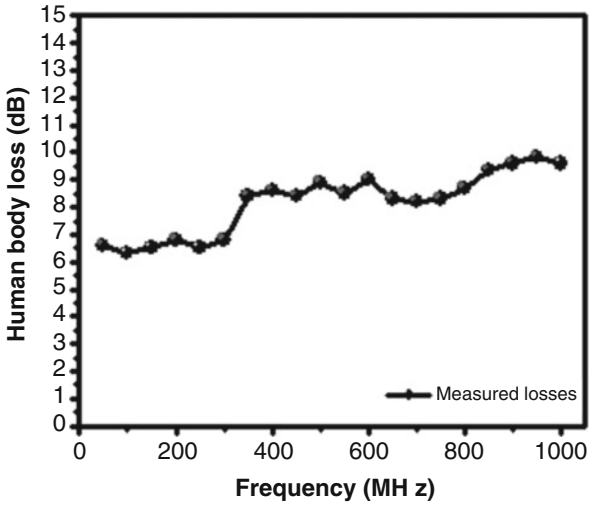


Fig. 11.29 Measured frequency–response characteristics of the human-body channel [32]

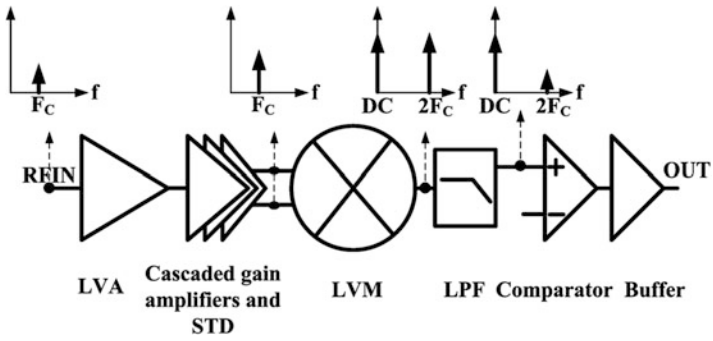


Fig. 11.30 Block diagram of the self-mixing receiver architecture [32]

amplifier. The OOK modulation is realized by switching on and off the ring oscillator that generates the 200-MHz carrier signal. The IBC SoC uses only one off-chip inductor for the power amplifier circuit. The transceiver can be solely powered by a solar cell at 0.5 V, while an additional molecular battery supplies the MCU and ADC operating at 1.8 V.

11.2.3.3 Experimental Results

Figure 11.31 shows the die photo of the CMOS SoC for IBC. The measured bit error rate (BER) versus the input power characteristic of the receiver is shown in Fig. 11.32. For a BER of 10^{-3}, the minimum input carrier power of −68 dBM is required for operation at 145 MHz, while the maximum input power can be up to −5 dBm. This corresponds to 63 dB of the dynamic range without any gain control

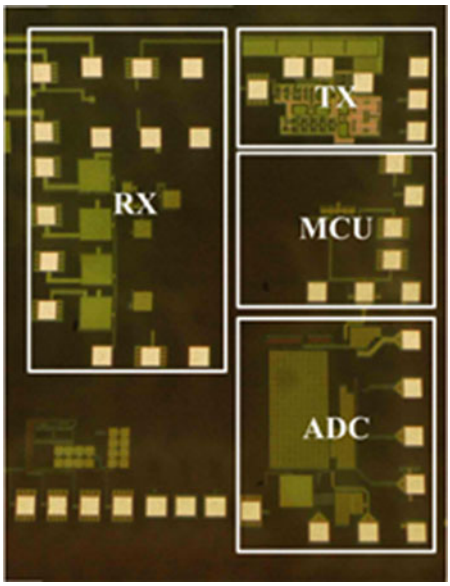


Fig. 11.31 Die photo of the CMOS SoC [32]

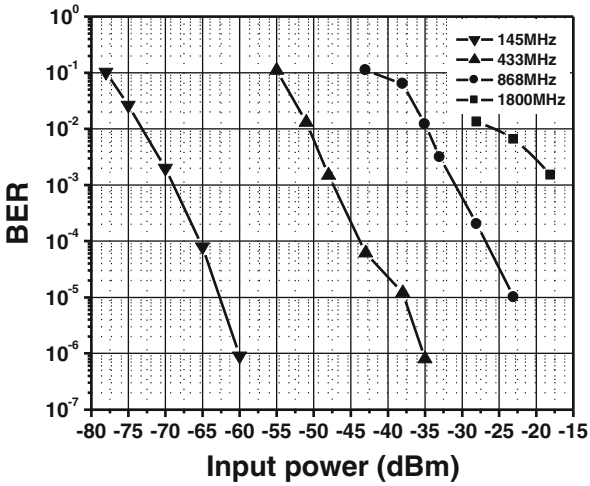


Fig. 11.32 Measured BER versus the input power characteristics of the receiver [32]

circuits, as a result of low-voltage multiplier topology. The maximum data rate is 2 Mb/s.

An experiment was set up to demonstrate the human-body communication network, as shown in Fig. 11.28. The biomedical signals of ECG were first collected by the ADC of the biomedical SoC, transmitted from the transmitter of the biomedical SoC through the human body via electrodes to the receiver of the RF gateway. The ECG signals were finally sent from the transmitter of the RF gateway

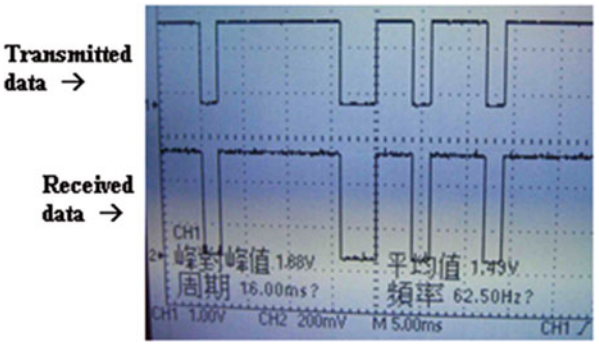


Fig. 11.33 Transmitted (by the transmitter of biomedical SoC) and received (by the receiver of the RF gateway) biomedical data observed by the oscilloscope [32]

Fig. 11.34 ECG signals on a cellular phone screen [32]

to a cellular phone. Figure 11.33 shows the biomedical data obtained by the receiver of the RF gateway. The final ECG signals received at the cellular phone were displayed on the cellular phone screen, as shown in Fig. 11.34. The well-recognized P, Q, R, S, and T waveforms clearly demonstrate a low-power (4.535 mW), small size (1.5 mm^2) IBC biomedical SoC solution for the next PC era.

We have presented several biomedical SoC prototypes that offer in vitro analytical and diagnostic tools outside and through the human body. In the following sections, we will further present three implantable SoC solutions that provide in vivo therapeutic and sensing applications inside the human body.

11.2.4 Biomedical SoC Inside the Body: Implantable Release-on-Demand Wireless CMOS Drug Delivery SoC

The method of drug delivery has a significant impact on the efficacy of drug therapy. Compared with the traditional non-invasive routines of administration,

such as peroral, transdermal, and inhalation routines, novel implantable drug delivery devices that can precisely control key therapy parameters can greatly enhance the efficacy of drug therapy [35]. However, current implantable drug delivery implementations are bulky with a low integration level because of process incompatibility between the drug reservoir and control IC. In this section, a fully integrated implantable CMOS drug delivery SoC with low cost, small size, and low-power consumption is presented for the next PC era. The device can deliver both liquid and solid drug formulations, and can be precisely controlled by physicians or patients non-invasively by wireless means [36].

11.2.4.1 System Architecture

Figure 11.35 shows the application scenario and schematic of the drug delivery system [36]. The entire drug assembly is enclosed in a lightweight, stable, biocompatible round titanium with a small drug outlet. The system is powered by a high-energy-density, rechargeable lithium-ion nanowire battery [37] (4.5 mm in diameter) with a capacity of 223 mAh. The miniature (3×3 mm) spiral loop antenna receives the wireless command signals, and serves as an energy converter that can pick up energy from external sources and store it to the rechargeable battery [38]. The design and operation principles of the drug reservoir and drug delivery SoC are explained in detail in the following two sub-sections.

11.2.4.2 Drug Reservoir Implementation

Each cell within the drug delivery array consists of a reservoir containing drug formulations, a metal membrane capping the reservoir, and metal trances directing

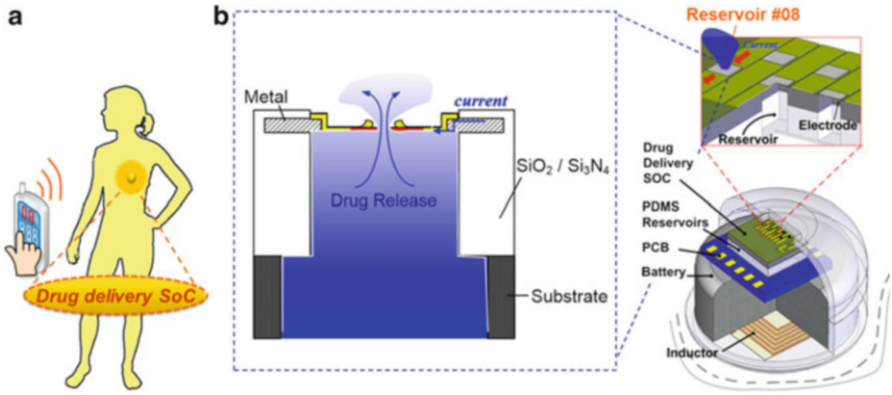


Fig. 11.35 (**a**) Application scenario and (**b**) schematic of the drug delivery system [36]

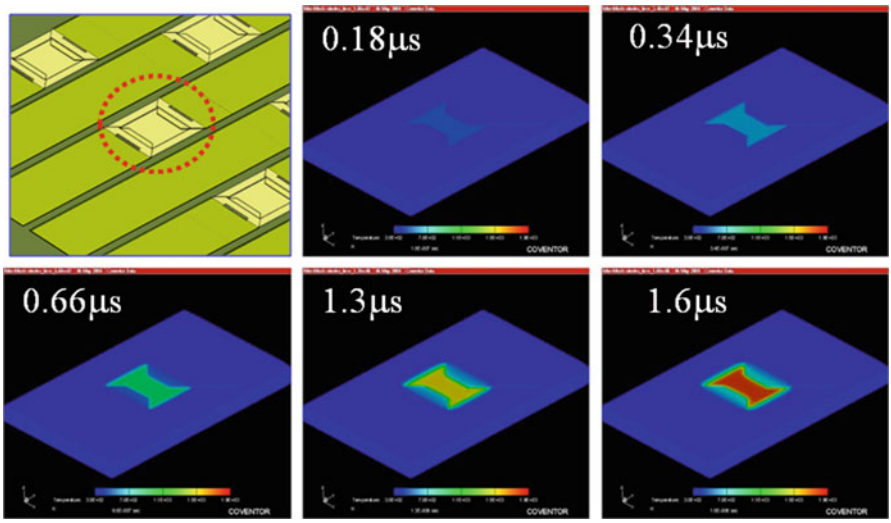


Fig. 11.36 Simulated results of the temperature distribution on a heated membrane in air [36]

electrical current to the top of the membranes. The membranes, composed of CMOS-compatible, biocompatible titanium (Ti), and platinum (Pt) films, are realized by post-IC photolithography and a lift-off process [39]. The cavities for reservoirs are formed by CMOS-compatible post-IC deep etching from the backside of the die, with a polydimethylsilicane (PDMS) layer to increase the volume capacity of the reservoirs [39]. The activation process of each individual cell reservoir is similar to the operation of an electrical fuse. As the electric current passes through the membrane, joule heating raises the temperature at the center of the membrane. Once the membrane is heated to the point of failure, the drug contained in each cell reservoir is released. Figure 11.36 shows the results of a simulation conducted for the temperature distribution on a heated membrane in air by using an FEM electrothermal simulator (ANSYS). The temperature at the center of the membrane increases rapidly and reaches the point of failure (about 200 K for Ti) in 1.6 μs with a 3 V activation voltage.

11.2.4.3 SoC Circuit Implementation

The wireless drug delivery SoC shown in Fig. 11.37 mainly consists of an OOK receiver and an MCU. A receiver based on the OOK modulation scheme can realize low-power consumption and small die size without the need for mixers and VCO. The OOK receiver is composed of a common-source feedback pre-amplifier, a cascaded amplifier, an envelope detector, and a comparator with an output buffer. The measured sensitivities of the OOK receiver at different frequencies are shown in Fig. 11.38. The receiver can achieve a sensitivity of −61 dBm for a data rate of

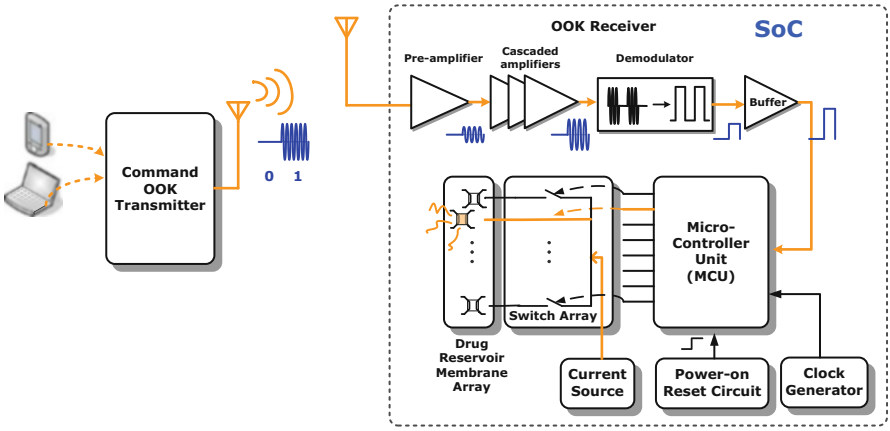


Fig. 11.37 System architecture of the drug delivery SoC [36]

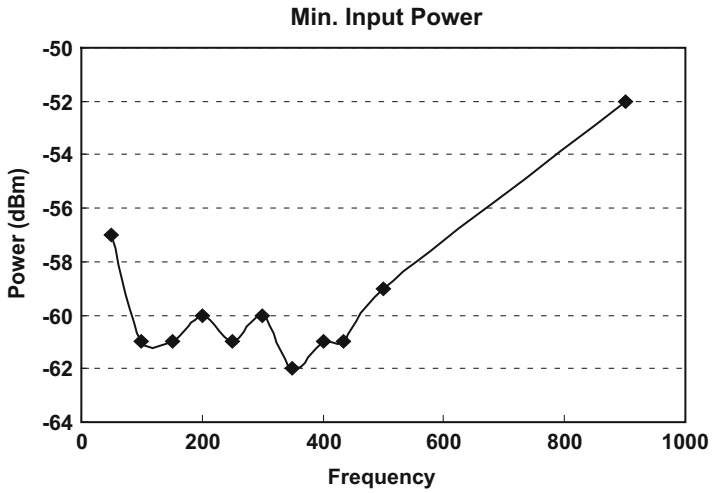


Fig. 11.38 Measured sensitivities of the OOK receiver at different frequencies [36]

5 kbps at 403 MHz, a frequency band complying with the Medical Implanted Communication System (MICS) standard for data transmission and reception in the body. To initiate the drug delivery operation, an external OOK command signal in RS232 format is wirelessly transmitted to the SoC, then acquired and demodulated by the OOK receiver. The MCU decodes the demodulated signal from the receiver and activates the selected drug cell by applying current to its membrane through a switch. This leads to the rupture of membrane and the release of drugs.

11.2.4.4 Implementation Results

Figure 11.39 shows the die photo of the drug delivery SoC, including eight individually addressable reservoirs. The volume of each reservoir is about 100 nL after the PDMS layer is bonded to the backside of the die. Figure 11.40 shows the images of the membrane before and after heating, which demonstrates the membrane addressed by the MCU ruptures after joule heating. Figure 11.41 shows the temporal response of the activation current injected into the membrane. The membrane ruptures in 50 ms after the activation current is applied. Detailed analysis shows that the temperature increase of the liquid drug is below 4 °C for a reservoir volume of 100 nL [36], which indicates a very low likelihood of deleterious thermal exposure to tissue or reservoir contents during membrane activation. To further confirm whether the concentration could be controlled by independent on-demand release of reservoir contents, an in vitro experiment of subsequent drug release is set up. The drug delivery system is immersed in the DI water, and blue dye is used as the reservoir content so that the concentration distribution of the released content can be observed and recorded by a microscope with a CCD video camera. Each membrane is then ruptured in sequence by

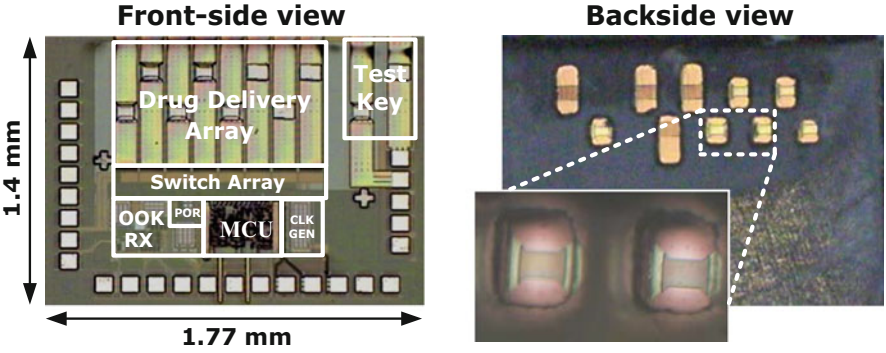


Fig. 11.39 Die photos of the drug delivery SoC [36]

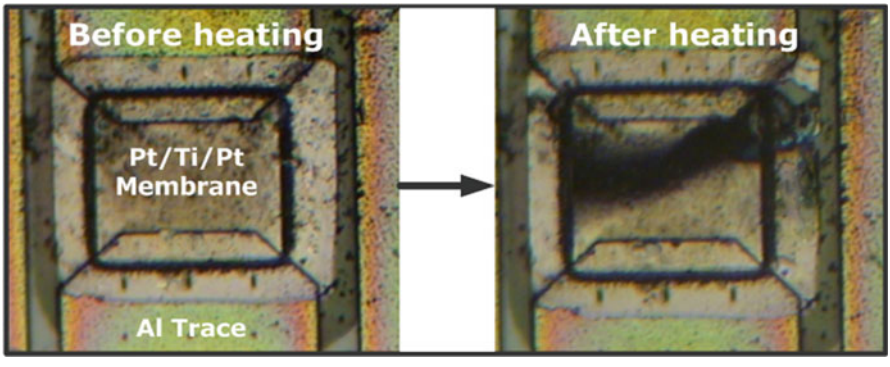


Fig. 11.40 Images of the membrane before and after heating [36]

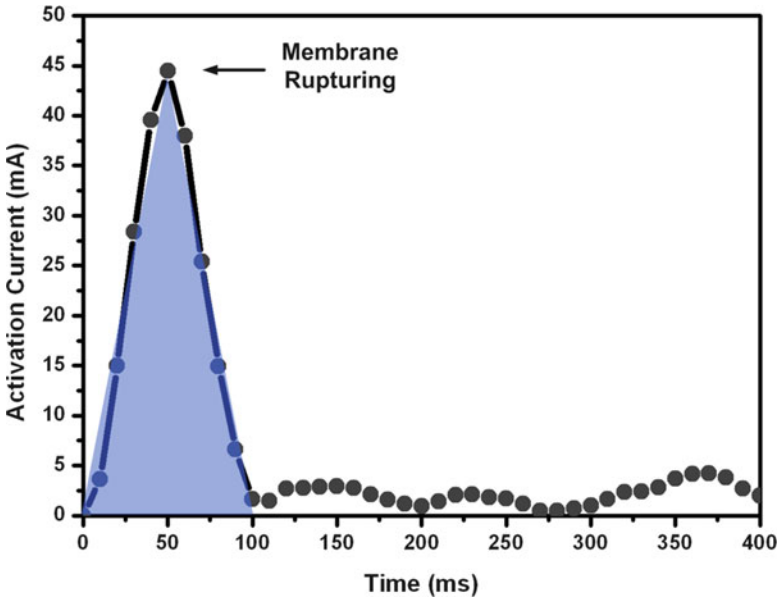


Fig. 11.41 Temporal response of the activation current [36]

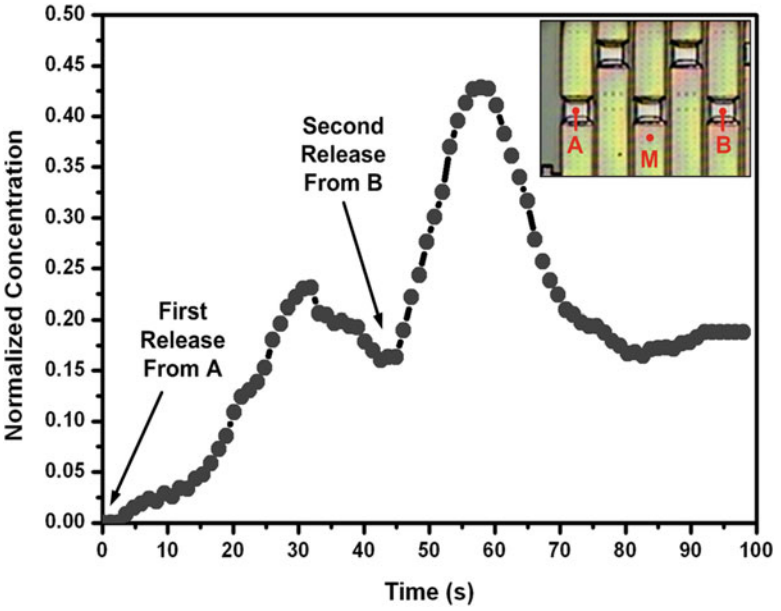


Fig. 11.42 Measurement results of normalized concentration as a function of time at spot M after the reservoir contents are released from the two ruptured openings, A and B, of drug reservoirs [36]

injecting the activation current. Figure 11.42 shows the measurement results of normalized concentration versus time at spot M illustrated in the inset figure after the reservoir contents are released from two ruptured openings of drug reservoirs,

A and B. We see that the concentration at M doubles after the adjacent reservoirs A and B are open. This clearly demonstrates that the implantable drug delivery SoC can be precisely controlled by the wireless commands, enabling future release-on-demand, localized therapies, and treatments for the next PC era.

11.2.5 Biomedical SoC Outside the Body: Implantable Pain-Control-on-Demand Batteryless Wireless CMOS SoC

Low back pain (LBP) is the fifth most common reason for physician visits in the USA [40, 41], while inflammation of the dorsal root ganglia (DRG) may cause approximately 40 % of LBP to be neuropathic pain [42–45]. Electrical stimulation to the central or peripheral neural conduction paths using pulsed radio frequency (PRF) pain therapy has been employed to relieve pain effectively while minimizing thermal damage. However, the pain relief after PRF therapy sustains only for a short period of time, i.e., about 3–6 months on average [46–48]. For offering continuous pain relief without repeated surgical procedures, a non-destructive and batteryless method using PRF for pain control is essential. Therefore, in this section, we present an implantable batteryless SoC that uses low-voltage PRF stimulation to avoid thermal damage while offering effective pain control on demand for the next PC era [49].

11.2.5.1 System Architecture

Figure 11.43 shows the application of an implantable pain-control-on-demand batteryless SoC. A patient with LBP can send a command from an external handheld device to the implanted SoC. The SoC will trigger the electrode stimulator to generate low-voltage PRF stimulation, providing an effective self-controlled analgesia anytime and anywhere. Figure 11.44 shows the block diagram of the proposed DRG stimulation system using an implantable batteryless SoC for pain control. The SoC consists of a radio-frequency-to-direct-current (RF-DC) circuit, a voltage limiter, voltage regulators, an RF receiver, a clock regenerator, a logic controller, and a PRF driver. The RF-DC circuit receives power from an external power source located outside the skin, and converts the RF signal RF_{in} into a DC voltage V_{DDr}. A low-frequency (1 MHz) spiral antenna is used for easy alignment and increased penetration depth. The following voltage limiter and regulator circuits together generate regulated supplies DV_{DD} and AV_{DD} for digital and analog circuits, respectively. The clock regenerator extracts the clock signal from the RF source to generate a l-MHz system clock. The PRF generator of the logic controller generates default bi-phasic PRF waveforms for the PRF drivers. The bi-phasic outputs are delivered to a pair of bi-polar electrodes placed into the surgically exposed L5 nerve of the lumbar region for stimulus in animal studies. The OOK

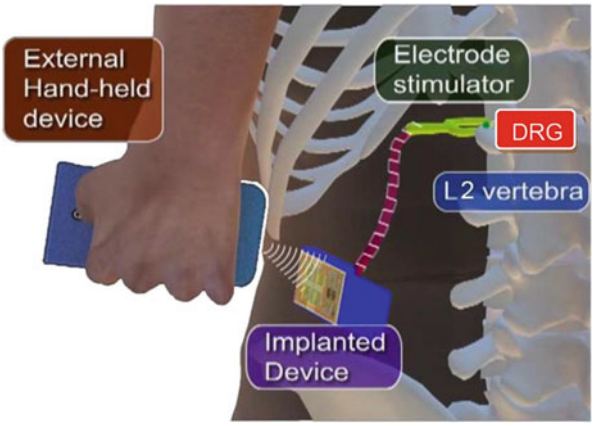


Fig. 11.43 DRG simulation for pain relief [49]

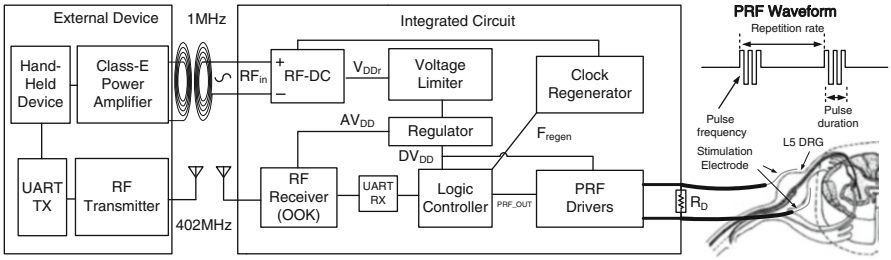


Fig. 11.44 Block diagram of the proposed DRG stimulator system for pain control [49]

receiver in [50, 51] is employed to acquire 402-MHz command signals complying with the MICS standard from an external handheld device, and then to direct the logic controller to output the specified PRF waveform.

11.2.5.2 Circuit Implementation

The RF-DC full-wave rectifier comprises four diode-connected MOS transistors. To avoid a reverse recovery current that causes additional power loss and power supply noise, the NMOS transistors are connected to ground via a substrate resistor R_{Sub}, and the bodies of the PMOS transistors are weakly connected to V_{DDr} by an additional resistor R_{bp}. Figure 11.45 shows the measured rectified V_{DDr} with respect to the inductive contact alignment when the VDD of the Class-E power amplifier in the external power source is set to 6 V. The minimal required V_{DDr} of the SoC is 2.2 V, corresponding to the maximal 18-mm gap distance. This distance is sufficient for the operation of the proposed batteryless wireless SoC implant under the skin. Figure 11.45 also shows that the maximal alignment offset of the antenna center is approximately 8 mm with a 10-mm gap distance.

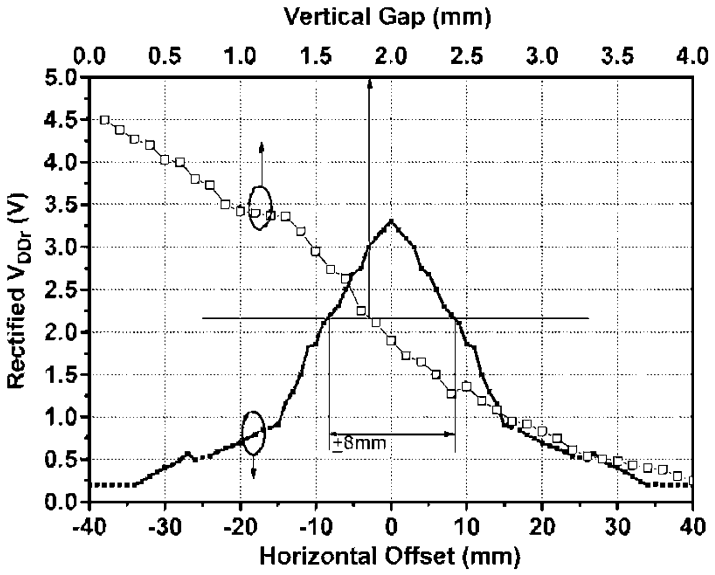


Fig. 11.45 Measured rectified V_{DDr} with respect to the vertical and horizontal offsets [49]

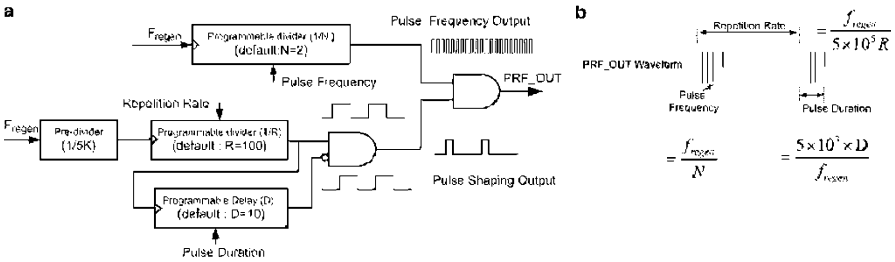


Fig. 11.46 (**a**) Block diagram of the PRF generator. (**b**) PRF waveform and parameter definitions [49]

The following voltage limiter composed of five serial connection diodes and a 200-kΩ resistor prevents the DC voltage from exceeding 5 V. The digital and analog regulators share the same bandgap reference [46]. A Schmitt trigger is used to realize the clock regenerator for better noise immunity and noise margin. A patient can specify a custom stimulation protocol by wirelessly issuing PRF parameters, such as pulse frequency and repetition rate, via an external handheld device. Figure 11.46 shows the block diagram of the PRF generator of the logic controller as well as the PRF waveform and parameter definitions. All three PRF parameters are programmable and can be reset by the OOK receiver.

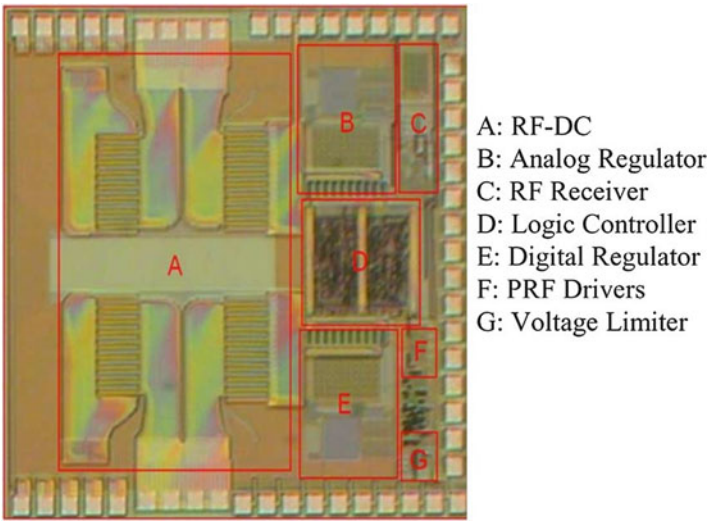


Fig. 11.47 Die photo of the pain-control-on-demand CMOS SoC [49]

11.2.5.3 Experimental Results

Figure 11.47 shows the die photo of the pain-control-on-demand wireless CMOS SoC. The RF-DC rectifier circuit occupies more than half of the chip area to maximize the efficiency and reduce the thermal effect. Figure 11.48 shows the DRG stimulator prototype powered by the external power source. The SoC loaded by a 10 kΩ resistor dissipates 12.48 mW with a chip temperature below 39 °C, as shown in the measured infrared (IR) thermography when activated. The packaged SoC chip was implanted into rats for the animal study. Before the implantation, the L5 nerve of the lumbar region was exposed to induce neuropathic pain by ligation. The bi-polar electrodes were then penetrated into the transverse process and placed beside DRG, as shown in Fig. 11.49. Once the external power source is close to the rat, the LED inside the rat lights up to demonstrate successful power delivery from the external power source to the SoC. The rats were grouped into the control group, two rats for which PRF was not applied, and the experimental group, four rats for which low-voltage PRF stimulation was applied for a 5-min duration. Von Frey (VF) monofilaments with different bending forces were utilized to stimulate the plantar surface of the foot to test mechanical allodynia. A higher VF score indicates high pain tolerance. All animals were tested before the surgery to collect their baseline values and were allowed to recover from surgical trauma before resuming the test on days 1, 2, 3, 5, and 7 to evaluate the VF score of both groups, as shown in Fig. 11.50. The experimental group with PRF stimulation had consistently higher pain tolerance than the control group without PRF stimulation. This clearly

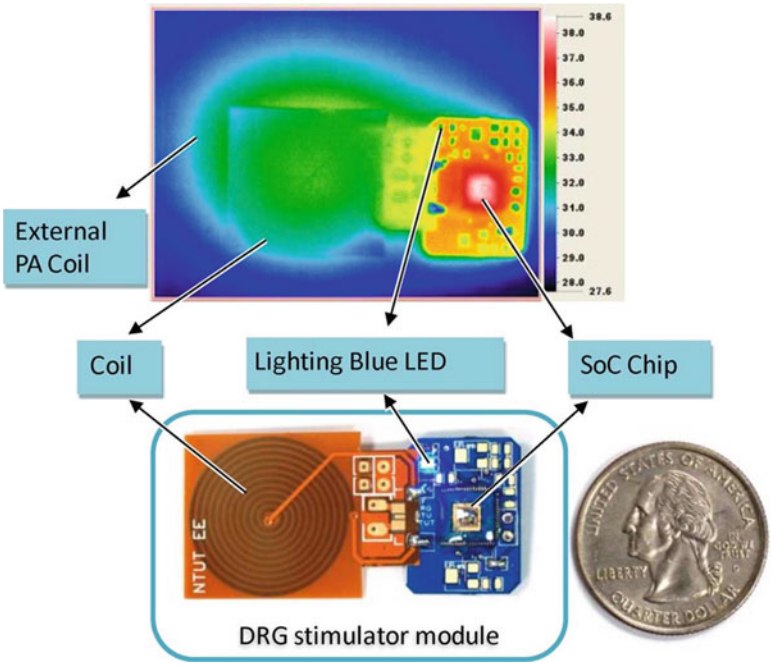


Fig. 11.48 DRG stimulator prototype and its measured IR thermography when activated [49]

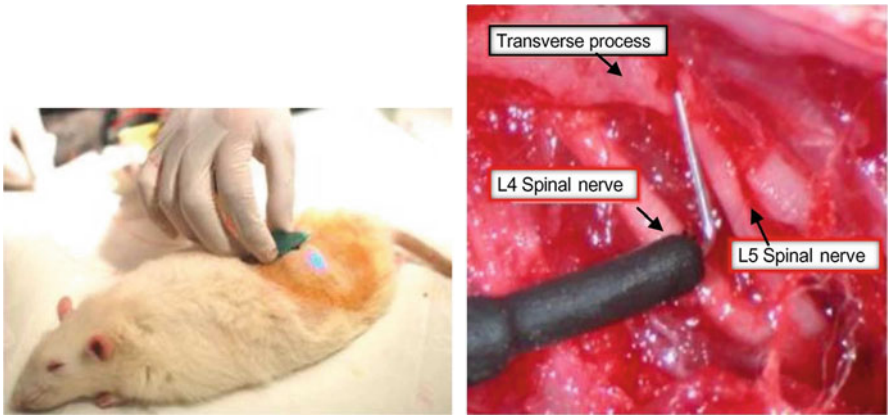


Fig. 11.49 Demonstration of the PRF treatment and placement of stimulation electrodes on the L5 spinal nerve [49]

demonstrates that the implantable batteryless wireless CMOS SoC using PRF can effectively reduce the sensation of pain on the DRG, offering a promising pain-control-on-demand solution for the next PC era.

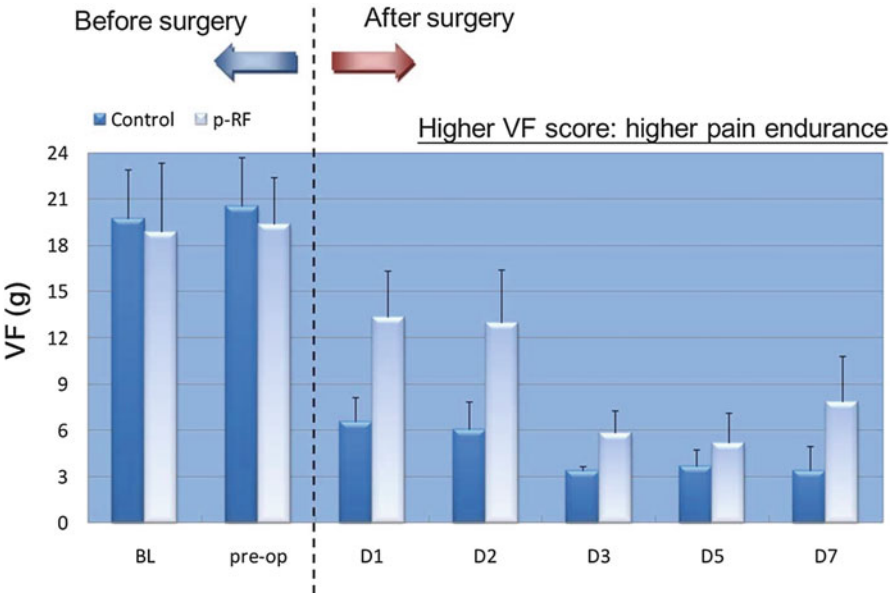


Fig. 11.50 Experimental results of the VF score variation before and after PRF stimulation to DRG [49]

11.2.6 Biomedical SoC Inside the Body: Implantable Batteryless Remotely Controlled Locomotive SoC

In the previous two sections, two implantable biomedical integrated SoC solutions for therapeutic and treatment applications for the next PC era were introduced. An implementable medical device with a controllable motion movement can potentially perform even advanced diagnosis and treatment inside the human body, such as tumor scan, drug delivery, and neuron stimulation, in a revolutionary way with minimum injury. Therefore, in this section, an implantable batteryless remotely controlled locomotive SoC that uses electrolytic bubbles as the propulsion mechanism is presented. Compared to the solution in the previous study [52], the solution introduced in this section is highly integrated without bulky external components such as magnets and on-board coils, and is capable of moving in four orthogonal directions with two speed controls [53].

11.2.6.1 System Architecture

Figure 11.51 shows the operation of the remotely controlled locomotive SoC. The process of electrolyzing water on chip generates micro-bubbles, which are adopted as the force to propel the chip. A user can control the micro-bubble emissions as well as the direction of motion by remotely commanding the SoC to appoint the

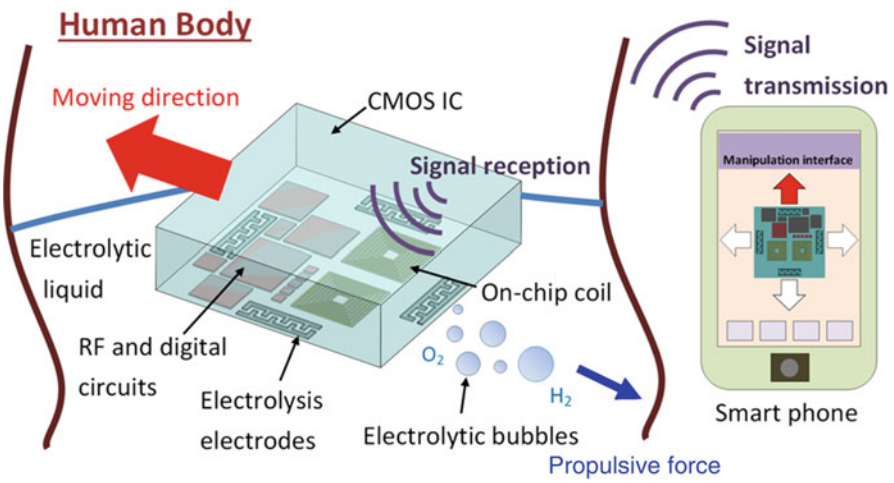


Fig. 11.51 Operation of the proposed remotely controlled locomotive SoC [53]

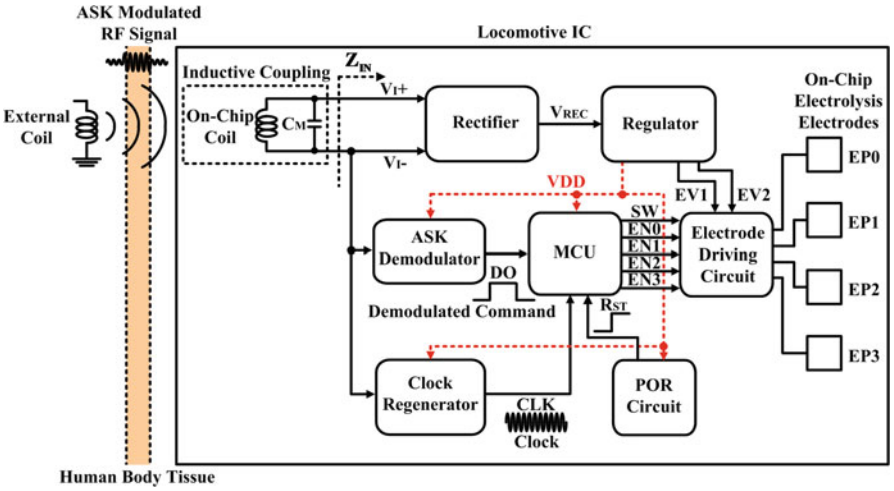


Fig. 11.52 Block diagram of the proposed CMOS locomotive SoC [53]

voltages to the corresponding on-chip electrolysis electrodes. The generated micro-bubbles composed of oxygen and hydrogen can be absorbed by the human body.

Figure 11.52 shows the block diagram of the CMOS locomotive SoC, which consists of an on-chip coil, a rectifier, a regulator, an amplitude-shift keying (ASK) demodulator, a clock generator, an MCU, a power-on-reset circuit, an electrode driving circuit, and on-chip electrolysis electrodes. An external 10-MHz RF signal with a 1-Mb/s ASK modulated command is inductively coupled from a transmitter to the chip through the on-chip coil. In addition to serving as locomotive control command, the received ASK signal is used for wireless powering and on-chip clock

generation. After rectification and regulation, the received signal is converted to the global supply voltage VDD (2 V) for the system. The electrode driving circuit is connected with the two voltages provided from the regulator outputs, EV1 (1.3 V) and EV2 (2 V), to generate bubbles at two different rates for two speed controls. The MCU controls the chip movement direction and adjusts the speed based on the demodulated command DO.

11.2.6.2 Circuit Implementation

Since the bubble generation rate is proportional to the applied voltage and the reaction area [54], four electrolysis electrodes with high-density interdigitated electrode pattern for high bubble generation rate are integrated on the four sides of the SoC, as shown in Fig. 11.51. Figure 11.53 shows the measured transient response of generated bubble volume and the corresponding recorded images. The linear increase in the volume of generated bubbles with time indicates a nearly constant bubble generation rate under a fixed electrolytic voltage.

The full-wave rectifier employing low-power bridge topology [55] converts the alternating ASK modulated RF signal to a DC voltage V_{REC}. The regulator consists of a start-up circuit, a bandgap reference, and an error amplifier [56]. The regulator provides the system supply VDD and two switchable electrolysis voltage EV1 and EV2 by resistor voltage divider circuits. The ASK demodulator is composed of an envelope detector, an LPF, and cascaded limiting amplifier [57]. Figure 11.54

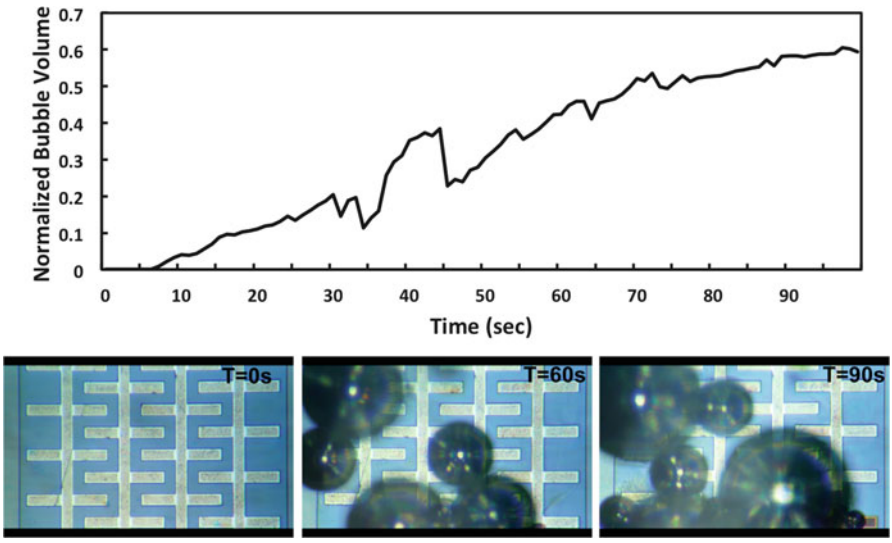


Fig. 11.53 Measured transient response of the generated bubble volume and the corresponding recoded images [53]

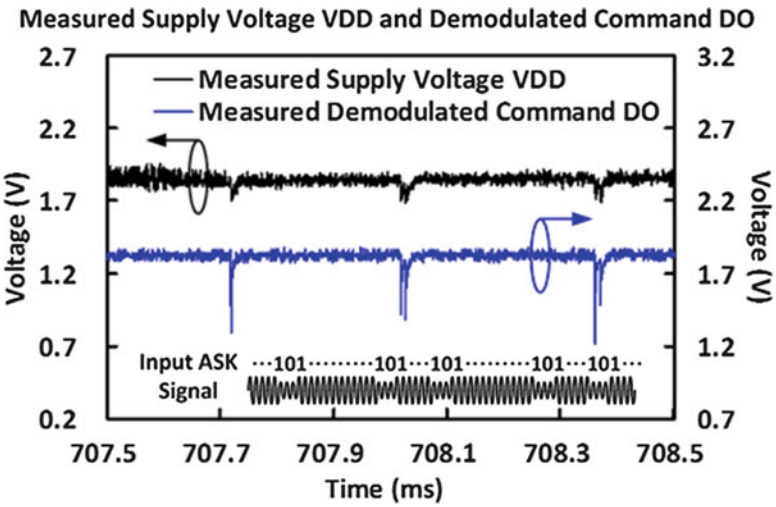


Fig. 11.54 Measured system supply voltage VDD and the demodulated command DO with the received ASK modulated RF signal inputs [53]

shows the measured system supply voltage VDD and the demodulated command DO with the received ASK modulated RF signal inputs.

The clock generator employing a four-stage cascaded self-biasing amplifier converts the ASK modulated signal to the full-swing clock signal CLK for the MCU. The electrode driving circuit consists of five analog switches and four analog buffers. The transmission-gate-based analog switch is controlled by the MCU to switch currents to the desired electrodes. Additional analog buffer comprising operational-amplifier-based unit-gain-buffer is employed to mitigate the loading effect due to human-body-fluid impedance variation with the concentration change of electrolyzed ions.

11.2.6.3 Experimental Results

Figure 11.55 shows the die photo of the locomotive SoC with on-chip coils and electrolysis electrodes. In order to verify the function of the proposed locomotive SoC, a bare chip is placed on the electrolyte surface, powered and controlled wirelessly to move in one direction by an ASK transmitter, as shown in Fig. 11.56. A saline solution with similar impedance to human-body fluid is used as the electrolyte. Substrate polishing is also used to reduce the chip mass for higher moving speed. The trajectory of the chip movement can be observed in Fig. 11.56. The locomotive SoC can move about 1.8 cm in 60 s, equivalent to a velocity of 0.3 mm/s. Figure 11.57 shows another experiment when a command is issued to change the direction during the chip movement. We can see that the locomotive SoC changes moving direction at $T = 10$ s after receiving a command of direction

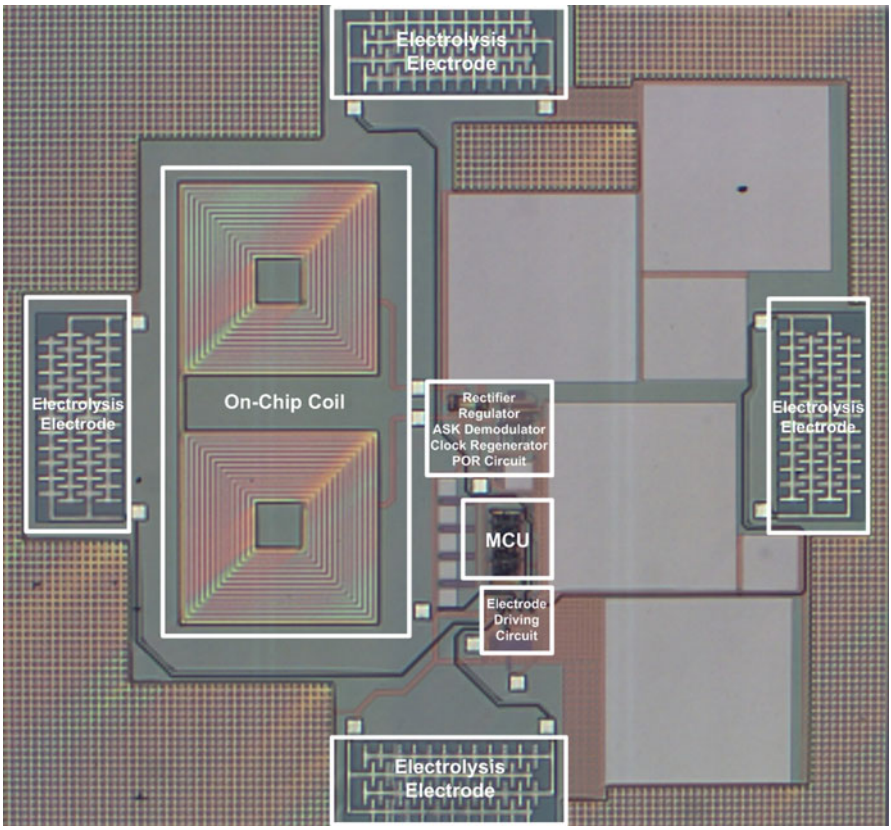


Fig. 11.55 Die photo of the proposed locomotive SoC [53]

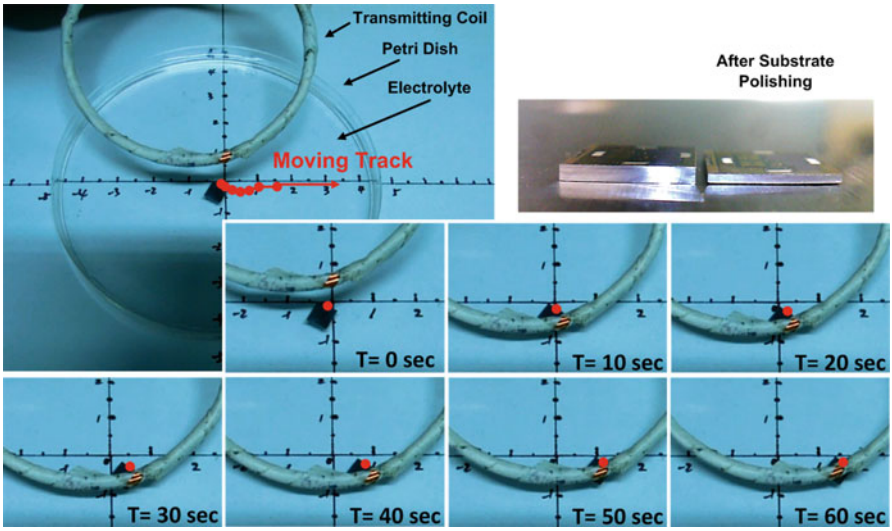


Fig. 11.56 Pictures of the chip movement with *red dots* marking the cruising track [53]

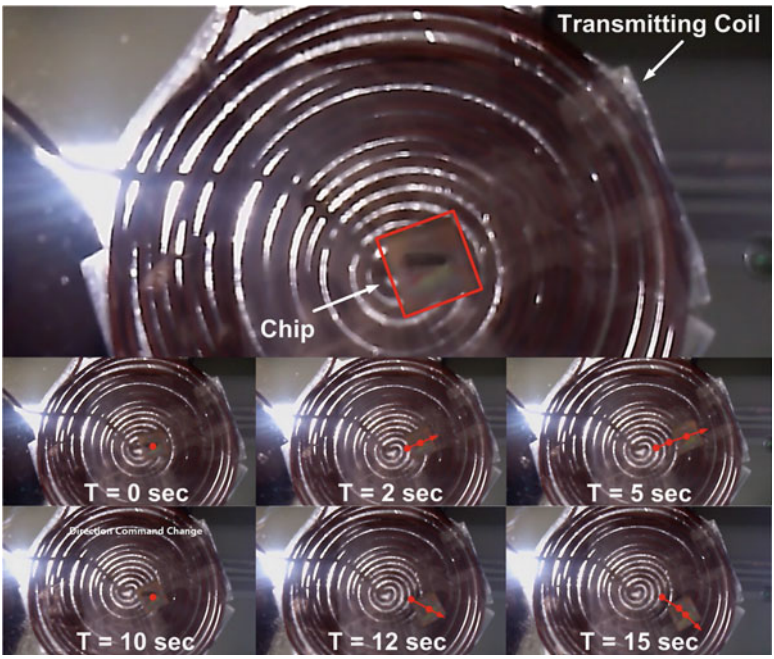


Fig. 11.57 Pictures of the chip movement with *red dots* marking the cruising track. The chip changes moving direction at T = 10 s after receiving a command of direction change [53]

change. The experimental results clearly demonstrate that the implantable batteryless remotely controlled locomotive SoC can help realize revolutionary diagnosis and treatment applications inside the human body in the next PC era.

11.2.7 Energy-Efficient Biomedical-Signal-Processing SoC

In the previous sections, the acquired physiological data or biomarker information from the SoC is directly transmitted to a remote computation platform for further processing and offline analysis. In other words, there is no biomedical signal processing involved on chip. However, sometimes on-chip biomedical processing is mainly necessary for the following two reasons:

1. Power and performance constraints: The bandwidth of the wireless data link might not be capable of supporting the direct transmission of rich biomedical data without consuming a huge amount of power. Additional on-chip biomedical processing can help reduce the data rate and minimize the overall power.
2. Application requirements: Some biomedical applications such as in the detection of sudden cardiac arrest (SCA) and seizures demand real-time responses to provide timely warning. Offline remote biomedical signal processing would not meet the latency requirement of these applications.

As a result, several energy-efficient SoC solutions [58–61] that can perform ECG, electroencephalogram (EEG), and electrocorticogram (ECoG) biomedical signal processing, neural spike-sorting functions in real time will be presented for the next PC era.

11.2.7.1 On-Chip ECG Signal Processing

In the USA, more than 5000 people experience SCA every week [58]. The survival rate of SCA victims drops by 10 % per minute without defibrillation, and more than 95 % of SCA victims die [62]. As a result, it is essential to provide timely warnings against fatal vascular signs. Figure 11.58 shows a heterogeneous ECG processor that can extract the abnormal ECG characteristics for ventricular fibrillation (VF), ventricular tachycardia (VT), and premature ventricular contraction (PVC) [58]. The heterogeneous ECG processor consists of an application processor (ASP) and an OpenRISC, a 32-bit general-purpose processor (GPP). The ASP first extracts the chaotic phase space differential (CPSD) [63] value from the raw ECG values. The GPP then decides whether there is a fatal sign based on the extracted CPSD values.

Figure 11.59 illustrates the CPSD algorithm [63] that has been developed to continuously detect critical cardiac conditions based on the time-delayed phase-space reconstruction method. A typical VF ECG signal spreads out on the phase space, while a normal ECG signal does not. Figure 11.59g shows the corresponding CPSD value and the VF threshold. The block diagram of the integrated ASP for CPSD acceleration is shown in Fig. 11.60. The ASP consists of four processing pipelines. In the first pipeline, a filter unit removes the low-frequency drifting voltage and 60-Hz power line noise. A phase space matrix (PM) constructor scans the filtered data, extracts the phase vectors, and then constructs the corresponding PM in the second pipeline. Two PMs are compared and their differences are accumulated in the third pipeline, while the CPSD value is calculated with the latest PM difference in the last pipeline.

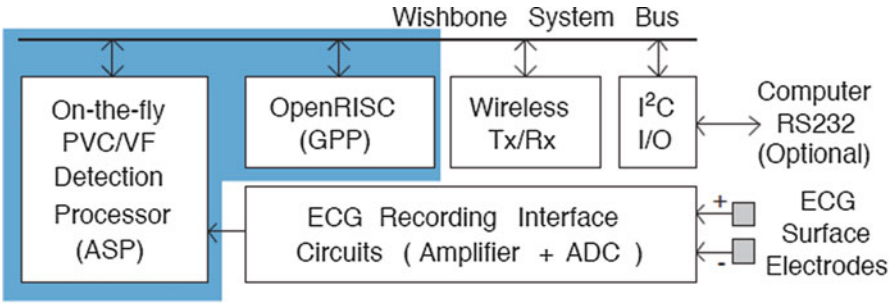


Fig. 11.58 Block diagram of the heterogeneous ECG processor [58]

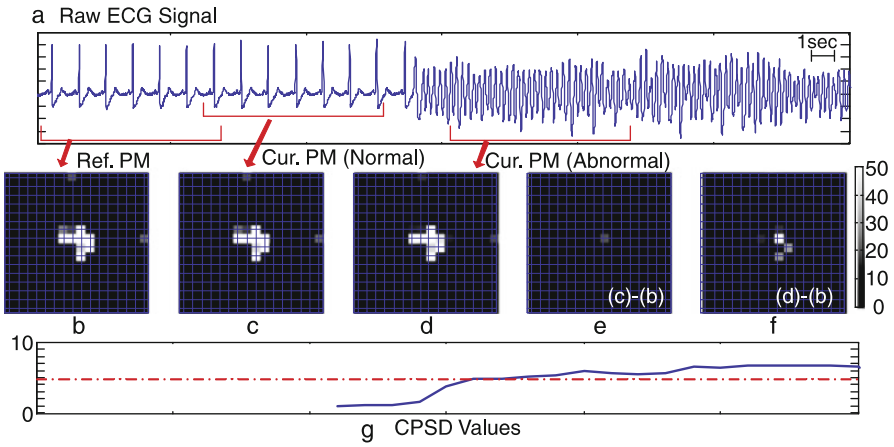


Fig. 11.59 CPSD algorithm. (**a**) Normal and abnormal ECG signals; (**b**) Phase space matrix (PM) constructed from normal ECG signal in the training phase as a reference during testing; (**c**, **d**) PMs constructed from normal and abnormal ECG signals during testing, respectively; (**e**, **f**) Difference in PMs of (**c**, **b**) as well as (**d**, **b**), respectively; (**g**) Corresponding CPSD value and the threshold for VF [58]

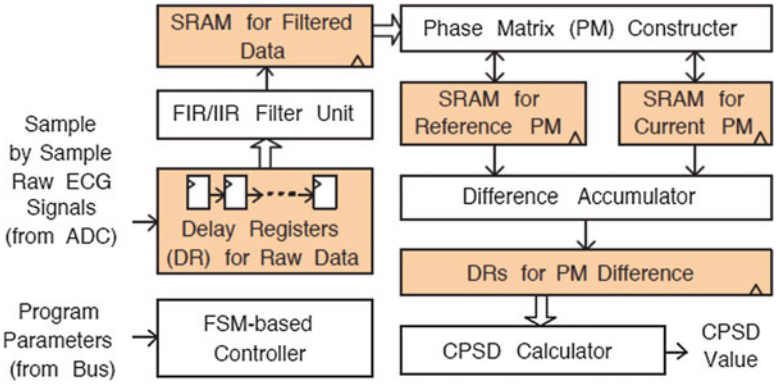


Fig. 11.60 Block diagram of the ASP for CPSD acceleration [58]

Figure 11.61 shows the die photo of the heterogeneous ECG processor. The processor can realize 98 % reduction in wireless transmission power by performing on-chip ECG signal processing. The heterogeneous architecture using both ASP and GPP can achieve 99 % power reduction compared with the one based on only GPP. The implementation results demonstrate that the on-chip ECG signal processing can provide real-time monitoring of heart conditions and give a timely warning against the fatal vascular signs for the next PC era.

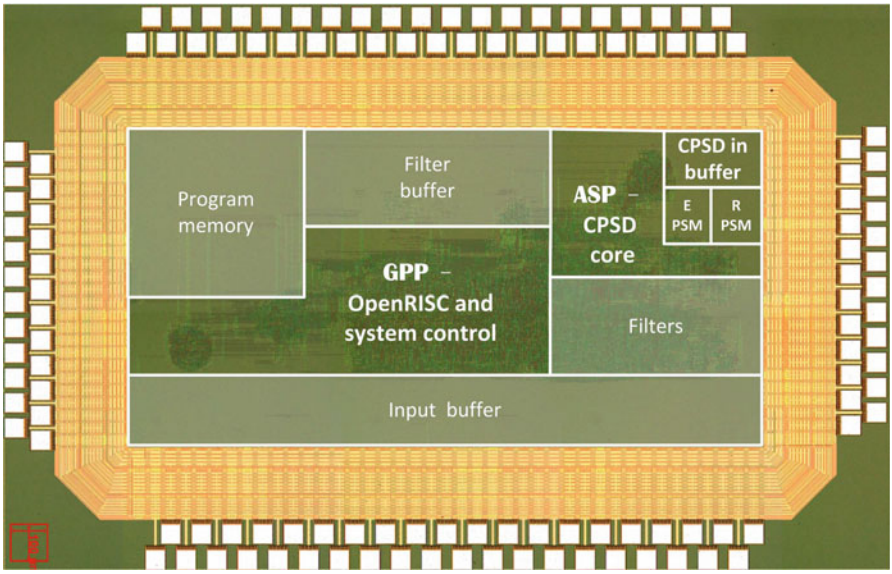


Fig. 11.61 Die photo of the heterogeneous ECG processor [58]

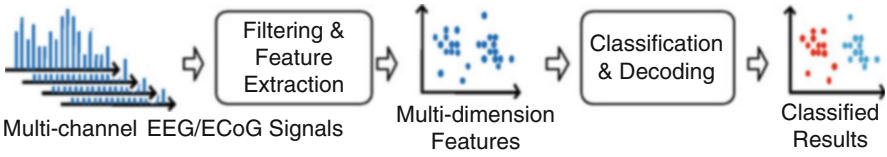


Fig. 11.62 Procedure of EEG/ECoG signal analysis [59]

11.2.7.2 On-Chip EEG/ECoG Signal Processing

On-line EEG/ECoG signal feature analysis can significantly enhance applications such as responsive neurostimulation (RNS) and brain–machine interface (BMI). Figure 11.62 shows the typical procedure of EEG/ECoG signal analysis involving pre-processing (PP) with filtering, feature extraction (FE), classification and decoding (CD). The corresponding EEG/ECoG signal processor consisting of three processing pipelines with heterogeneous hardware units is shown in Fig. 11.63 [59].

The PP pipeline performs both temporal and spatial linear filtering to remove the artifact and assemble the signal-of-interest. The PP architecture based on multiplication and accumulation (MAC) units cascaded with register array can significantly reduce memory-access power. In the FE stage, four types of features—temporal-domain characteristics, spatial-domain cross-channel correlations, frequency-domain spectrum features, and non-linear chaotic values—are extracted in parallel for 16 channels to increase both the accuracy and robustness of the processor.

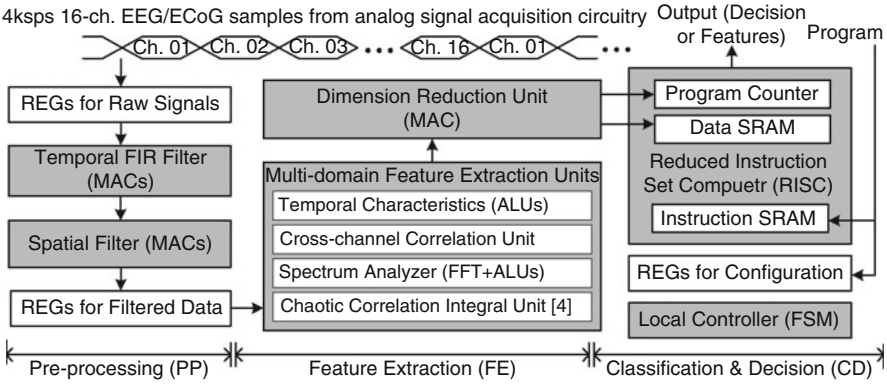


Fig. 11.63 Block diagram of the EEG/ECoG signal processor [59]

The dimension reduction unit using MAC then reduces the dimension of the feature space with the offline training algorithms. The FE architecture based on dedicated accelerators and reconfigurable logics provides both efficiency and flexibility. A reduced instruction set computer is employed on the CD stage to preserve the full programmability for the CD algorithm. The processor can output the decision every 0.1 s for real-time applications.

Figure 11.64 shows the die photo of the EEG/ECoG signal processor. Processing folding among 16 channels is utilized to further reduce the area and power. Figure 11.65 shows the experiment of early seizure detection using the EEG/ECoG signal processor. The processor analyzes the ECoG signals of a rat and detects the early stage of a chemically induced seizure. The decreased brain chaoticity, increased temporal energy, and oscillation frequency of the rhythmic discharge could be clearly observed sequentially on the feature space. The experimental results demonstrate that the on-chip EEG/ECoG signal processing can provide real-time monitoring of brain condition and early detection of seizure in the next PC era.

11.2.7.3 On-Chip Neural Spike Sorting

A closed-loop BMI system demands on-chip real-time neural signal processing for two reasons. First, since each implanted electrode may recode spike signals from multiple surrounded neurons, a real-time neural signal processing, spike sorting, is necessary to classify different spike signals based on their source neurons. Second, as mentioned before, real-time neural signal processing extracting only useful spike information can substantially reduce the communication data rate. Figure 11.66 shows the procedure of a typical neural spike-sorting process involving spike detection, spike alignment, and feature extraction and classification, which is similar to the procedure of EEG/ECoG signal analysis. A higher spike sampling rate (SR) usually leads to a better sorting performance, but consumes larger power.

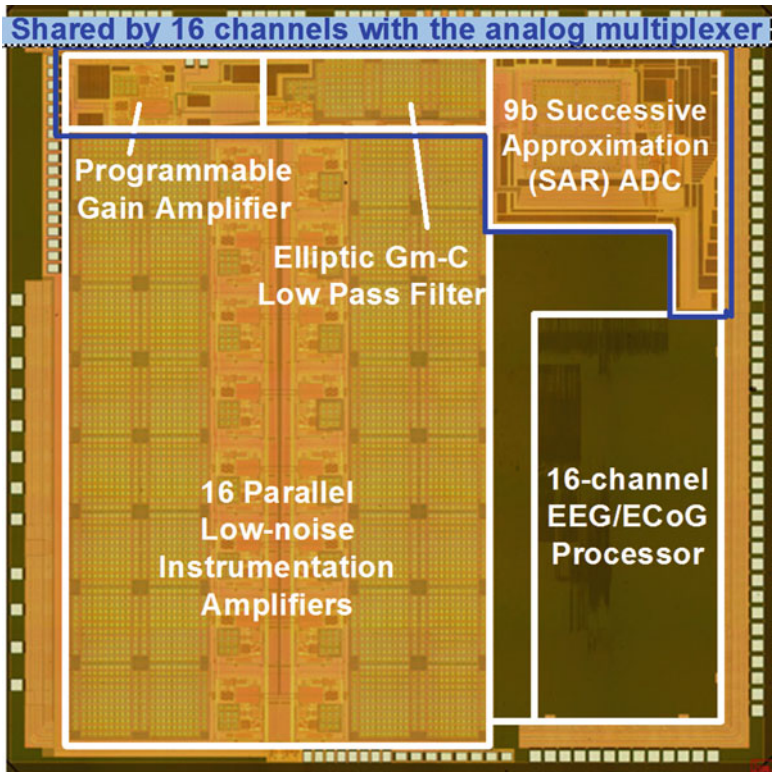


Fig. 11.64 Die photo of the EEG/ECoG signal processor [59]

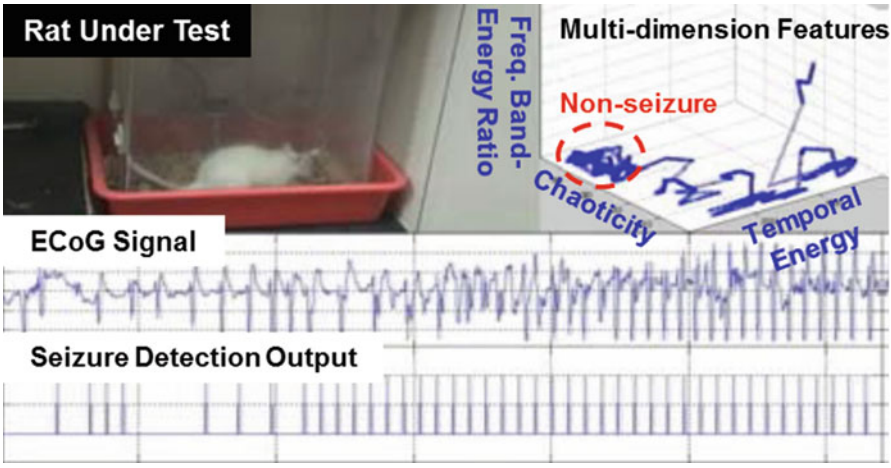


Fig. 11.65 Early detection of seizure using the EEG/ECoG signal processor [59]

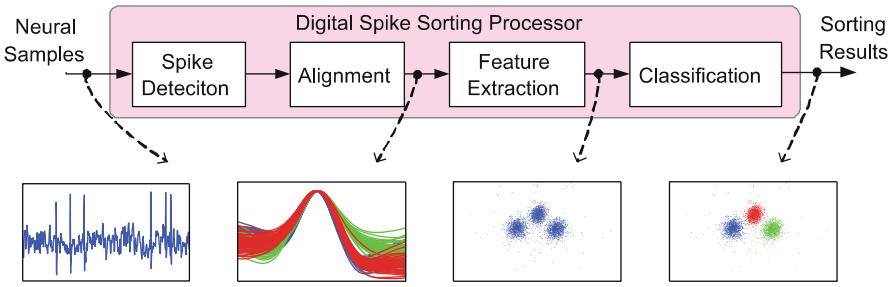


Fig. 11.66 Procedure of neural spike sorting [60]

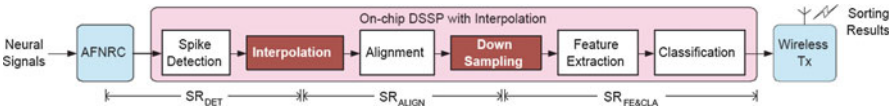


Fig. 11.67 On-chip neural spike-sorting processor with interpolation [60]

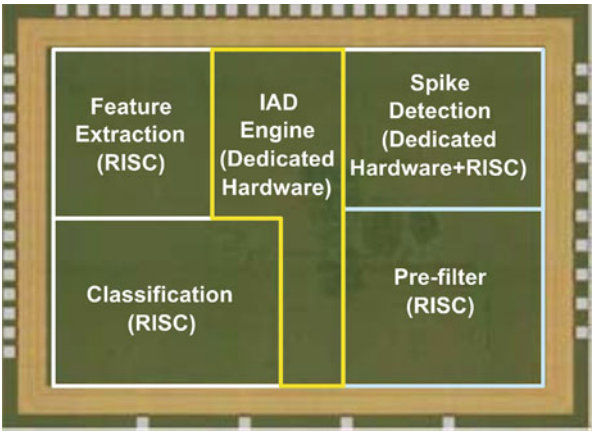


Fig. 11.68 Die photo of neural spike-sorting processor with interpolation [60]

Figure 11.67 shows an on-chip neural spike-sorting processor that can overcome this power–performance tradeoff by employing additional interpolation hardware [60]. The processor consists of three stages with different SRs: SR_{DET}, SR_{ALIGN}, and $SR_{FE\&CLA}$. Since the spike detection usually employs simple energy detection, a low SR (SR_{DET}) is sufficient in the first stage. To reduce the sampling skew and enhance the neuron separation accuracy, the second stage utilizes interpolation to align the spike with a higher SR (SR_{ALIGN}). Finally, the feature extraction and classification processing operate at a lower SR ($SR_{FE\&CLA}$) to minimize the power consumption after down-sampling. Figure 11.68 shows the die photo of a neural spike-sorting processor with interpolation. An IAD engine performs interpolation, alignment, and down-sampling, as shown in Fig. 11.67. Figure 11.69 shows the

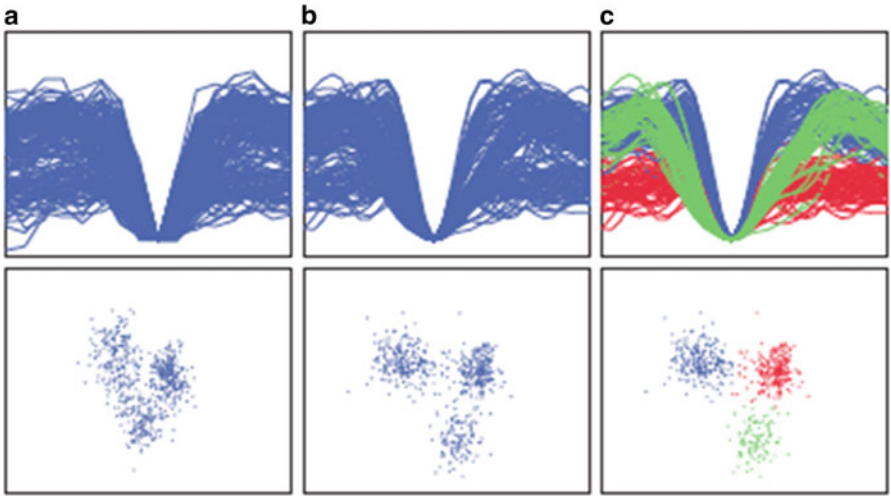


Fig. 11.69 Experimental results of spike-sorting processor (**a**) without interpolation and (**b**) with interpolation. (**c**) The results of (**b**) after K-means classification algorithm [60]

experimental results of the spike-sorting processor with the pre-recorded neural signals from rat hippocampus [64, 65]. It would be difficult to distinguish the boundary of different clusters of the detected spikes on the feature space owing to the low SR (12.5 ksps) and large sampling skew of the original data when the IAD engine is turned off, as shown in Fig. 11.69a. The boundary of the clusters becomes much more distinguishable after the IAD engine is turned on with $SR_{ALIGN} = 100$ ksps and $SR_{FE\&CLA} = 50$ ksps, as shown in Fig. 11.69b.

To further minimize the power consumption by the spike-sorting function for an implantable neural sensor, a robust and energy-efficient spike-sorting architecture exploiting an asynchronous timing strategy can be employed, as shown in Fig. 11.70 [61]. An asynchronous self-timed four-phase dual-rail handshaking protocol can reliably govern the communication between each module of the processor. As a result, each self-timed module can operate at its own speed reliably and achieve best-effort performance while minimizing the leakage. Figure 11.71 shows the die photos of synchronous and asynchronous spike-sorting processors. The implementation results show that the asynchronous spike-sorting processor realizes a 2.3 times reduction in power compared to the traditional synchronous approach. This demonstrates that the on-chip neural signal spike sorting can enable future closed-loop BMI systems with minimum energy and power overhead for the next PC era.

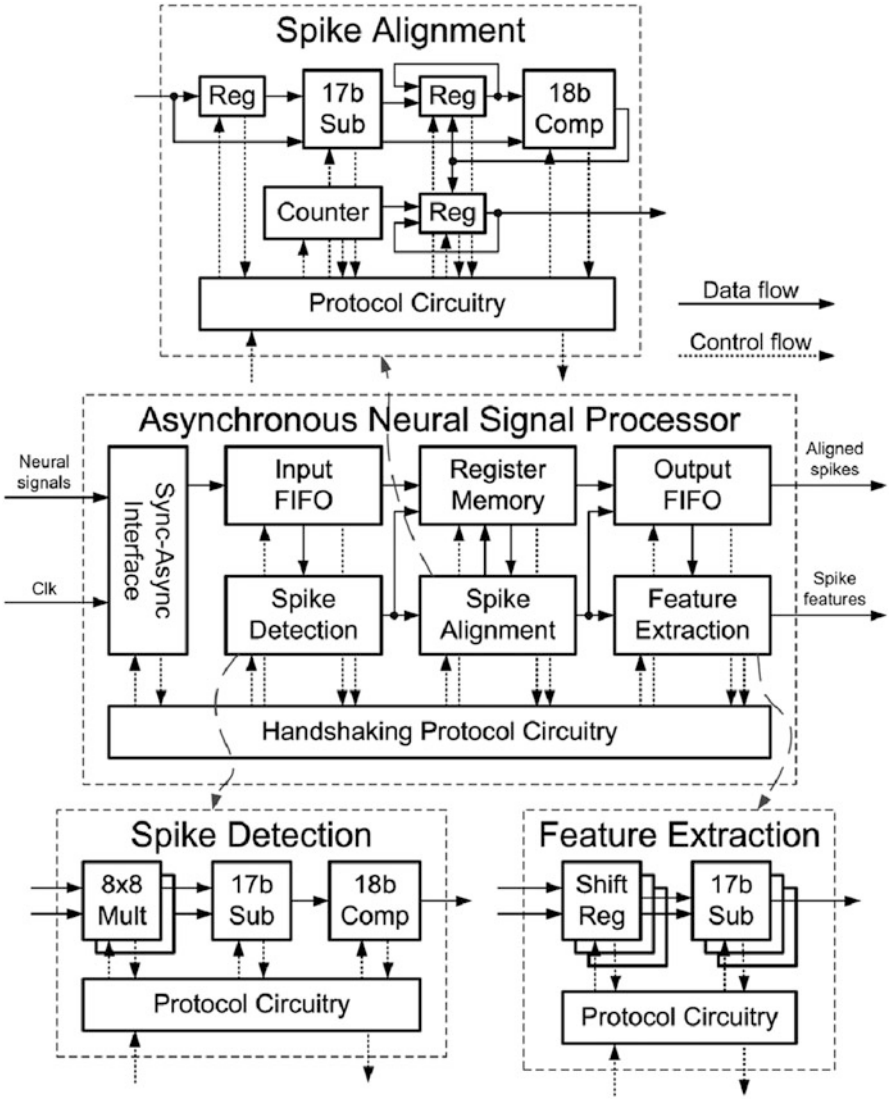


Fig. 11.70 Block diagram and circuit schematics of an asynchronous spike-sorting processor [61]

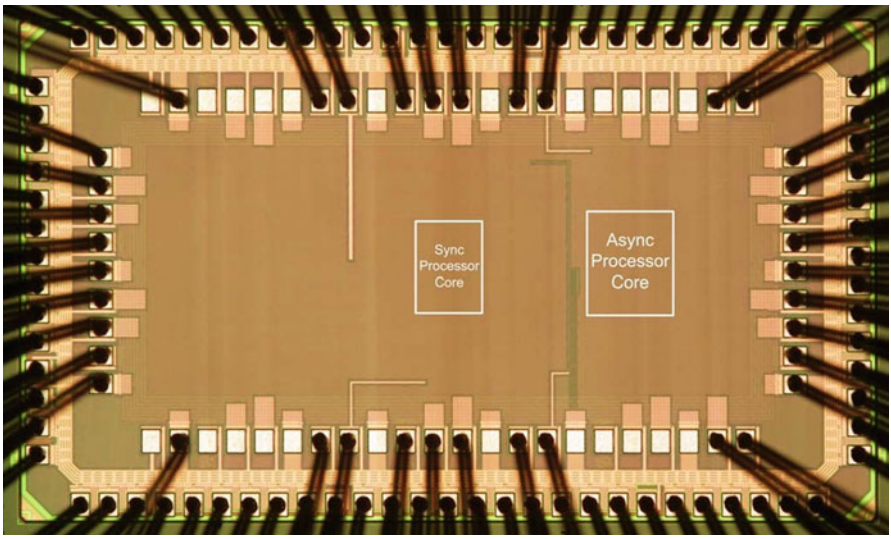


Fig. 11.71 Die photos of synchronous and asynchronous spike-sorting processors [61]

References

1. Wikipedia Contributors (2016) Patient protection and affordable care act. Wikipedia, The Free Encyclopedia, 20 Feb 2016, 15:13 UTC. https://en.wikipedia.org/w/index.php?title=Patient_Protection_and_Affordable_Care_Act&oldid=705954359. Accessed 25 Feb 2016
2. Wikipedia Contributors (2016) Health care in the United States. Wikipedia, The Free Encyclopedia, 24 Feb 2016, 13:01 UTC. https://en.wikipedia.org/w/index.php?title=Health_care_in_the_United_States&oldid=706640161. Accessed 25 Feb 2016
3. National health expenditure data: NHE fact sheet. Centers for Medicare and Medicaid Services. Accessed 26 Feb 2008
4. Keehan S, Sisko A, Truffer C, Smith S, Cowan C, Poisal J, Clemens MK, National Health Expenditure Accounts Projections Team (2008) Health spending projections through 2017: the baby-boom generation is coming to Medicare. Health Aff 27(2):w145–w155
5. http://www.forbes.com/sites/brucejapsen/2013/09/13/obamacare-will-boost-consumer-medical-device-market-to-10-billion/#2120bece2715
6. Thorpe KE, Howard DH (2006) The rise in spending among Medicare beneficiaries: the role of chronic disease prevalence and changes in treatment intensity. Health Aff 25(5):w378–w388
7. Huang C-R, Chang J-Y, Chiang C-L (2008) Telecare and telehealth care network in Taiwan. The 6th conference of the international society for gerontechnology (ISG08). Telemonitoring and telecare 3, Pisa, June 2008, p 1–5
8. Anker SD, Koehler F, Abraham WT (2011) Telemedicine and remote management of patients with heart failure. Lancet 378(9792):731–739
9. Chiang L-C, Chen W-C, Dai Y-T, Ho Y-L (2012) The effectiveness of telehealth care on caregiver burden, mastery of stress, and family function among family caregivers of heart failure patients: a quasi-experimental study. Int J Nurs Stud 49(10):1230–1242
10. Ho Y-L, Yu J-Y, Lin Y-H, Chen Y-H, Huang C-C, Hsu T-P, Chuang P-Y, Hung C-S, Chen M-F (2014) Assessment of the cost-effectiveness and clinical outcomes of a fourth-generation synchronous telehealth program for the management of chronic cardiovascular disease. J Med Internet Res 16(6), e145

11. Ho T-W, Huang C-W, Lin C-M, Lai F, Ding J-J, Ho Y-L, Hung C-S (2015) A telesurveillance system with automatic electrocardiogram interpretation based on support vector machine and rule-based processing. JMIR Med Inform 3(2), e21
12. Welch G, Garb J, Zagarins S, Lendel I, Gabbay RA (2010) Nurse diabetes case management interventions and blood glucose control: results of a meta-analysis. Diabetes Res Clin Pract 88 (1):1–6
13. Garcia-Aymerich J, Hernandez C, Alonso A, Casas A, Rodriguez-Roisin R, Anto JM, Roca J (2007) Effects of an integrated care intervention on risk factors of COPD readmission. Respir Med 101(7):1462–1469
14. Maron DJ, Boden WE, O'Rourke RA, Hartigan PM, Calfas KJ, Mancini GB, Spertus JA, Dada M, Kostuk WJ, Knudtson M, Harris CL, Sedlis SP, Zoble RG, Title LM, Gosselin G, Nawaz S, Gau GT, Blaustein AS, Bates ER, Shaw LJ, Berman DS, Chaitman BR, Weintraub WS, Teo KK, COURAGE Trial Research Group (2010) Intensive multifactorial intervention for stable coronary artery disease: optimal medical therapy in the COURAGE (Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation) trial. J Am Coll Cardiol 55(13):1348–1358
15. Clark RA, Inglis SC, McAlister FA, Cleland JG, Stewart S (2007) Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. BMJ 334(7600):942–945
16. Rosamond W, Flegal K, Friday G, Furie K, Go A, Greenlund K, Haase N, Ho M, Howard V, Kissela B, Kittner S, Lloyd-Jones D, McDermott M, Meigs J, Moy C, Nichol G, O'Donnell CJ, Roger V, Rumsfeld J, Sorlie P, Steinberger J, Thom T, Wasserthiel-Smoller S, Hong Y, American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2007) Heart disease and stroke statistics—2007 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 115(5):e69–e171
17. Chen Y-H, Lin Y-H, Hung C-S, Huang C-C, Yeih D-F, Chuang P-Y, Ho Y-L, Chen M-F (2013) Clinical outcome and cost-effectiveness of a synchronous telehealth service for seniors and nonseniors with cardiovascular diseases: quasi-experimental study. J Med Internet Res 15 (4), e87
18. Davidson PM, Dracup K, Phillips J, Padilla G, Daly J (2007) Maintaining hope in transition: a theoretical framework to guide interventions for people with heart failure. J Cardiovasc Nurs 22(1):58–64
19. Riegel B, Dickson VV (2010) Self-care of heart failure: a situation-specific theory of health transition. In: Meleis AI (ed) Transitions theory: middle-range and situation-specific theories in nursing research and practice. Springer, Heidelberg
20. Kuo P-H, Kuo J-C, Hsueh H-T, Hsieh J-Y, Huang Y-C, Wang T, Lin Y-H, Lin C-T, Yang Y-J, Lu S-S (2015) A smart CMOS assay SoC for rapid blood screening test of risk prediction. Solid-state circuits conference (ISSCC), 2015 I.E. International, San Francisco, CA, 2015, p 1–3
21. Manickam A, Chevalier A, McDermott M, Ellington AD, Hassibi A (2010) A CMOS electrochemical impedance spectroscopy biosensor array for label-free biomolecular detection. ISSCC Dig Tech Papers, Feb 2010, p 130–131
22. Gambini S, Skucha K, Liu P, Jungkyu K, Krigel R, Mathies R, Boser B (2012) A CMOS 10k pixel baseline-free magnetic bead detector with column-parallel readout for miniaturized immunoassays. ISSCC Dig Tech Papers, Feb 2012, p 126–128
23. Kuo J-C, Kuo P-H, Hsueh H-T, Ma C-W, Lin C-T, Lu S-S, Yang Y-J (2014) A capacitive immunosensor using on-chip electrolytic pumping and magnetic washing techniques for point-of-care applications. 2014 I.E. 27th international conference on micro electro mechanical systems (MEMS), San Francisco, CA, p 809–12
24. Kuo P-H, Hsieh J-Y, Huang Y-C, Huang Y-J, Tsai R-D, Wang T, Chiu H-W, Lu S-S (2014) A remotely controlled locomotive IC driven by electrolytic bubbles and wireless powering. ISSCC Dig Tech Papers, Feb 2014, p 322–323

25. Liu PP, Skucha K, Duan Y, Megens M, Jungkyu K, Izyumin II, Gambini S, Boser B (2012) Magnetic relaxation detector for microbead labels. IEEE J Solid State Circuits 47 (4):1056–1064
26. Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds) (2015) SEER cancer statistics review, 1975-2012. National Cancer Institute, Bethesda, MD. http://seer.cancer.gov/csr/1975_2012/, based on November 2014 SEER data submission, posted to the SEER website, Apr 2015
27. Lung cancer fact sheet (2016) http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/learn-about-lung-cancer/lung-cancer-fact-sheet.html
28. Mazzone PJ (2012) Exhaled breath volatile organic compound biomarkers in lung cancer. J Breath Res 6(2):1–8
29. Tzeng T-H, Kuo C-Y, Wang S-Y, Huang P-K, Huang Y-M, Hsieh W-C, Huang Y-J, Kuo P-H, Yu S-A, Lee S-C, Tseng YJ, Tian W-C, Lu S-S (2016) A portable micro gas chromatography system for lung cancer associated volatile organic compound detection. IEEE J Solid State Circuits 51(1):259–272
30. Chang C-Y, Kuo C-Y, Huang P-K, Tian W-C, Lu C-J (2013) Volatile organic compounds sensor with stacked interdigitated electrodes coated with monolayer-protected gold nanoclusters. 2013 transducers and Eurosensors XXVII: the 17th international conference on Solid-state sensors, actuators and microsystems (TRANSDUCERS and EUROSENSORS XXVII), Barcelona, p 1170–1173
31. Wang K-C, Wang S-Y, Kuo C-H, Tseng YJ (2013) Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. Anal Chem 85(2):1231–1239
32. Lin Y-T, Lin Y-S, Chen C-H, Chen H-C, Yang Y-C, Lu S-S (2011) A 0.5-V biomedical system-on-a-chip for intrabody communication system. IEEE Trans Ind Electron 58 (2):690–699
33. Ruiz JA, Xu J, Shimamoto S (2006) Propagation characteristics of intra-body communications for body area networks. 3rd IEEE consumer communications and networking conference 2006 (CCNC 2006), p 509–513
34. Hsiao S-Y, Wu C-Y (1998) A parallel structure for CMOS four-quadrant analog multipliers and its application to a 2-GHz RF downconversion mixer. IEEE J Solid State Circuits 33 (6):859–869
35. Smith S, Tang TB, Terry JG, Stevenson JT, Flynn BW, Reekie HM, Murray AF, Gundlach AM, Renshaw D, Dhillon B, Ohtori A, Inoue Y, Walton AJ (2007) Development of a miniaturised drug delivery system with wireless power transfer and communication. IET Nanobiotechnol 1(5):80–86
36. Huang Y-J, Liao H-H, Huang P-L, Wang T, Yang Y-J, Wang Y-H, Lu S-S (2012) An implantable release-on-demand CMOS drug delivery SoC using electrothermal activation technique. ACM J Emerg Technol Comput Syst 8(2):1–22
37. Chan CK, Peng H, Liu G, McIlwrath K, Zhang XF, Huggins RA, Cui Y (2008) High-performance lithium battery anodes using silicon nanowires. Nat Nanotechnol 3(1):31–35
38. Simons RN, Miranda FA, Wilson JD, Simons RE (2006) Wearable wireless telemetry system for implantable bio-MEMS sensors. Proceedings of the 28th IEEE EMBS annual international conference, New York, 30 Aug–3 Sept 2006
39. Wang T, Chen H-C, Chiu H-W, Lin Y-S, Huang GW, Lu S-S (2006) Micromachined CMOS LNA and VCO by CMOS-compatible ICP deep trench technology. IEEE Trans Microw Theory Tech 54(2):580–588
40. Deyo RA, Mirza SK, Martin BI (2006) Back pain prevalence and visit rates: estimates from U.S. national surveys, 2002. Spine (Phila Pa 1976) 31(23):2724–2727
41. Hart LG, Deyo RA, Cherkin DC (1995) Physician office visits for low back pain. Frequency, clinical evaluation, and treatment patterns from a U.S. national survey. Spine (Phila Pa 1976) 20(1):11–19

42. Bogduk N, McGuirk B (2012) Medical management of acute chronic low back pain: an evidence-based approach. Elsevier, Amsterdam
43. Zhang JM, Song XJ, LaMotte RH (1999) Enhanced excitability of sensory neurons in rats with cutaneous hyperalgesia produced by chronic compression of the dorsal root ganglion. J Neurophysiol 82(6):3359–3366
44. Zhang JM, Li H, Brull SJ (2000) Perfusion of the mechanically compressed lumbar ganglion with lidocaine reduces mechanical hyperalgesia and allodynia in the rat. J Neurophysiol 84 (2):798–805
45. Xiang Z, Xiong Y, Yan N, Li X, Mao Y, Ni X, He C, LaMotte RH, Burnstock G, Sun J (2008) Functional up-regulation of P2X 3 receptors in the chronically compressed dorsal root ganglion. Pain 140(1):23–34
46. Richebé P, Rathmell JP, Brennan TJ (2005) Immediate early genes after pulsed radiofrequency treatment: neurobiology in need of clinical trials. Anesthesiology 102(1):1–3
47. Simopoulos TT, Kraemer J, Nagda JV, Aner M, Bajwa ZH (2008) Response to pulsed and continuous radiofrequency lesioning of the dorsal root ganglion and segmental nerves in patients with chronic lumbar radicular pain. Pain Physician 11(2):137–144
48. Lin M-L, Chang C-H, Lin C-W, Chiu H-W, Wen Y-R, Lin S-H (2009) Implantable pulsed-RF on dorsal root ganglion for treatment of neuropathic pain—animal study. WIP, New York
49. 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 (2010) 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 4 (6):350–359
50. Chen C-H, Hwang, R-Z, Huang L-S, Lin S, Chen H-C, Yang Y-C, Lin Y-T, Yu S-A, Wang Y-H, Chou N-K, Lu S-S (2006) A wireless bio-MEMS sensor for c-reactive protein detection based on nanomechanics. ISSCC Dig Tech Papers, Feb 2006, p 2298–2307
51. Chen C-H, Hwang R-Z, Huang L-S, Lin S, Chen H-C, Yang Y-C, Lin Y-T, Yu S-A, Wang Y-H, Chou N-K, Lu S-S (2009) A wireless bio-MEMS sensor for C-reactive protein detection based on nanomechanics. IEEE Trans Biomed Eng 56(2):462–470
52. Yakovlev A, Pivonka D, Meng T, Poon A (2012) A mm-sized wirelessly powered and remotely controlled locomotive implantable device. ISSCC Dig Tech Papers, Feb 2012, p 302–304
53. Hsieh J-Y, Kuo P-H, Huang Y-C, Huang Y-J, Tsai R-D, Wang T, Chiu H-W, Wang Y-H, Lu S-S (2014) A remotely-controlled locomotive IC driven by electrolytic bubbles and wireless powering. IEEE Trans Biomed Circuits Syst 8(6):787–798
54. Chan S-C, Chen C-R, Liu C-H (2010) A bubble-activated micropump with high-frequency flow reversal. Sensors Actuators A Phys 163(2):501–509
55. Sedra AS, Smith KC (2004) Microelectronic circuits, 5th edn. Oxford University Press, New York
56. Razavi B (2001) Design of analog CMOS integrated circuits. McGraw-Hill, New York
57. Huang P-L, Kuo P-H, Huang Y-J, Liao H-H, Yang JY, Wang T, Wang Y-H, Lu S-S (2012) A controlled-release drug delivery system on a chip using electrolysis. IEEE Trans Ind Electron 59(3):1578–1587
58. Chen H-H, Chiang C-Y, Chen T-C, Liu C-S, Huang Y-J, Lu S-S, Lin C-W, Chen L-G (2011) Analysis and design of on-sensor ECG processors for realtime detection of cardiac anomalies including VF, VT, and PVC. J Signal Process Syst 65(2):275–285
59. Chen T-C, Lee T-H, Chen Y-H, Ma T-C, Chuang T-D, Chou C-J, Yang C-H, Lin T-H, Chen L-G (2010) 1.4μW/channel 16-channel EEG/ECoG processor for smart brain sensor SoC. 2010 I.E. symp VLSI circuits dig, Aug–Sept 2010, p 21–22
60. Chen T-C, Ma T-C, Chen Y-Y, Chen L-G (2012) Low power and high accuracy spike sorting microprocessor with on-line interpolation and re-alignment in 90 nm CMOS process. Conf proc IEEE Eng Med Biol Soc, Aug–Sept 2012, p 4485–4488
61. Liu T-T, Rabaey JM (2013) A 0.25 V 460 nW asynchronous neural signal processor with inherent leakage suppression. IEEE J Solid State Circuits 48(4):897–906

62. de Vreede-Swagemakers JJ, Gorgels AP, Dubois-Arbouw WI, van Ree JW, Daemen MJ, Houben LG, Wellens HJ (1997) Out-of-hospital cardiac arrest in the 1990's: a population-based study in the Maastricht area on incidence, characteristics and survival. J Am Coll Cardiol 30(6):1500–1505
63. Liu C-S, Lin Y-C, Chuang Y-H, Hsiao T-C, Lin C-W (2009) Chaotic phase space differential (CPSD) algorithm for real-time detection of VF, VT, and PVC ECG signals. In: 4th European conference of the international federation for medical and biological engineering, vol. 22. Springer, Heidelberg, p 18–21
64. Henze DA, Harris KD, Borhegyi Z, Csicsvari J, Mamiya A, Hirase H, Sirota A, Buzsáki G (2009) Simultaneous intracellular and extracellular recordings from hippocampus region CA1 of anesthetized rats. CRCNS.org. http://dx.doi.org/10.6080/K02Z13FP
65. Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsáki G (2000) Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. J Neurophysiol 84(1):401–414
66. Lin Y-L, Kyung C-M, Yasuura H, Liu Y (2015) Smart sensors and systems. Springer, Heidelberg

Chapter 12
Functional Nanofibers for Flexible Electronics

Suiyang Liao, Ya Huang, and Hui Wu

Abstract In the field of flexible electronics, elastomeric substrates take the place of their rigid counterparts and boost fascinating uses such as solar cells, electronic skins, and flexible displays. Nano-level dimensions lead to novel properties of materials, which is the case in nanofibers with a large aspect ratio and a high surface–volume ratio. Researchers are striving to apply nanofibers of polymers, metals, carbon materials, composites, and even ceramics to flexible electronics. Electrospinning, as a method to prepare nanofibers, is low cost and versatile, which are the two advantages for massive production. This review recapitulates the most recent progress of fabricating flexible electronics based on electrospun nanofibers and reveals the problematic issues faced by both electrospinning and flexible electronics.

Keywords Flexible electronics • Nanofibers • Optoelectronics • Smart sensor

12.1 Introduction

Modern society has been changed dramatically by portable electronic equipment like smartphones such as iPhones, personal computers, and tablets such as iPads. Most of these devices leave us with an impression of rigidity. Rigid electronics, if we may name the conventional electronics in this way, together with flexible electronics, will continue changing the world and provide us with convenience. Rigid electronics will keep thriving by decreasing the feature size of integrated circuits [1], by developing more efficient methods to produce silicon wafers with higher diameter, and by lightening and minimizing energy storage systems and through similar innovations. Although the future of flexible electronics is fascinating, it's quite challenging as well.

Flexible electronics' emergence dates back to the 1960s and this concept can be interpreted in various ways, such as plastic electronics, addressing its mechanical

S. Liao • Y. Huang • H. Wu (✉)
State Key Laboratory of New Ceramics and Fine Processing, School of Materials Science and Engineering, Tsinghua University, Beijing 100084, China
e-mail: huiwu@tsinghua.edu.cn

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_12

advantages, printable electronics, addressing its general manufacturing, and organic or polymer electronics, addressing the substance it relies on [2]. Generally, an electronic system is constructed with electronic components and devices, flexible substrates, interconnects, a bonding layer and an encapsulating layer. As regards flexible electronic systems, they are almost the same as a conventional electronic system, except for the elastomeric substrates. Theoretically, thinness can pave the way to obtain elastomeric substrates for flexible devices. The failure of materials results from reasons such as mechanical fragility, thermal mismatch between substrates and functional layers, and erosion from ambient chemicals, with mechanical fragility being the most severe problem for films. To be more specific, device failure is usually caused by a strain that is higher than the threshold. For example, when a film is bent, the largest strains occur on both surfaces with tensional strain and compressive strain being proportional to the thickness of the film. Since the threshold is a material constant, the thinner the layer is, the harder it is to collapse. However, the price paid for thinness is bad resistance against diffusion of oxygen and moisture. Meanwhile, warping and buckling tend to occur in thin films more severely when they bear thermal impact, and since mostly multi-layer devices are constituted with materials of different coefficients of thermal expansion, thermal mismatch often causes delamination. Thus, the single change in structure not only leads to numerous innovative applications but also brings about massive challenges, lying in both material preparation and device fabrication.

But is thinness the only solution? Absolutely not. There exists another philosophy to tackle the trouble. Since substrates merely work as carriers for functional parts and the failure of electronics comes from the disability of the functional parts, which in most scenarios, results from being damaged by over-stretched substrates, the stability and lifespan of the whole system can be extended once those functional parts are located on rigid substrates. Please note that we are not going back to rigid electronics but talking about a hybrid of the rigid and the flexible. Ships and boats might be torn apart by storms and waves, but buildings on the islands remain still. So are the functional parts on rigid islands, which are distributed in a flexible medium that undergoes external force field.

Many methods have been investigated for flexible electronics. In this review, we will summarize the state-of-the-art progress of electrospun nanofibers based on flexible electronic components and devices. The next section will give a brief introduction to the electrospinning method. The following section shows how the electrospun nanofibers are applied to fabricating components or devices such as transistors, transparent electrodes, sensors, and energy devices. The last section will examine the method critically to present its disadvantages and further more we will give our perspective on how the method can been revised to face the challenges.

12.2 Electrospinning

Several methods have been well established to prepare nanofibers, but each has its pros and cons. Organic nanofibers can be prepared by methods such as drawing, template-synthesis [3, 4], phase-separation [5–8], self-assembly [9–11], and electrospinning. More methods can be applied to prepared inorganic nanofibers, including metal nanofibers, carbon nanofibers, and ceramic nanofibers. To name a few, we have also template-synthesis [12–17], chemical vapor deposition (CVD) [18–22], solvent-thermal method, biomolecule self-assembling [23], and electrospinning. For more information, the reader might be interested in and refer to the review article by Nayak et al. [24].

Electrospinning is facile, and the facility makes low cost. Electrospinning is also versatile, not only because it is non-selective on the raw materials, but to the extent of the control it has on the morphology and structures of the products. Researchers, via electrospinning, can handle most forms of one-dimensional nanostructures, such as nanowires [25], nanofibers, nanobelts [26, 27], nanorods [28–30], nanogrids [31], and core-sheath structures [32–36]. Self-contained theories have been constructed, and the results altogether form a system to guide the electrospinning process [37, 38]. Unable to match the imagination of material scientists, the typical electrospinning method has been revised by changing the shape of the nozzle, by controlling the electric field, or by using various collecting electrodes.

A systematic study on how the electrospun fibers will behave when the geometric shape and feature size of the insulating region are varied has been presented [39]. An auxiliary electrode can be applied in the process [38, 40–42]. When the electric field of the counter electrode changes, the degree of orientation of electrospun nanofibers to the field direction of a target electrode changes [43, 44]. Using a three-pole electrospinning device with a channel electrode, researchers can produce well-aligned nanofibers [45]. The shape of the collecting electrode can be made cylindrical and rotatable [44, 46, 47]. A concentric nozzle, an inner nozzle and an outer nozzle, has been used in concentric electrospinning, to produce a core-shell nanostructure [48, 49]. A wire electrode has been used to sweep through a bath containing a polymeric solution in contact with a high voltage [50]. Setting the collecting electrode beside instead of under the nozzle and using a strong electric field to overcome the gravity, researchers can collect aligned fibers between the electrodes [51].

Though the electrospinning process is amazing, we are not going to go further but address some fundamental theoretical results and introduce the very conventional setup and procedure, and for more information on the method itself, you may be interested in two review articles, one by Greiner and Wendorff and the other by Li [52, 53].

Some transformative theoretical work regarding the method should be noted. Taylor was the first one to give the simulation result of how the shape of the charged droplet changed under the electric force [54]. Yarin et al. proposed a theory of stable shapes of droplets affected by an electric field and compared with data

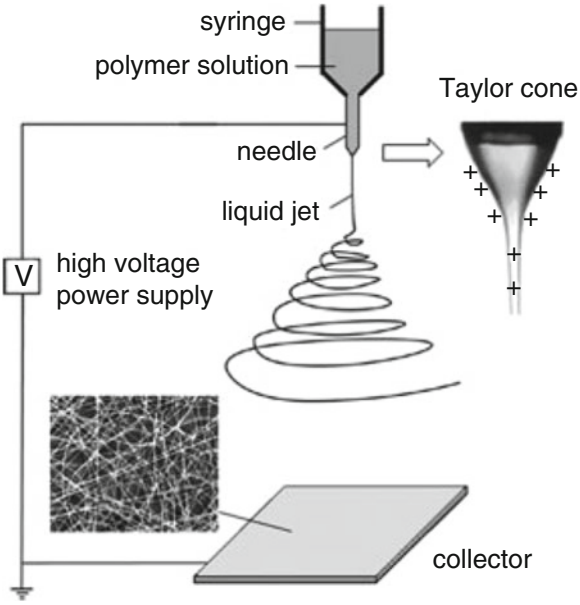


Fig. 12.1 A conventional setup of electrospinning. Reproduced with permission [53]. 2004, Wiley-VCH

acquired in their experimental work on electrospinning of nanofibers from polymer solutions and melts [55]. Carrying the same charges, when sprayed out of the spinneret, fibers tend to reject itself and whip. Shin et al. work showed that the fluid instability is a key element in the spinning process and revealed by a linear instability analysis that describes the jet behavior in terms of known fluid properties and operating conditions [56]. Kiselev et al. showed how elimination of the whipping motion of electrospinning fibers leads to nearly perfect alignment of fibers collected onto fast-rotating cylindrical collectors [57].

Now, we would like to demonstrate a typical electrospinning process. Figure-12.1 is a conventional setup.

The electric field strength applied is 100–500 kV m^{-1}, and the distance from the tip to the counter electrode is 10–25 cm. There are three main steps. Firstly, experimenters prepare the precursor solution, which will flow out at a certain rate controlled by the syringe pump. Then a high voltage is applied and the electric field provides a driving force to deform the charged droplet to become a Taylor cone. Finally, a jet is sprayed out of the spinneret and then deposited on a counter electrode randomly because of the whip motion. Readers should notice that much work has been done to eliminate the randomness, mostly by modifying the counter electrodes as mentioned above. Besides, in an actual experiment, methods such as direct-dispersing [58, 59], gas–solid reaction [60–62], in situ photoreduction [63, 64], sol-gel method [65, 66], emulsion [67–69], co-evaporation method and coaxial method are combined with the electrospinning process to improve the stability and properties of the one-dimensional nanostructure and after the deposition of the electrospun fibers, post treatments such as calcination, carbonation, and

activation are generally applied to obtain nanofibers of ceramics, carbon materials, and activated carbon, respectively. For more information about the methods to combine with electrospinning, readers will find the review by Lu et al. to be interesting [70].

12.3 Flexible Components and Devices Based on Electrospun Nanofibers

As mentioned above, a generic flexible electronic system is made up of five building blocks: electronic components and devices, flexible substrates, interconnects, a bonding layer, and an encapsulating layer [71]. While the rest of our review will be devoted to components and devices, we still want to debrief our readers about substrates, since within the realm we concern and on the level of lab research, the selection of substrates are covered by most of the published work. Thin glass, plastic films, and metal foils are three types of substrates most frequently used. None of them is perfect. The biggest problem for thin glass is their brittleness due to crack propagation. Plastic films suffer a lot from easy penetration of oxygen and water, which might harm the functional part in the long run, and the thermal mismatch stress, which limits the processing temperature and the serving ambience. Most of the published work have chosen polymer foils as substrates, such as polyethylene terephthalate (PET), polyethylene naphthalate (PEN), and polyimide (PI), not bad as the starting point towards industrial products. The last type of metal foils, with stainless steel be the most popular, typically have bad optical transmittance and most lethally, the surface roughness can cause the system to fail. Although, crack propagation can be avoided, to some extent, by coating plastic layer, bad resistance against chemicals by coating a barrier, and fine polishing can reduce the surface roughness, extra treatments are required. One intrinsic property of metal substrates is electric conductivity and since a counter electrode is required in any electrospinning setup, it might be interesting to deposit functional nanofibers on such substrates to accomplish direct fabrication of flexible devices, while in most of the related work, a transfer process is involved. The rest of our review will concentrate on how the method of electrospinning has been used in fabricating flexible electronic components and devices, and in improving the properties of the rest four building blocks.

Electronic components and devices can be subdivided into numerous categories. We would like to distinguish components from devices so that we can construct our statement in a more logical way. Components are the most basic elements in electronics and devices are integrations of certain components. Electrospinning is versatile but the functional nanofibers prepared in this way cannot cover all the materials used in flexible electronics. Here, in this review, we summarize the recent significant electrospinning progresses in components such as transistors, transparent electrodes and in devices such as sensors, solar cells, batteries, and supercapacitors.

12.3.1 Components for a Typical Electrical System: Field-Effect Transistors

Basically, a field-effect transistor (FET) operates as a capacitor where one plate is a conducting channel between two ohmic contacts: the source and drain electrodes [72]. In 1930, Lilienfeld first proposed the principle of the FET [73]. In 1987, after Koezuka et al. reported on a structure based on electrochemically polymerized polythiophene, organic FETs, or OFETs, were identified as promising candidates for building electronic devices [74, 75]. Many modern applications, including radio-frequency identification, next-generation displays, chemical sensors, and nonvolatile memories, are based on OFETs.

Although the electrospinning technique is suitable for almost all sorts of materials, most researches on fabrications of transistors tend to use polymers and organic–organic composites, which results in OFETs. Another advantage for OFETs is that they adopt the architecture of the thin-film transistor (TFT), thus flexibility achieved easily [72]. Still, carbon materials have been used via electrospinning. Chang et al. presented a simple yet scalable process by means of direct-write electrospun fibers to accomplish both doping and patterning of graphene simultaneously [74]. By combining near-field electrospinning with different types of functional polymer fibers to modulate the electrical properties, complementary graphene FETs including n- and p-type graphene FETs, pn, and other types of electronic junctions can be constructed on the same substrate under ambient pressure and room temperature. As mentioned, post-treatments are necessary steps to obtain nanofibers of ceramics, metals, or carbon materials, from the precursor, which would raise the cost and hinder the massive production. Choosing organic nanofibers, researchers easily made it a one-step device fabrication [76].

Using OFETs as building blocks, many modern applications become reality, such as nonvolatile memories [77–79]. Chang et al. reported on the fabrication and characterization of transistor memories on flexible PEN substrate using the electrospun nanofibers of poly(3-hexylthiophene) (P3HT) semiconductor: gold nanoparticles (Au NPs) hybrid [77]. Functionalized with self-assembled monolayer (SAM) of para-substituted amino (Au-NH_2), methyl (Au-CH_3) or trifluoromethyl (Au-CF_3) tail groups on the benzenethiol moiety, the ~10 nm sized Au NPs induced interface with the P3HT matrix, and it is at the interface that charge traps influence the accumulated density of holes. Due to the valence band set by the interface, it takes extra external electric field to offset, and this is the mechanism of the enhancement/depletion mode of OFETs. They used coaxial electrospinning to get the hybrid nanofibers, which seems to be relatively complicated, and it is a drawback. As for the fabrication of device, the flexible PEN with a 100-nm-thick Au gate electrode was subsequently deposited by thermal deposition, the atomic layer deposition (ALD) method was adopted to prepare 100-nm-thick Al_2O_3 as dielectric. After transferring the nanofiber onto the substrate, thick gold source/drain electrodes were thermally deposited, resulting in a prototypically memory shown in Fig. 12.2a with an optical picture inserted. Further, they studied the device

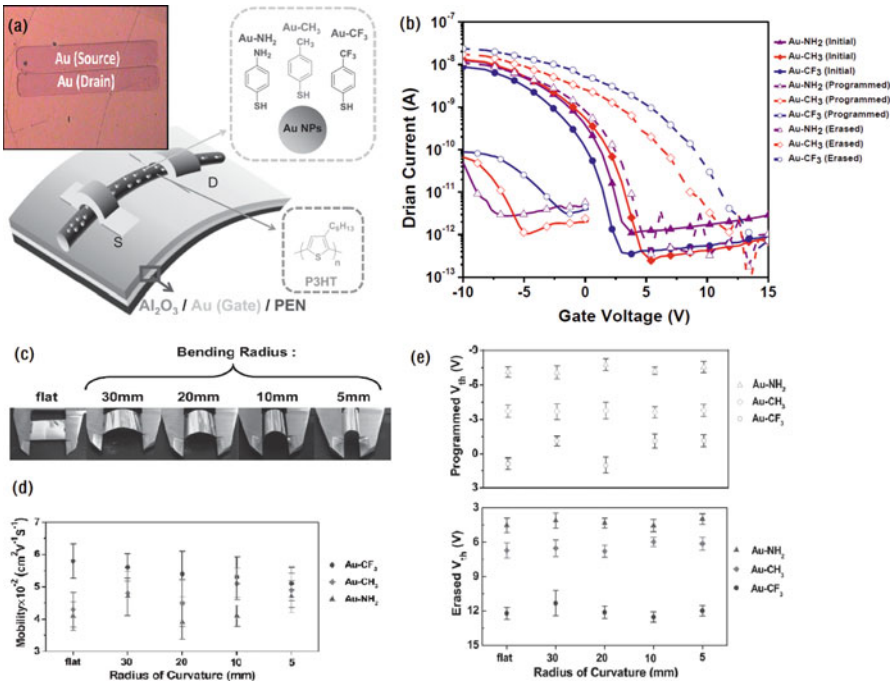


Fig. 12.2 (**a**) Schematic configuration of the hybrid nanofiber-based transistor memory devices, chemical structures of P3HT, and surface-modified Au NPs; the inset is the optical microscope image of the hybrid nanofiber-based transistor memories, (**b**) Transfer characteristics of P3HT:Au hybrid nanofiber-based transistor memories device (programmed state: $V_g = -5$ V, 1 ms; erased state: $V_g = 5$ V, 1 ms; $V_d = -5$ V), (**c**) Hybrid nanofiber-based transistor memory devices under flat and various bending radii, (**d**) Variation on the mobility of the flexible hybrid nanofiber-based transistor memory devices, (**e**) Variation of programmed and erased threshold voltages of the flexible hybrid nanofiber-based transistor memory devices. Reproduced with permission [77]. 2013, Wiley-VCH

performances including operation voltages, retention, and endurance ability as well as stability against external mechanical stimulus. Figure 12.2b shows the transfer characteristics I_d–V_g (drain current–gate voltage curve) sweeping from 15 V to −10 V at a step length of 5 V. The test of flexibility was conducted by flexing the memory with a Vernier caliper, as shown in Fig. 12.2c. First of all, the device architectures were not cracked or deformed. Secondly, the statistical data on threshold voltages under programmed/erased state, mobility, and ON/OFF state current with various bending radii or repeated cycles were collected and analyzed. As Fig. 12.2d, e showed, the mobility and the threshold voltages remained similar without any notable fluctuation under various curvature radii of 30, 20, 10, and 5 mm or under a 1000-bending cycles at a 2-bends/s rate.

Aside from the complicated processing, doubt can be cast on the thermally stability of such an organic system.

12.3.2 Transparent Electrodes: A Component of Cutting-Edge Technology

In modern devices such as touch panels, solar cells, organic light emitting diode (OLED), and liquid crystal display (LCD), traditional electrodes don't suit any more due to the requirements of both transparency and conductivity [80–84]. The conventional material is indium doped tin oxide (ITO), which has been studied for over 60 years. The technology of ITO is mature but ITOs' application and development are highly limited by two characteristics: high cost and brittleness. Apparently, ITOs don't meet the requirements of flexible electronics. Because of the innate brittleness resulting from their ceramic nature, doped metal oxides such as ITO, aluminum doped zinc oxides (AZO), gallium doped zinc oxides (GZO), and fluorine doped tin oxides (FTO) are no longer the materials of first choice in flexible electronics. Scientists have been focusing on polymers, since the discovery of conductive polymers in 1970s [85], carbon nanostructures, and metal nanostructures. Electrospinning, along with other inspiring post treatments, has been adopted in this area as well.

Because of the molecular structure and the nature of metal bonding, respectively, polymers and metals are the two most promising candidates with intrinsic ductility and flexibility. Additionally, metals present great conductivity. So, electrospinning has mostly been conducted to fabricate metal networks for transparent conductive electrodes with flexibility and efforts to obtain the electrodes based on conducting polymers have also been made [80, 86–88]. Moreover, the electrospinning method has also worked as an associating step in other electrode fabrication [89–91].

For example, Hsu et al. combined electrospinning with electroless deposition [86]. The process flow was shown schematically in Fig. 12.3a. Two key steps guaranteed great sheet resistance-transparency (R_s–T) performance. Tin (II) chloride was dissolved in the polyvinyl butyral (PVB) solution as the precursor. Before electrospinning, the substrates were spin-coated with hydrophobic polymer film to reduce metal precursor onto the substrates. After electrospinning, the nanowires were immersed in silver nitrate aqueous solution so that Ag^+ got reduced to form Ag seed layer, which contributed to selectively metal deposition. The selectively deposition was achieved by the hydrophobic film and the metal seed layer to reach high R_s–T performance, which was shown in Fig. 12.3b. This could be attributed to the reduced number of junctions, since the electrospun nanofibers tend to exhibit large aspect ratio, and low junction resistance, because the electroless deposition naturally "fused" the junctions, as shown in Fig. 12.3c. Besides, they studied the relationship between the diameter of Cu nanowires and the electroless deposition time. Results in Fig. 12.3d proved that the diameter could be controlled to adjust the R_s–T performance. To demonstrate the mechanical flexibility, a bending test was applied to evaluate, as shown in Fig. 12.3e. Figure 12.3f quantitatively demonstrated the flexibility and the durability.

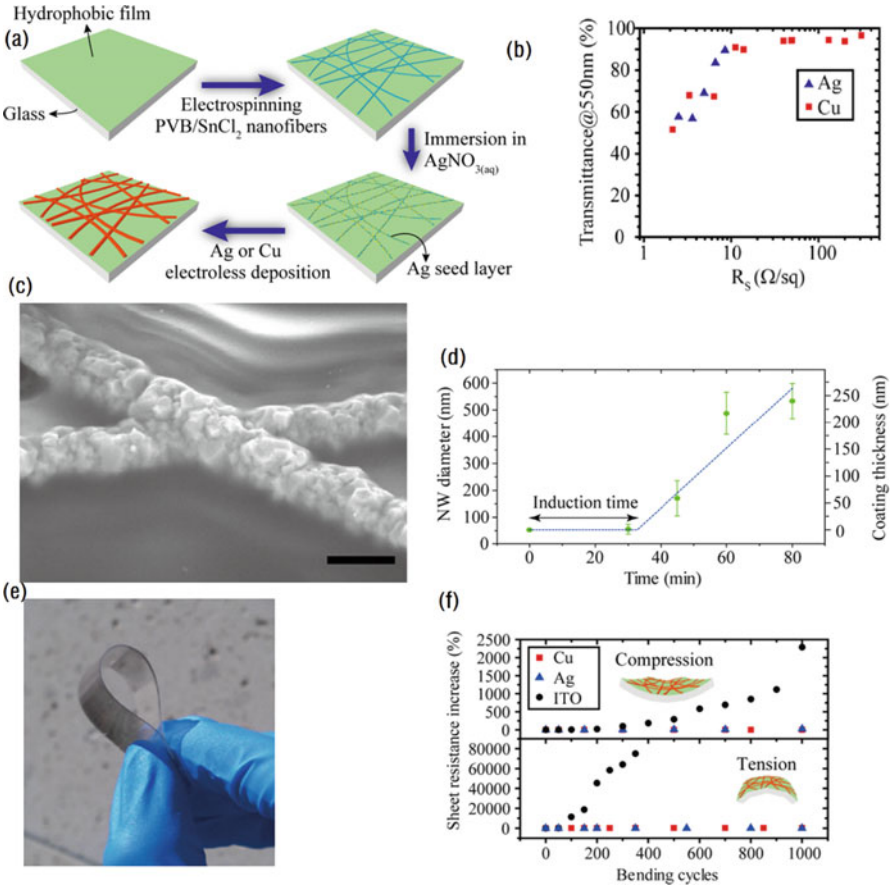


Fig. 12.3 (**a**) The process flow of electrolessly deposited metallic electrospun nanowire transparent electrodes, (**b**) R_s–T relationships of the electrolessly deposited metal nanowires, (**c**) the fused junctions decreased the resistance, (**d**) the thickness of the Cu NWs was time-dependent and the growth exhibited a linear increase after an induction period, (**e**) photograph of a copper nanowire transparent electrode deposited on a flexible PET substrate, (**f**) the results of cycling stability tests: with a bending radius of 4 mm, the metal NWs maintained its initial conductivity after 1000 bending cycles. Reproduced with permission [86]. 2014, American Chemical Society

Copper nanofiber networks is one of the most promising candidates to replace ITO due to the advantages of low cost and moderate flexibility, etc. However, one factor that must be considered when the lab products are pushed forward into industry is their resistance against tough serving environment. To prevent the increase of sheet resistance caused by thermal oxidation or chemical corrosion, Cui's group developed a way of passive coating, based on atomic layer deposition, during which process AZO and Al_2O_3 were deposited on the active metal nanofibers [92].

Regarding polymer-based electrodes, Ramakrishna et al. fabricated a counter electrode for flexible dye-sensitized solar cells (DSC) by directly depositing conductive polyaniline doped with 10-camphorsulfonic acid (PANI × CAS) blended with polylactic acid (PLA) composite films on flexible indium tin oxide-coated PEN substrate [93]. And the photoelectric conversion efficiency of the DSCs achieved 3.1 % under 1 sun illumination of 100 nW cm^{-2}. As an indispensable building block of solar cell, transparent and conductive electrodes have been widely used and since this is a device level topic, we will cover this later with more detailed description.

Another recent work worth mentioning is a transparent electrode based on metal nanotrough network [89]. Using the electrospun nanofiber network as template, Wu et al. adopted some standard thin-film coating methods to deposit functional materials onto the nanofiber network. A schematic demonstration of the fabricating process can be found in Fig. 12.4a, during which they produced random or uniaxially aligned nanofiber networks (Fig. 12.4b). There are several eye-catching points in their work, but here we list three of them. First, nanotroughs' resistance against tougher mechanic environment is better most other typical nanostructures, and this is where the flexibility comes, as shown in Fig. 12.4c, d. Second, the high resistance of most nanostructure is due to the junction and many researches have been conducted to tackle this problem. In this work, due to the anisotropy of physical depositing processes, or directionality of most thin-film coating methods such as thermal evaporation, electron-beam evaporation, or magnetron sputtering, the deposition of functional materials happens preferentially on one side of the network. With the other organic side being exposed, it is easy to remove the organic template by dissolving, thus the contact resistance issue can be perfectly solved, as shown in Fig. 12.4e. Third, the process is suitable for a wide range of materials since, as mentioned above, most physical depositions can be applied. They have fabricated nanotrough network based on materials including silicon, indium tin oxide, and metals such as gold, silver, copper, platinum, aluminum, chromium, nickel, and their alloys. When put into actual usage, requirements differ according to different particular applications. The tradeoff between transmittance and conductivity can be found in Fig. 12.4f. As for the device fabrication, a flexible touch screen and a transparent conductive tape were built to demonstrate its novel performance.

12.3.3 Devices for Interaction: Sensors

Sensors detect the information such as strain, some gas, or a specific chemical from the environment and convert it into information of another form, mostly the change of voltage and current. In the field of flexible electronics, sensors are even more important especially when they are used in electronic skins [94]. To take advantage of the large specific surface of nanostructures, functional materials have been electrospun to be highly sensitive sensors. Electrospinning, as a novel method to

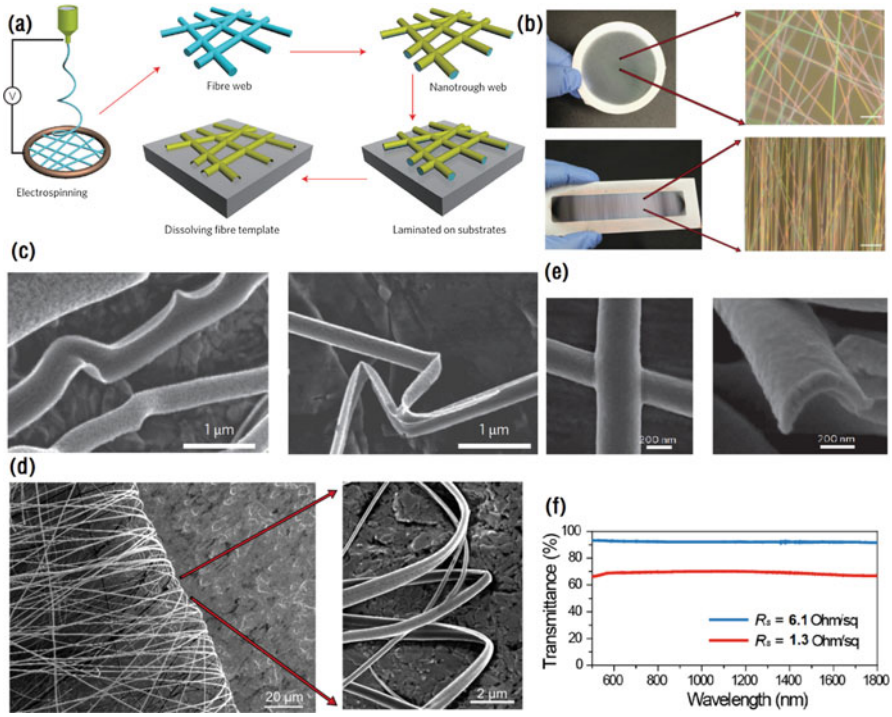


Fig. 12.4 (**a**) Schematic of the polymer-nanofiber templating process for fabricating nanotroughs, (**b**) electrospun fibers deposited on different fiber collectors, a copper ring with a diameter of 5 cm and two parallel electrodes with a gap of 3 cm, (**c**) SEM images of gold nanotroughs on aluminum foil after folding, (**d**) transmittance spectra of two gold nanotrough network samples scanned, (**e**) SEM images of a freestanding metal nanotrough network after been folded from 500 to 1800 nm, showing very flat spectrum for the entire wavelength, (**f**) a *top-view* SEM images of a junction between two gold nanotroughs (*left*) and an SEM image of the cross-section of a single gold nanotrough, revealing its concave shape (*right*). Reproduced with permission [89]. 2013, Nature Publishing Group

prepare 1-D nanomaterials, has been widely used in the process of sensor fabrication. Materials that have been sufficiently investigated and applied to fabricate strain sensors, gas sensors, biosensors, UV sensors, humidity sensors, pressure sensors, and sensors for the purpose of detecting certain chemicals such as Hg^+ [95], H_2S [96, 97], and H_2O_2 [98, 99]. For a better comprehension about how this method is applied in sensors, Table 12.1 is given to show readers a bigger picture of the recent development of the area, in which all the cited articles are published during the past 6 years. Besides, the review by Ding et al. gave an overview on gas sensors using electrospun nanofibers comprising polyelectrolytes, conducting polymer composites, and semiconductors based on various sensing techniques such as acoustic wave, resistive, photoelectric, and optical techniques [145].

Table 12.1 Sensors based on different electrospun materials

	Strain	Gas	Bio	UV	Humidity	Pressure	Ion	H_2S/H_2O_2
Metal	[100]							
Composite	[49, 101–107]			[108, 109]	[110]	[111]		[97, 99]
Carbon	[112]		[113]		[114, 115]			
Ceramics		[116–123]	[124]	[125]	[118, 126, 127]			[96, 98]
Polymer	[101, 102, 128–130]	[131, 132]	[133–136]	[137–139]	[140–142]	[143, 144]	[95]	

Things can be much easier if a mere sensor is the ultimate goal. However, it will really take some finesses to reach flexibility. To the best of our knowledge, most researches towards building sensors, no matter flexible or not, via electrospinning, have been dedicated to strain sensors [49, 101, 102, 128, 146], due to the objective requirements to detect rather small deformation. But still, other sorts of sensors have been successfully fabricated by researchers [147–150]. For example, Mandal et al. showed that the piezoelectricity of as-electrospun poly (vinylidene fluoride-trifluoroethylene) (P(VDF-TrFE)) nanofiber webs opens up new possibilities for their use as a flexible nanogenerators and nano-pressure sensors [143].

Park et al. established a micropatterning strategy, shown in Fig. 12.5a which would find its application in stretchable circuits and strain sensors [128]. In their experiment, the poly(4-vinylpyridine) (P4VP) of 30 wt% dissolved in dimethylformamide (DMF) was electrospun and collected on a PET film to obtain the nanofiber mat. Then, certain areas were shielded by Ni shadow mask and the rest areas were irradiated with UV/O3 (UVO). And since the shielded areas would not be dissolved in ethanol, the micropattern would be formed after the mat was dipped in a precursor solution prepared by dissolving $HAuCl_4 \times 3H_2O$ in 0.02 M ethanol. Constant tensile force of 20N with 20 μm/s was applied until $\varepsilon = 0.37$ to obtain the stress–strain curve in Fig. 12.5b. The similarity of the overall shapes of the three indicated that the nanofiber mat maintained its overall mechanical behavior. By controlling the reduction times, the strain sensitivity of the composite mat can be controlled, as shown in Fig. 12.5c. Since electric circuits require invariance in resistivity under a mechanical strain, while strain sensors require a large variance in resistivity under an external strain, the composite mat could find applications in both. As shown in Fig. 12.5d, both the strain-invariant and the strain-sensitive patterns were employed to fabricate a strain sensor. Finally, an array of strain sensors and circuits from a single material was created. The bending measurement was carried out at a frequency of 1 Hz (32 cycles for each strain value in the range 0.03–0.17) and the pattern was completely stable over 200 bending cycles.

12.3.4 Devices for Energy Generation and Storage: Solar Cells, Batteries, and Supercapacitors

Energy is one of the most crucial issues we have been faced up with ever after the millennium. With the explosion of population, the drastic development of industry, and the increasing concerns about environment, the final choice of methods of energy generation will have to strike a balance among all stakeholders. From the perspective of benefits among countries, energy is the course of strained international relationship since the amount of fossil fuels, roughly including gas, oil, and coals, is limited [151]. And furthermore, with fossil fuels being the main energy sources, legitimate share of carbon dioxide emission becomes the arguing point of many international summits. From the standing of commerce, an accessible energy

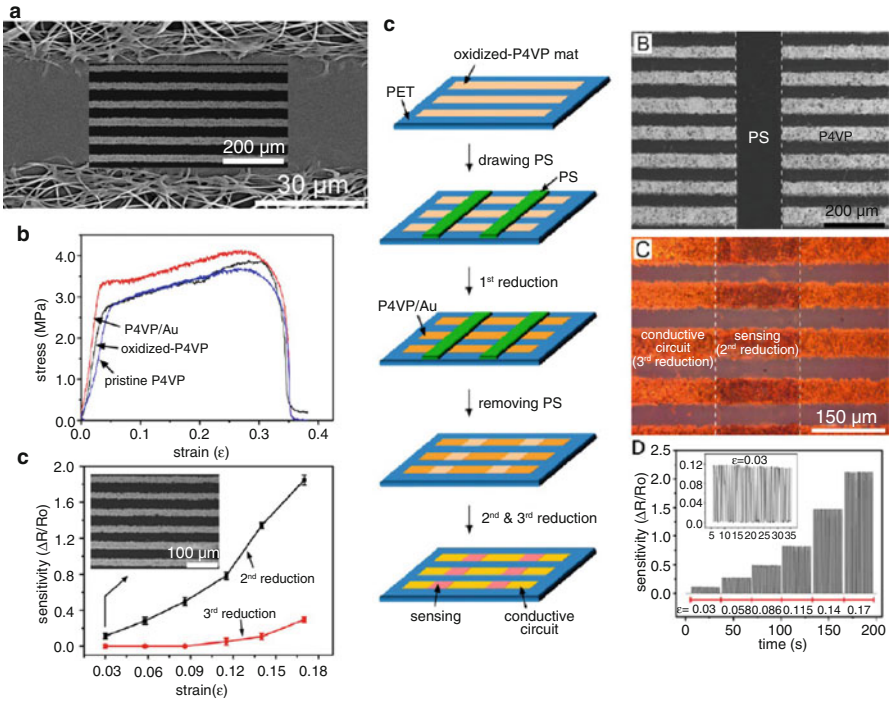


Fig. 12.5 (**a**) The micropattern of the composite nanofiber mat (**b**) stress–strain curves for pristine P4VP fiber mat, oxidized P4VP fiber mat, and P4VP/Au composite fiber mat, (**c**) sensitivity changed with strain for the composite pattern obtained after the second and third reduction cycles, (**d**) a schematic fabricating process, an SEM image of the patterned P4VP mat, an optical microscopy image of P4VP/Au composite patterns and a real-time sensitivity change in strain in the range $\varepsilon = 0.03$–0.17. Reproduced with permission [128]. 2013, American Chemical Society

source, from which an electronic gadget can get recharged fast, can free people from carrying mobile power and anyone will be amazed by how much space, and accordingly how much weight and it is heavier than it seems to be, the batteries have taken if they open any Apple devices [152–154]. It has to be admitted that this is really a waste of space and we can hardly imagine how thin those tablets and cellphones can be and what fascinating functions can be added into them if the space can be spared. It can be fairly addressed that not only the evolution of Apple devices but the success of Tesla electronic cars count on more advanced energy solutions, as Tesla's Chief Technology Officer envisioned that the combination of batteries and solar cells would lead to cheap electricity [155, 156].

Let's start from the energy source. It can be easily found that the account of renewable sources will increase largely in the next two decades, and the share of renewable energy in world energy demand will increase from 10 % in 2010 to 14 % in 2035 [157]. And as the most approachable sustainable energy source, solar energy plays a huge part in the sector of renewable energy. Although we have

mentioned the application as solar cells in the section of transparent conductive electrodes, which is an indispensable component of the device, this section will go one step further on this energy generating method. Among the recent work, transparent conductive electrodes built via electrospinning have been one of the building blocks of solar cells. Wu et al. constructed a copper nanofiber-based transparent conductive electrode with a three-step processing demonstrated schematically in Fig. 12.6a [80]. In step one, copper acetate/poly(vinyl acetate) (PVA) solution as the precursor was electrospun onto a glass substrate. Step two was heat treatment in air at 500 °C for 2 h to remove the polymer and transform copper ion into dark brown CuO. The last step saw the reduction of CuO by annealing in an H_2 atmosphere at 300 °C for 1 h. Although the fabrication process is not as simple as some one-step construction, its products does have some excellent property that is unachievable through other methods, once again, the junction resistance. It is never hard to understand why the fused junction, shown in Fig. 12.6b, can be obtained once we know well about the processing. And since electrospinning is highly tunable, experimenters controlled the electrospinning time to adjust the fiber density to reach variable transmittance/resistance to meet the requirements of specific applications, which can be found in Fig. 12.6c. Besides, due to the large aspect ratio of the copper nanofibers, the electrode was bendable and possessed some stretching ability, as shown in Fig. 12.6d, e. In the level of device fabrication, a Poly-3-hexylthiophene (P3HT):[6,6]-phenyl-C61-butyric acid (PCBM) solar cell, shown in Fig. 12.6f, was fabricated using routine methods. They spin-coated a 50-nm layer of poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS) onto the electrode and then the sample was transferred to a nitrogen filled glove box and annealed at 110 °C for 10 min. The active layer (P3HT:PCBM 1:1 weight ratio, 25 mg/mL in dichlorobenzene, film thickness: ~240 nm) and metal electrode (7 nm Ca/200 nm Al) deposition and device testing were performed inside a nitrogen glove box. The power conversion efficiency of the device is 3.0 %, which is comparable to devices made on glass/ITO substrates.

As for energy storage, battery is the most frequently used technique for current supplying. In a conventional system, chemical reactions are generally involved in the process of energy storing and releasing. The two electrodes of an electrochemical cell are separated by an electrolyte causing the anode/anodic half-cell reaction to take place at the corresponding electrode to produce a potential difference. Electrochemical cells are connected in series or in parallel to be a battery. Li ion batteries outperform other battery technologies such as lead-acid batteries and Ni-Cd batteries, due to the high energy density and potential towards flexibility [158–160]. In Li ion batteries, the presence of Li in its ionic state rather than metallic state solves the dendrite problem, and that contributes to a better safety compared to Li-metal batteries [161]. Functional nanofibers are widely used in the field, especially on lithium ion batteries. Researchers' work has been involved in improving lithium ion batteries' electrolyte [162, 163], electrodes [164–170], and separators [165, 171] by electrospinning. Silicon is a promising replacement of graphite in the current graphite/$LiCoO_2$ batteries with ten times higher theoretical capacity, abundance, non-toxicity, and low cost. But further development is limited

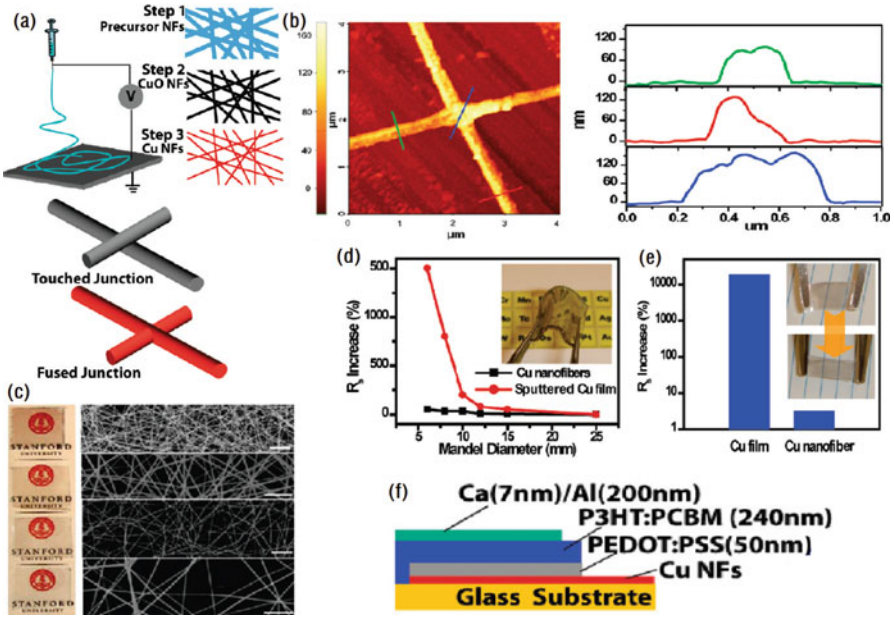


Fig. 12.6 (**a**) Schematic of materials preparation method. *Left column*: Schematic of an electrospinning setup, shown without a syringe pump. *Right column*: the fabrication process of Cu nanofibers. In the first step, CuAc2/PVA composite fibers were prepared by electrospinning. In step 2, the fibers were calcinated in air to get CuO nanofibers. In step 3, the CuO nanofibers were reduced to Cu nanofibers by annealing in an H_2 atmosphere, (**b**) AFM image of a junction between two nanofibers. The curved lines show the heights of two nanofibers and the cross junction, respectively, (**c**) Digital photos of a series of Cu nanofiber transparent electrodes with different fiber densities. Each sample has a size of 2 cm by 2.5 cm. The right column shows corresponding SEM images. Scale bar is 2 μm, (**d**) The transparent electrodes based on Cu nanofiber networks show much better flexibilities than sputtered Cu films on PDMS substrates. The Mandrel diameter is the bending radius, (**e**) Cu nanofiber networks show much smaller changes in terms of sheet resistance upon stretching with 10 % strain. The sheet resistance was measured after the films were released back to their original lengths, (**f**) a schematic demonstration of the P3HT:PCBM solar cell using a Cu nanofiber film as the transparent electrode. Reproduced with permission [80]. 2010, American Chemical Society

by the poor conductivity of Si and the bad cycling stability. Using a simultaneous electrospraying and electrospinning technique, Xu et al. synthesized a flexible 3D Si/C fiber paper electrode by simultaneously electrospraying nano-Si-polyacrylonitrile (PAN) clusters and electrospinning PAN fibers followed by carbonization. The setup, demonstrated in Fig. 12.7a, can incorporate Si nanoparticles into a carbon textile matrix uniformly, and the occupied fiber cages in the network provided free space to accommodate the volume expansion of nano-Si, as shown in Fig. 12.7b, c. Figure 12.7d, e illustrates both the as-synthesized Si/PAN paper and the post-carbonization Si/C composite film exhibited good flexibility. To prove its electrochemical performance as anodes, they built a coin cell with lithium as counter electrodes. It turned out to demonstrate a very high overall capacity of

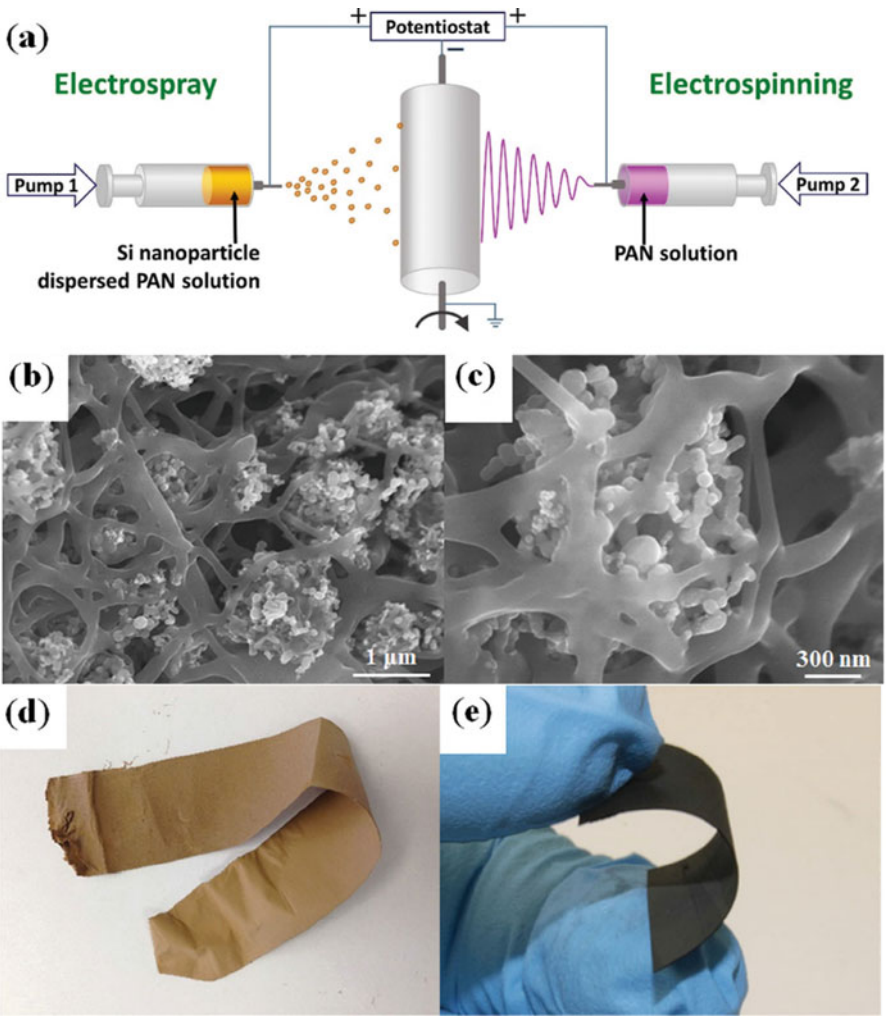


Fig. 12.7 (**a**) The synthesis process and the architecture of the flexible 3D Si/C fiber paper electrode, (**b**, **c**) *top-view* SEM images of the 3D Si/C fiber paper electrode, (**d**) Photographs of the electrospun/sprayed flexible paper electrode before carbonization, (**e**) Photographs of the electrospun/sprayed flexible paper electrode after carbonization. Reproduced with permission [172]. 2014, Wiley-VCH

~1600 mAh g^{-1} with capacity loss less than 0.079 % per cycle for 600 cycles, and excellent rate capability. This work presents a good solution for nano-Si loading and indeed, the paper electrode bears fair flexibility or bendability. But whether it can function well as anodes when deformed is not mentioned and its device flexibility is yet to be proven.

Supercapacitors, or electrochemical supercapacitors, is one of the most effective and practical technologies for energy storage. They are different from traditional

dielectric capacitors and batteries, but can also be able to supply a high power output as well as store energy. The development of supercapacitors suffers from two disadvantages: low energy density and high production cost [173]. The property and structure of the electrode materials are vital in the improvement of energy density. Electrospinning can tackle the two problems simultaneously. The method has been applied to construct nanostructures with high surface area and high porosity from carbon materials [174–178], polymers [48, 175, 179–182], and metal oxides [175, 181] for the electrode application. One way to further decrease the cost is to choose the right material. Lignin is the second most abundant natural polymer after cellulose and based on alkali lignin. Lai et al. prepared mechanically flexible mats for the electrode use in supercapacitors [178]. The spin precursors were aqueous mixtures of lignin and PVA with different mass ratios of 30/70, 50/50, and 70/30. The lignin/PVA composite nanofibers were collected on aluminum foil covering the roller. After stabilization and carbonization, product mats in Fig. 12.8a were obtained. The PVA mats were brittle as can be seen in the inset of Fig. 12.8b, while the presence of lignin strengthened the mechanical performance. Figure 12.8c was an SEM image of 70/30 lignin/PVA composite nanofiber mat and

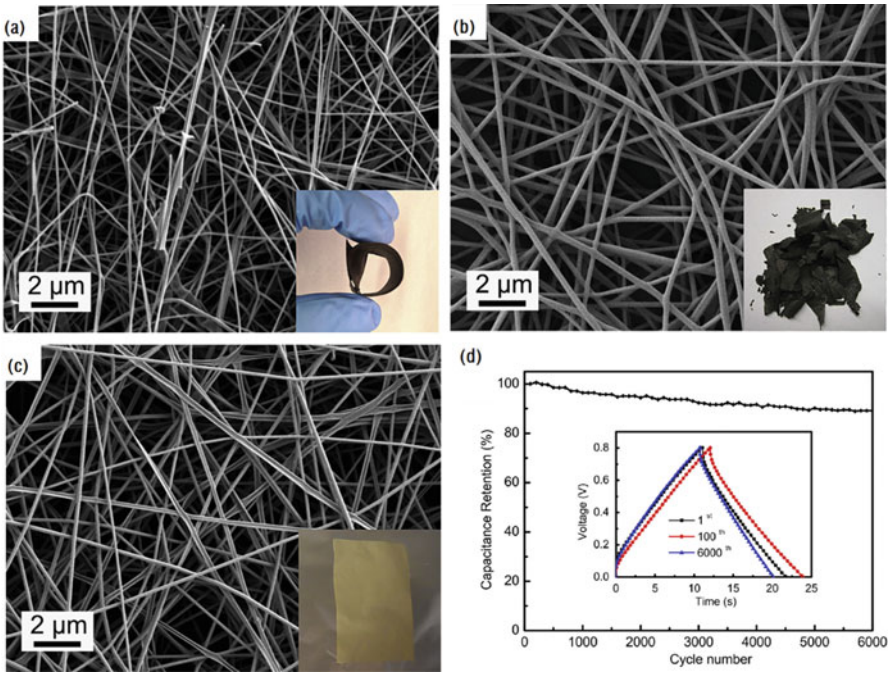


Fig. 12.8 (**a**) Electrospun carbon nanofibers made from 70/30 lignin/PVA, the inset showing the flexibility, (**b**) electrospun carbon nanofibers made from neat PVA, the inset showing the brittleness, (**c**) 70/30 lignin/PVA composite nanofiber mat from 12 wt% aqueous solution, (**d**) cycling stability of the mat in (**c**) at the current density of 2000 mA g^{-1}, the inset showing the charge/discharge curves of the 1st, 1000th, and 6000th cycles at 2000 mA g^{-1}. Reproduced with permission [178]. 2014, Elsevier

the diameter was ~140 nm. In the next stage of their experiment, electrochemical capacitive performance of the prepared electrospun carbon nanofibers (ECNF) were investigated by using cyclic voltammetry, galvanostatic charge/discharge, and electrochemical impedance spectroscopy. The nanofiber mats also exhibited excellent cycling stability as shown in Fig. 12.8d. The work revealed that high amount of lignin in the precursor nanofibers resulted in smaller average pore size, larger pore volume, and higher specific surface area in the electrospun mats.

12.4 Conclusions and Outlook

In this review, a brief introduction to flexible electronics has been given; the biggest difference in flexible electronics from the conventional electronics is the substrate. This single change leads to fascinating properties, as well as a series of problems. Electrospinning is one of the most promising methods of fabricating nanostructures. Instead of showing details in the procedure, we listed the main steps in a classic process. Most importantly, as researchers have revised the method to meet their needs, we have summarized those special electrospinning methods by citing recent related work to provide examples. In order to maintain functionality with flexible substrates, certain materials have been used and special structures are applied in building components or devices. We have summarized the significant progresses of field-effect transistors, transparent electrodes, sensors, solar cells, batteries, and supercapacitors, as they are in the most recent research.

But problems remain and we are faced with challenges. Great achievements in controlling the morphology of nanostructures via electrospinning have been made, but more precise control such as the spacing between two separated nanofibers is a dream still to come true. The versatility of electrospinning methods in flexible electronics can be questioned because though all categories of materials have been successfully electrospun into functional structures, most work has been conducted on polymers and the least focused material is ceramics. We believe one of the challenges is whether scientists can further revise the method to overcome the innate brittleness of ceramics and include more materials. One last issue emerges after reflection on the imbalance of the application of electrospinning to the building blocks of flexible electronic systems. As has been indicated, some related research has produced flexible materials, instead of flexible devices, and future work, if possible, should go one step further into the device level. At this moment, most products' flexibility can only be addressed as the ability to bend. To improve the stretching ability via electrospinning is rarely explored. Electrospinning might perform well in fabricating fiber networks conforming irregular surfaces. Electrospinning can be time-consuming, especially when a dense network is demanded. The time cost can be a disadvantage in industry. Besides, to accelerate its commercialization, how to modify this method so that it can be compatible with roll-to-roll technologies is very important. To start with, researchers want to consider the possibility of removing the transfer process but doing in situ

fabrication, or at least improving the efficiency of transferring. Since transferring, in most cases, is due to the limited thermal budget of plastic substrates, this is not only about electrospinning, but also some other disciplines.

References

1. Ferain I, Colinge CA, Colinge J-P (2011) Nature 479:310
2. Crabb RL, Treble FC (1967) Nature 213:1223
3. Long YZ, Duvail JL, Chen ZJ, Jin AZ, Gu CZ (2009) Polym Adv Technol 20:541
4. Pan L, Qiu H, Dou C, Li Y, Pu L, Xu J, Shi Y (2010) Int J Mol Sci 11:2636
5. Sun L-L, Zhao Y, Zhong W-H (2011) Macromol Mater Eng 296:992
6. Zhao J, Han W, Tang M, Tu M, Zeng R, Liang Z, Zhou C (2013) Mater Sci Eng C 33:1546
7. Ichimori T, Mizuma K, Uchida T, Yamazaki S, Kimura K (2013) J Appl Polym Sci 128:1282
8. Shao J, Chen C, Wang Y, Chen X, Du C (2012) React Funct Polym 72:765
9. Lee CC, Grenier C, Meijer EW, Schenning APHJ (2009) Chem Soc Rev 38:671
10. Hartgerink JD, Beniash E, Stupp SI (2001) Science 294:1684
11. Fu IW, Nguyen HD (2015) Biomacromolecules 16:2209
12. Xie G, Li X, Jiao H (2008) J Dispers Sci Technol 29:120
13. Dewangan K, Patil GP, Kashid RV, Bagal VS, More MA, Joag DS, Gajbhiye NS, Chavan PG (2014) Vacuum 109:223
14. Cheng Y, Li T, Fang C, Zhang M, Liu X, Yu R, Hu J (2013) Appl Surf Sci 282:862
15. Xie G, Wang Z, Li G, Shi Y, Cui Z, Zhang Z (2007) Mater Lett 61:2641
16. Huang B, Li C, Wang J (2013) J Magn Magn Mater 335:28
17. Liang H-W, Guan, Q.-F, Chen L-F, Zhu Z, Zhang W-J, Yu S-H (2012) Angew Chem 124:5191; Angew Chem Int Ed 51:5101
18. Hoshi F, Tsugawa K, Goto A, Ishikura T, Yamashita S, Yumura M, Hirao T, Oura K, Koga Y (2001) Diam Relat Mater 10:254
19. Thakur DB, Tiggelaar RM, Gardeniers JGE, Lefferts L, Seshan K (2009) Surf Coat Technol 203:3435
20. Somani SP, Somani PR, Tanemura M, Lau SP, Umeno M (2009) Curr Appl Phys 9:144
21. El Mel AA, Gautron E, Choi CH, Angleraud B, Granier A, Tessier PY (2010) Nanotechnology 21:435603
22. Butt FK, Cao C, Khan WS, Ali Z, Mahmood T, Ahmed R, Hussain S, Nabi G (2012) Mater Chem Phys 136:10
23. Gottlieb D, Morin SA, Jin S, Raines RT (2008) J Mater Chem 18:3865
24. Nayak R, Padhye R, Kyratzis IL, Truong Y, Arnold L (2012) Text Res J 82:129
25. Khalil A, Lalia BS, Hashaikeh R, Khraisheh M (2013) J Appl Phys 114:171301
26. Lu B, Guo X, Bao Z, Li X, Liu Y, Zhu C, Wang Y, Xie E (2011) Nanoscale 3:2145
27. Ko YW, Teh PF, Pramana SS, Wong CL, Su T, Li L, Madhavi S (2015) ChemelEctroChem 2:837
28. Fujihara K, Kumar A, Jose R, Ramakrishna S, Uchida S (2007) Nanotechnology 18:365709
29. Cherian CT, Sundaramurthy J, Kalaivani M, Ragupathy P, Kumar PS, Thavasi V, Reddy MV, Sow CH, Mhaisalkar SG, Ramakrishna S, Chowdari BVR (2012) J Mater Chem 22:12198
30. Fang X, Olesik SV (2014) Anal Chim Acta 830:1
31. Jusang L, Gouma PI (2011) J Nanomater 2011:863631
32. Su Y, Li X, Tan L, Huang C, Mo X (2009) Polymer 50:4212
33. Sun ZC, Zussman E, Yarin AL, Wendorff JH, Greiner A (2003) Adv Mater 15:1929
34. Li D, Xia YN (2004) Nano Lett 4:933
35. Haslauer CM, Moghe AK, Osborne JA, Gupta BS, Loboa EG (2011) J Biomater Sci Ed 22:1695

36. Liu G, Tang Q, Yu Y, Li J, Luo J, Li M (2014) Polym Adv Technol 25:1596
37. Carnell LS, Siochi EJ, Wincheski RA, Holloway NM, Clark RL (2009) Scr Mater 60:359
38. Mit-uppatham C, Nithitanakul M, Supaphol P (2004) Macromol Symp 216:293
39. Li D, Ouyang G, McCann JT, Xia YN (2005) Nano Lett 5:913
40. Wu Y, Carnell LA, Clark RL (2007) Polymer 48:5653
41. Kim GH, Yoon H (2008) Appl Phys Lett 93:023127
42. Arras MML, Grasl C, Bergmeister H, Schima H (2012) Sci Technol Adv Mater 13:035008
43. Kim GH (2006) J Polym Sci B Polym Phys 44:1426
44. Lee H, Yoon H, Kim G (2009) Appl Phys A Mater Sci Process 97:559
45. Jafari A, Jeon J-H, Oh I-K (2011) Macromol Rapid Commun 32:921
46. Sutka A, Kukle S, Gravitis J, Milasius R, Malasauskiene J (2013) Adv Mater Sci Eng 2013:932636
47. Kim G, Cho Y-S, Kim WD (2006) Eur Polym J 42:2031
48. Laforgue A, Power J (2011) Sources 196:559
49. Tiwari MK, Yarin AL, Megaridis CM (2008) J Appl Phys 103:044305
50. Forward KM, Rutledge GC (2012) Chem Eng J 183:492
51. Wu H, Lin D, Zhang R, Pan W (2007) J Am Ceram Soc 90:632
52. Greiner A, Wendorff JH (2007) *Angew Chem* 119:5731; *Angew Chem Int Ed* 46:5670
53. Li D, Xia YN (2004) Adv Mater 16:1151
54. Taylor GI, McEwan AD, Fluid J (1965) J Fluid Mech 22:1
55. Yarin AL, Koombhongse S, Reneker DH (2001) J Appl Phys 104:4836
56. Shin YM, Hohman MM, Brenner MP, Rutledge GC (2001) Appl Phys Lett 75:1149
57. Kiselev P, Rosell-Llompart J (2012) J Appl Polym Sci 125:2433
58. Hwang HJ, Barakat NAM, Kanjwal MA, Sheikh FA, Kim HY, Abadir MF (2010) Macromol Res 18:551
59. Kim G-M, Wutzler A, Radusch H-J, Michler GH, Simon P, Sperling RA, Parak WJ (2005) Chem Mater 17:4949
60. Xu J, Cui X, Zhang J, Liang H, Wang H, Li J (2008) Bull Mater Sci 31:189
61. Zhang C, Liu Q, Zhan N, Yang Q, Song Y, Sun L, Wang H, Li Y (2010) Colloids Surf A Physicochem Eng Asp 353:64
62. Hai-Ying W, Yang Y, Xiao-Feng L, Ce W (2006) Chem J Chinese Univ 27:1785
63. Dong F, Li Z, Huang H, Yang F, Zheng W, Wang C (2007) Mater Lett 61:2556
64. Anka FH, Perera SD, Ratanatawanate C, Balkus KJ (2012) Mater Lett 75:12
65. Caruso RA, Schattka JH, Greiner A (2001) Adv Mater 13:1577
66. Patel AC, Li S, Wang C, Zhang W, Wei Y (2007) Chem Mater 19:1231
67. Xu X, Yang L, Xu X, Wang X, Chen X, Liang Q, Zeng J, Jing X (2005) J Control Release 108:33
68. Qi H, Hu P, Xu J, Wang A (2006) Biomacromolecules 7:2327
69. Xu X, Zhuang X, Chen X, Wang X, Yang L, Jing X (2006) Macromol Rapid Commun 27:1637
70. Lu X, Wang C, Wei Y (2009) Small 5:2349
71. Wong WS, Salleo A (2009) Flexible electronics: materials and applications. Springer Science & Business Media, New York, NY
72. Horowitz G (1998) Adv Mater 10:365
73. Barbe DF, Westgate CR (1970) J Phys Chem Solids 31:2679
74. Chang J, Liu Y, Heo K, Lee BY, Lee S-WW, Lin L (2014) Small 10:1920
75. Tsumura A, Koezuka H, Ando T (1988) Synth Met 25:11
76. Gonzalez R, Pinto NJ (2005) Synth Met 151:275
77. Chang H-C, Liu C-L, Chen W-C (2013) Adv Funct Mater 23:4960
78. Jian P-Z, Chiu Y-C, Sun H-S, Chen T-Y, Chen W-C, Tung S-H (2014) ACS Appl Mater Interfaces 6:5506
79. Lin Y-W, Lin C-J, Chou Y-H, Liu C-L, Chang H-C, Chen W-C (2013) J Mater Chem C 1:5336

80. Wu H, Hu L, Rowell MW, Kong D, Cha JJ, McDonough JR, Zhu J, Yang Y, McGehee MD, Cui Y (2010) Nano Lett 10:4242
81. Fu W, Liu L, Jiang K, Li Q, Fan S (2010) Carbon N Y 48:1876
82. Chien Y-M, Lefevre F, Shih I, Izquierdo R (2010) Nanotechnology 21:134020
83. Wang X, Zhi L, Muellen K (2008) Nano Lett 8:323
84. Yamamoto N, Makino H, Yamada T, Hirashima Y, Iwaoka H, Ito T, Ujihara A, Hokari H, Morita H, Yamamoto T (2010) J Electrochem Soc 157:J13
85. Shirakawa H, Louis EJ, MacDiarmid AG, Chiang CK, Heeger AJ (1977) J Chem Soc Chem Commun 578. http://pubs.rsc.org/en/content/articlehtml/1977/c3/c39770000578
86. Hsu P-C, Kong D, Wang S, Wang H, Welch AJ, Wu H, Cui Y (2014) J Am Chem Soc 136:10593
87. Hyeon JY, Choi J-M, Park Y, Kang J, Sok J (2013) Appl Sci Converg Technol 22:156
88. Lee Y-I, Choa Y-H (2012) Korean J Mater Res 22:562
89. Wu H, Kong D, Ruan Z, Hsu P-C, Wang S, Yu Z, Carney TJ, Hu L, Fan S, Cui Y (2013) Nat Nanotechnol 8:421
90. Fuh Y-K, Lien L-C (2013) Nanotechnology 24:055301
91. He T, Xie A, Reneker DH, Zhu Y (2014) ACS Nano 8:4782
92. Hsu P-C, Wu H, Carney TJ, McDowell MT, Yang Y, Garnett EC, Li M, Hu L, Cui Y (2012) ACS Nano 6:5150
93. Peng S, Zhu P, Wu Y, Mhaisalkar SG, Ramakrishna S (2012) RSC Adv 2:652
94. Hammock ML, Chortos A, Tee BC-K, Tok JB-H, Bao Z (2013) Adv Mater 25:5997
95. Senthamizhan A, Celebioglu A, Uyar T (2014) J Mater Chem A 2:12717
96. Zheng W, Lu X, Wang W, Li Z, Zhang H, Wang Z, Xu X, Li S, Wang C, Colloid Interface J (2009) Science 338:366
97. Ji L, Medford AJ, Zhang X (2009) Polymer 50:605
98. Wang B, Luo L, Ding Y, Zhao D, Zhang Q (2012) Collioids Surf B 97:51
99. Miao Y-E, He S, Zhong Y, Yang Z, Tjiu WW, Liu T (2013) Electrochim Acta 99:117
100. Chang F-K (2013) Structural health monitoring 2013: a roadmap to intelligent structures. In: Proceedings of the ninth international workshop on structural health monitoring, DEStech Publications, Inc, Palo Alto, CA 10–12 Sept 2013
101. Liu N, Fang G, Wan J, Zhou H, Long H, Zhao X (2011) J Mater Chem 21:18962
102. Sun B, Long Y-Z, Liu S-L, Huang Y-Y, Ma J, Zhang H-D, Shen G, Xu S (2013) Nanoscale 5:7041
103. Rong H, Yunze L, Chengchun T, Hongdi Z (2014) Adv Mater Res 853:79
104. Dhakras D, Borkar V, Ogale S, Jog J (2012) Nanoscale 4:752
105. Li C, Chartuprayoon N, Bosze W, Low K, Lee KH, Nam J, Myung NV (2014) Electroanalysis 26:711
106. Darbandi SMA, Nouri M, Mokhtari J (2012) Fibers Polym 13:1126
107. Tiwari MK, Yarin AL, Megaridis CM, Yarin AL (2008) Electrospun Nanocomposites as Flexible Sensors. In: Proceedings of the ASME international manufacturing science and engineering conference, Evanston, IL, 2008, vol 2, p 281
108. Anitha S, Brabu B, Rajesh KP, Natarajan TS (2013) Mater Lett 92:417
109. Min X, Xiaoxu W, Yong Z, Zhengtao Z, Hao F (2014) Appl Phys Lett 104:133102
110. Yue X-J, Hong T-S, Xu X, Li Z (2011) Chinese Phys Lett 28:090701
111. Merlini C, Barra GMO, Araujo TM, Pegoretti A (2014) RSC Adv 4:15749
112. Cai J, Chawla S, Naraghi M (2014) Carbon N Y 77:738
113. Liu Y, Teng H, Hou H, You T (2009) Biosens Bioelectron 24:3329
114. He Y, Zhang T, Zheng W, Wang R, Liu X, Xia Y, Zhao J (2010) Sens Actuators B Chem 146:98
115. Zhang L, Wang X, Zhao Y, Zhu Z, Fong H (2012) Mater Lett 68:133
116. Yang D-J, Kamienchick I, Youn DY, Rothschild A, Kim I-D (2010) Adv Funct Mater 20:4258
117. Zhang Z, Li X, Wang C, Wei L, Liu Y, Shao C (2009) J Phys Chem C 113:19397

118. Choi J-K, Hwang I-S, Kim S-J, Park J-S, Park S-S, Jeong U, Kang YC, Lee J-H (2010) Sens Actuators B Chem 150:191
119. Lim SK, Hwang S-H, Chang D, Kim S (2010) Sens Actuators B Chem 149:28
120. Wu W-Y, Ting J-M, Huang P-J (2009) Nanoscale Res Lett 4:513
121. Landau O, Rothschild A, Zussman E (2009) Chem Mater 21:9
122. Cao M, Wang Y, Chen T, Antonietti M, Niederberger M (2008) Chem Mater 20:5781
123. Bai S, Chen S, Zhao Y, Guo T, Luo R, Li D, Chen A (2014) J Mater Chem A 2:16697
124. Zhang Y, Wang Y, Jia J, Wang J (2012) Sens Actuators B Chem 171:580
125. Huang S, Matsubara K, Cheng J, Li H, Pan W (2013) Appl Phys Lett 103:141108
126. Wang Z, Li Z, Jiang T, Xu X, Wang C (2013) ACS Appl Mater Interfaces 5:2013
127. Horzum N, Tascioglu D, Okur S, Demir MM (2011) Talanta 85:1105
128. Park M, Im J, Park J, Jeong U (2013) ACS Appl Mater Interfaces 5:8766
129. Lin D-P, He H-W, Huang Y-Y, Han W-P, Yu G-F, Yan X, Long Y-Z, Xia L-H (2014) J Mater Chem C 2:8962
130. Niu H, Wang H, Zhou H, Lin T (2014) RSC Adv 4:11782
131. Pinto NJ, Rivera D, Melendez A, Ramos I, Lim JH, Johnson ATC (2011) Sens Actuators B Chem 156:849
132. Viter R, Chaaya AA, Iatsunskyi I, Nowaczyk G, Kovalevskis K, Erts D, Miele P, Smyntyna V, Bechelany M (2015) Nanotechnology 26:105501
133. Luo Y, Nartker S, Miller H, Hochhalter D, Wiederoder M, Wiederoder S, Setterington E, Drzal LT, Alocilja EC (2010) Biosens Bioelectron 26:1612
134. Zampetti E, Pantalei S, Scalese S, Bearzotti A, De Cesare F, Spinella C, Macagnano A (2011) Biosens Bioelectron 26:2460
135. Mercante LA, Pavinatto A, Iwaki LEO, Scagion VP, Zucolotto V, Oliveira ON Jr, Mattoso LHC, Correa DS (2015) ACS Appl Mater Interfaces 7:4784
136. Kerr-Phillips T, Srinivas ARG, Travas-Sejdic J (2014) Int J Nanotechnol 11:626
137. Zheng Y, Cheng L, Yuan M, Wang Z, Zhang L, Qin Y, Jing T (2014) Nanoscale 6:7842
138. Khatri Z, Ali S, Khatri I, Mayakrishnan G, Kim SH, Kim I-S (2015) Appl Surf Sci 342:64
139. Chaaya AA, Bechelany M, Balme S, Miele P (2014) J Mater Chem A 2:20650
140. Lin Q, Li Y, Yang M (2012) Sens Actuators B Chem 161:967
141. Li P, Li Y, Ying B, Yang M (2009) Sens Actuators B Chem 141:390
142. Corres JM, Garcia YR, Arregui FJ, Matias IR (2011) Sens J IEEE 11:2383
143. Mandal D, Yoon S, Kim KJ (2011) Macromol Rapid Commun 32:831
144. Merlini C, Almeida RDS, D'Avila MA, Schreiner WH, de Oliveira Barra GM (2014) Mat Sci Eng B Solid 179:52
145. Ding B, Wang M, Yu J, Sun G (2009) Sensors 9:1609
146. Huang R, Long Y, Tang C, Zhang H, Rong H, Yunze L, Chengchun T, Hongdi Z (2014) Adv Mater Res 853:79
147. Wu J, Yin F (2013) J Electroanal Chem 694:1
148. Chen D, Guo X, Wang Z, Wang P, Chen Y, Lin L (2011) Polyaniline nanofiber gas sensors by direct-write electrospinning. In: IEEE 24th international conference on micro electro mechanical systems (MEMS), Cancun, 23–27 Jan 2011, pp 1369–1372
149. Saetia K, Schnorr JM, Mannarino MM, Kim SY, Rutledge GC, Swager TM, Hammond PT (2014) Adv Funct Mater 24:492
150. Xi M, Wang X, Zhao Y, Zhu Z, Fong H (2014) Appl Phys Lett 104:133102
151. Macfarlane AM (2007) Elements 3:165
152. Swider M (2015) Apple Watch battery size half as big as top Android Wear watch. http://www.techradar.com/news/wearables/apple-watch-battery-size-mah-1291964. Accessed Oct 2015
153. Sande S (2015) New iPad battery has 70% more capacity. http://www.engadget.com/2012/03/09/new-ipad-battery-has-70-more-capacity/. Accessed: Oct 2015

154. Eaton K (2015) One reason apple may make a bigger iPhone: battery size. http://www.fastcompany.com/3006913/tech-forecast/one-reason-apple-may-make-bigger-iphone-battery-size. Accessed: Oct 2015
155. Battisti G, Giulietti M (2015) Tesla is betting on solar, not just batteries. https://hbr.org/2015/07/tesla-is-betting-on-solar-not-just-batteries. Accessed: Oct 2015
156. Wang U (2015) Tesla CTO: batteries + solar will lead to cheap electricity within 10 years. http://www.forbes.com/sites/uciliawang/2015/07/14/tesla-cto-batteries-solar-will-lead-to-cheap-electricity-within-10-years/. Accessed Oct 2015
157. Yuan Z, Chen B (2012) AIChE J 58:3370
158. Hu L, Wu H, La Mantia F, Yang Y, Cui Y (2010) ACS Nano 4:5843
159. Hu L, La Mantia F, Wu H, Xie X, McDonough J, Pasta M, Cui Y (2011) Adv Energy Mater 1:1012
160. Yang Y, Jeong S, Hu L, Wu H, Lee SW, Cui Y (2011) Proc Natl Acad Sci U S A 108:13013
161. Tarascon JM, Armand M (2001) Nature 414:359
162. Wu N, Jing B, Cao Q, Wang X, Kuang H, Wang Q (2012) J Appl Polym Sci 125:2556
163. Zhou L, Wu N, Cao Q, Jing B, Wang X, Wang Q, Kuang H (2013) Solid State Ionics 249:93
164. Lee J, Jo C, Park B, Hwang W, Lee HI, Yoon S, Lee J (2014) Nanoscale 6:10147
165. Zhang X, Ji L, Toprakci O, Liang Y, Alcoutlabi M (2011) Polym Rev 51:239
166. Ji L, Yao Y, Toprakci O, Lin Z, Liang Y, Shi Q, Medford AJ, Millns CR, Zhang X, Power J (2010) Sources 195:2050
167. Hu L, Wu H, Gao Y, Cao A, Li H, McDough J, Xie X, Zhou M, Cui Y (2011) Adv Energy Mater 1:523
168. Teh PF, Pramana SS, Sharma Y, Ko YW, Madhavi S (2013) ACS Appl Mater Interfaces 5:5461
169. Mizuno Y, Hosono E, Saito T, Okubo M, Nishio-Hamane D, Oh-ishi K, Kudo T, Zhou H (2012) J Phys Chem C 116:10774
170. Yuan T, Zhao B, Cai R, Zhou Y, Shao Z (2011) J Mater Chem 21:15041
171. Liu Z, Jiang W, Kong Q, Zhang C, Han P, Wang X, Yao J, Cui G (2013) Macromol Mater Eng 298:806
172. Xu Y, Zhu Y, Han F, Luo C, Wang C (2015) Adv Energy Mater 5:1400753
173. Wang G, Zhang L, Zhang J (2012) Chem Soc Rev 41:797
174. Kim B-H, Yang KS, Woo H-G (2011) Electrochem Commun 13:1042
175. Zhang F, Yuan C, Zhu J, Wang J, Zhang X, Lou XWD (2013) Adv Funct Mater 23:3909
176. Kim C, Choi YO, Lee WJ, Yang KS (2004) Electrochim Acta 50:883
177. Wang Y-S, Li S-M, Hsiao S-T, Liao W-H, Chen P-H, Yang S-Y, Tien H-W, Ma C-CM, Hu C-C (2014) Carbon N Y 73:87
178. Lai C, Zhou Z, Zhang L, Wang X, Zhou Q, Zhao Y, Wang Y, Wu X-F, Zhu Z, Fong H, Power J (2014) Sources 247:134
179. Chen X, Zhao B, Cai Y, Tade MO, Shao Z (2013) Nanoscale 5:12589
180. Lee P-C, Han T-H, Hwang T, Oh J-S, Kim S-J, Kim B-W, Lee Y, Choi HR, Jeoung SK, Yoo SE, Nam J-D, Pyoung-Chan L, Tai-Hoon H, Taeseon H, Joon-Suk O, Se-Joon K, Byung-Woo K, Youngkwan L, Hyouk RC, Sun KJ, Seung EY, Jae-Do N (2012) J Memb Sci 409–410:365
181. Li X, Wang G, Wang X, Li X, Ji J (2013) J Mater Chem A 1:10103
182. Joo YL, Wiesner U, Park JH. Ordered porous nanofibers, methods, and applications: U.S. Patent Application 14/378,249[P]. 2013-2-14

Chapter 13
Urine Microchip Sensing System

Ching-Hsing Luo, Mei-Jywan Syu, Shu-Chu Shiesh, Shin-Chi Lai, Wei-Jhe Ma, Yi-Hsiang Juan, and Wen-Ho Juang

Abstract The use of biosensors in intelligent electronics is a hot and popular topic. Various professionals, multidisciplinary, and cross-domain integrations in terms of chemical engineering, electrical engineering, medicine, industrial design, and manufacturers have been developed in recent years. In this chapter, a urine sensing system is introduced. The development of a urine detection device is mainly aimed at the essential indexes of chronic kidney diseases for homecare. Renal failure and complications include acute or chronic urinary tract obstruction, hepatic failure, and nephritic syndrome. At present, more than 2000 people per million worldwide have to rely on hemodialysis. The majority of patients with end stage renal disease have limited mobility, and patients must go to the hospital for diagnosis. Traditional urine detection requires a great deal of duty time prior to a patient obtaining the report. Therefore, this urine sensing system is expected to achieve fast detection and lower cost for patients.

Keywords Urinary creatinine • glomerular filtration rate (GFR) • chronic kidney disease (CKD) • urine albumin-to-creatinine ratio (UACR) • Microalbumin • Bio-electrochemical sensors • Microchip sensing system • Urine • Potentiostat •

C.-H. Luo (✉)
Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan

Institute of Medical Science and Technology, National Sun Yat-Sen University, Kaohsiung 804, Taiwan
e-mail: robin@ee.ncku.edu.tw

M.-J. Syu
Department of Chemical Engineering, National Cheng Kung University, Tainan 701, Taiwan

S.-C. Shiesh
Department of Medical Laboratory Science and Biotechnology, National Cheng Kung University, Tainan 701, Taiwan

S.-C. Lai
Department of Computer Science and Information Engineering, Nan Hua University, Chiayi 622, Taiwan

W.-J. Ma • Y.-H. Juan • W.-H. Juang
Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_13

Front-end readout circuit • Successive approximation ADC (analog-to-digital converter) • Mixed-signal IC design

The rest of this chapter is organized as follows: Sect. 13.1 introduces various bio-electrochemical sensors for detecting albumin and creatinine on a micro-sensing array chip for chronic kidney disease (CKD). Section 13.2 focuses on a bio-electrochemical acquisition device for measurement and testing with the electrochemical sensors. The device includes front-end readout circuits, an analog-to-digital converter, and microcontroller. Furthermore, Section 13.3 discusses clinic validation of the proposed sensing system using various analyses and evaluations and provides a discussion of clinical practicalities.

The urine detection system design is presented as the smart sensor and system shown in Fig. 13.1. The system includes bio-electrochemical sensors and a readout circuit system. The former senses the electrochemical signal with a high-accuracy design for creatinine and albumin. The latter reads the sensing signals and converts them into the chemical concentrations displayed on a PC or a mobile device. In addition, clinical validation is required to testify the accuracy and reliability of the proposed urine microchip sensing system via hospital-level analysis technologies.

The proposed system reduces the cost of sensing and makes homecare detection of CKD realistic. Patients can determine their renal performance on a daily basis to avoid their kidney(s) getting worse or irreversibly destroyed. Thus, this system not only reduces consumption of medical resources but also reduces the burden on the medical-care system.

This chapter is organized as follows: Sect. 13.1.1 describes the design of the biosensors and their characteristics, such as the use of creatinine and albumin for the CKD index. Section 13.1.2 introduces the front-end readout circuit, the analog-to-digital (ADC) design, the micro-control unit (MCU), and the mixed-signal IC design. It briefly reviews several readout circuit structures and describes the basic ADC structure and operation. In addition, the MCU design and mixed-signal IC

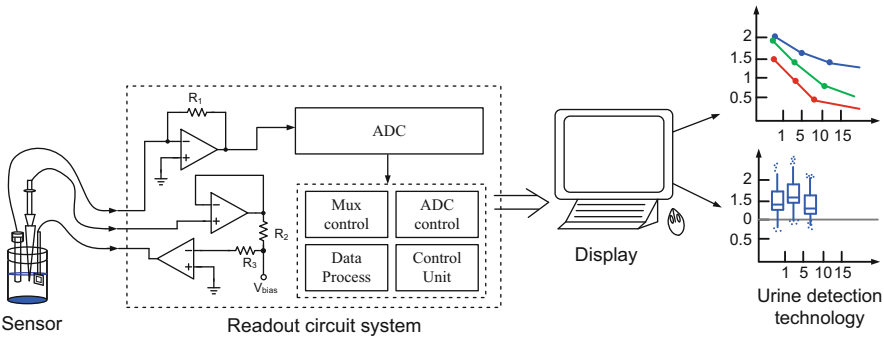


Fig. 13.1 Urine microchip sensing system

design are both introduced. Section 13.1.3 provides the clinical validation for CKD, estimation of the glomerular filtration rate (eGFR), and the urine albumin-to-creatinine ratio (UACR).

13.1 Biosensing: Creatinine and Albumin

Biosensors have been investigated for several decades [1]. There are a variety of reaction mechanisms, adsorption mechanisms, transducers, microelectromechanical systems (MEMS) devices, and electric circuit designs comprising the modules of a biosensor. Traditionally, it is the immobilized enzyme(s) that are utilized for biosensors because of the specificity of enzymes towards unique substrates. However, there have already been other methods proposed due to the fact that immobilized enzymes are more expensive and exhibit shorter-term stability compared to these other approaches. The signal transducer can include (1) electrochemicals such as amperometric, potentiometric, capacitance, and impedance, (2) a field-effect transistor (FET), (3) optical fiber, (4) a piezoelectric quartz crystal, (5) fluorescence, (6) surface-enhanced Raman scattering, and (7) surface plasmon resonance (SPR), etc.

In this session, the content of urine that is frequently highly correlated to kidney, liver, and related diseases is introduced. Particularly, the albumin/creatinine ratio (ACR, μg/mg) instead of urinary creatinine (or CCR, creatinine clearance rate of 24 h) or urinary albumin concentration alone is becoming accepted as a more reliable index for the detection of microalbuminuria [2].

13.1.1 Urinary Creatinine

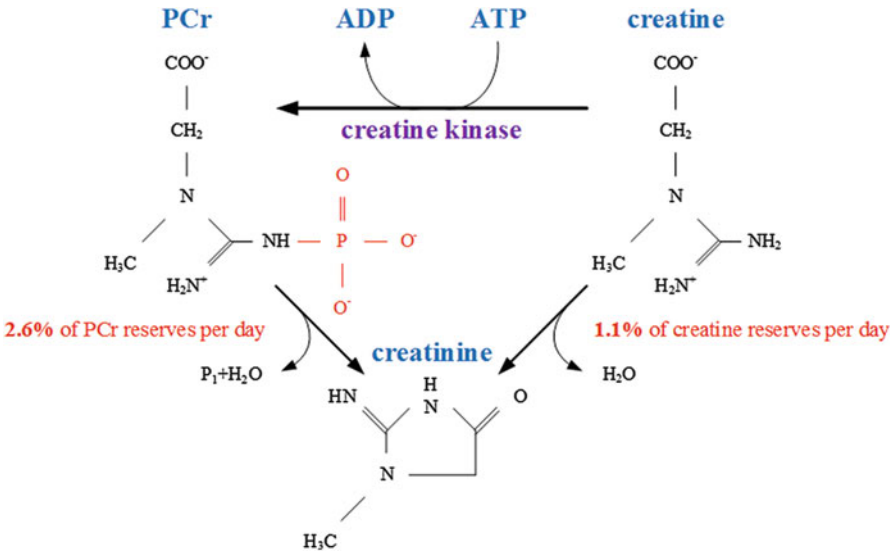
The measurement of serum creatinine is useful for kidney function evaluations, such as during kidney failure, urinary tract obstruction, kidney infection, or damage to the kidney system. However, creatinine concentration in urine varies from 500 to 2000 mg/d. The values also depend on the sex, age, and weight. The most common measurement is CCR, which is an assessment of glomerular filtration rate (GFR). CCR is calculated from a ratio of urine creatinine to serum creatinine concentration with urine volume. CCR requires a timed urine collection (usually 24-h), which is very inconvenient and may have the error of incomplete urine collection. The eGFR is a calculation using serum creatinine values in combined with other parameters. When eGFR <60 mL/min/1.73 m^2 for more than 3 months, it indicates CKD. Several formulae are used for the calculation, such as Cockcroft-Gault equation, modification of diet in renal disease-simplified GFR (MDRD-S-GFR), and CKD-EPI equation.

The other indicator for CKD is the excretion of urine albumin. Instead of the CCR for 24-h collection of urine, UACR can be used to estimate excretion of urine

albumin using first morning urine. This avoids the inconvenience caused from 24-h timely collection of urine. ACR, calculated from the ratio of urine albumin to urine creatinine, could be a more efficient CKD indicator for kidney function. When the UACR value is greater than 30 mg/g but less than 300 mg/g, this indicates microalbuminuria. Consequently, both urine albumin and urine creatinine are important measurements regarding the above urinary diseases and a point-of-care testing must be established for the early diagnosis and better prognosis.

13.1.2 Measurement of Creatinine Concentration

Creatinine is a metabolite from creatine generated from the catalysis of creatine kinase (CK) with the substrate of creatine phosphate (PCr). The brief reaction can be expressed as below.



The most common method to measure creatinine concentration is using the Jaffé reaction [3–5]. Jaffé observed that when mixing picric acid with creatinine under alkali conditions, a red-orange creatinine–picrate complex forms. Thus, the reaction, and therefore the creatinine concentration, can be measured by a spectrophotometer. However, the precision of this method can be interrupted by the other metabolites.

Other than the Jaffé method, creatinine can also be detected using the multi-enzyme complex or a single enzyme with or without immobilization measuring for the catalyzed hydrolysis of creatinine [6]. The two kinds of catalyzed reactions are described as below.

Some companies provide test strips for the measurement of creatinine level. Certain companies offer a creatinine assay kit that uses an enzyme mixture of creatininase and creatinase. The required volume of samples is 2–50 μL, and the incubation time is 1 h.

1. Using the enzyme complex of creatinine amidohydrolase (or creatininase), creatine amidino-hydrolase (or creatinase), and sarcosine oxidase (SO) [7].
2.

$$\text{Creatinine} + H_2O \xleftrightarrow{\text{creatininase}} \text{Creatine}$$

$$\text{Creatine} + H_2O \xleftrightarrow{\text{creatinase}} \text{Sarcosine} + \text{urea}$$

$$\text{Sarcosine} + O_2 \xleftrightarrow{\text{sarcorsine oxidase}} \text{Formaldehyde} + \text{Glycine} + H_2O_2$$

Using an enzyme complex makes the creatinine sensor quite expensive. The amperometric mode is used to measure the end product H_2O_2, which is correlated to the concentration of creatinine. Thus, the creatinine concentration can be detected [8]. However, to be able to detect the H_2O_2, a higher potential of around 0.7 V would be required. On the contrary, in such a case, interference, such as the presence of uric acid in the urine, could become more severe.

Meyerhoff and Rechnitz used an enzyme activator, tripolyphosphate, to improve the activity of creatinine iminohydrolase (EC 3.5.4.2.1), thus allowing the creatinine concentration to be calibrated according to the potentiometric change [9].

In addition, there are other similar enzyme mixtures being considered, for instance, creatininase/creatinase is used by companies as the assay kit. With the aid of a coloring agent or a fluorescent dye, variations in the concentration can be obtained by an optical change instead of an electric signal. The equations are expressed as follows:

$$\text{Creatinine} \xrightarrow{\text{creatininase}} \text{Creatine} \xrightarrow{\text{creatinase}} \text{Sarcosine} \xrightarrow{\text{oxidation}} \text{Color/Fluorescence}$$

By a single enzyme, creatinine deiminase (or creatinine iminohydrolase) [10, 11]

$$\text{Creatinine} \xrightarrow{\text{creatinine deiminase}} \textit{N}\text{-methylhydantoin} + OH^- + NH_4^+$$

In this case, the dissolved ammonium ion in the solution is the end product, although few studies have reported the use of ammonia gas as the target analyte. Thus, creatinine concentration can be calibrated against the potentiometric signal, which is caused by the formation of an ammonium ion. The detection of creatinine concentration using the enzymatic reaction can often be achieved with (1) a spectrophotometer or even together with commercial assay kits (the detection is often made by adding a coloring agent with the enzymatic reaction); (2) a test strip; (3) an

immobilized enzyme biochip; or (4) a fluorometer (the detection must be aided with a fluorescent dye).

Jurkiewicz et al. used flow injection analysis together with an immobilized enzyme electrode for the detection of urea and creatinine [12]. Creatinine deiminase was immobilized on controlled-pore glass beads, whereas urease was immobilized on a nylon open tubular reactor. Both urea and creatinine, after the hydrolysis by urease and creatinine deiminase, respectively, released ammonia. Thus, a flow-through ammonium ion-selective electrode with an inner solid graphite-epoxy composite was prepared for the purpose of detection.

The immobilized enzyme chips can also include an enzyme field-effect transistor (ENFET). The creatinine hydrolysis enzyme complex or a single enzyme is coated onto the FET micro-device. Together with an appropriately designed sensing material, the enzyme(s) can be fabricated on the electrode of the FET device. There are many approaches to immobilize the enzyme(s) for the catalyzation of the creatinine hydrolysis reaction. The immobilization methods may also include magnetic nanomaterials or the conduction of a polymer composite to entrap the regarding enzyme(s) [13].

However, the enzyme(s) for the hydrolysis of creatinine are very expensive. Therefore, for decades, other methods have always been investigated with the effort focused on reducing the cost of creatinine sensors (or analysis). Among all of these approaches, molecular imprinting (MI) technology has been applied to create a specific cavity for creatinine. It is accomplished by uniformly mixing monomers, crosslinkers, and creatinine (as the template molecule) in a solvent. Then, by providing an initiator and energy, the molecularly imprinted polymers (MIP) for specific binding of creatinine can be synthesized. Creatinine MIP-related research has been conducted by different researchers from different perspectives [14–27].

Prasad's team reported the use of an MIP-modified hanging mercury drop electrode (HMDE) for the differential pulse, cathodic stripping voltammetric (DPCSV) detection of creatinine concentration [28, 29]. The creatinine sensor was fabricated by the drop coating of a creatinine-imprinted polymer/dimethylformamide (DMF) solution onto the surface of a HMDE. The creatinine was pre-concentrated and oxidized in the MIP layer. The sensor was highly selective for creatinine with extremely little interference. The results confirmed that the imprinted polymer successfully acts as a recognition element of the creatinine sensor. The proposed MIP-modified HMDE can be used to determine creatinine in human serum accurately and rapidly. Lakshmi et al. also reported that poly (*p*-aminobenzoicacid-*co*-1, 2-dichloroethane) film was casted on the surface of a HMDE for the selectivity and sensitivity evaluation of creatine in real samples [30]. Via non-covalent (electrostatic) imprinting, the MIPs could specifically recognize creatine. Creatine concentration could be detected by a differential pulse, cathodic stripping voltammetric signal. The modified sensor was reproducible and selective with a limit of detection of 0.11 ng/mL creatine. The addition of urea, creatinine, and phenylalanine showed no significant binding to the as-prepared MIP film.

Sergeyeva et al. proposed a calorimetric test-system for the measurement of creatinine concentration in aqueous solution. A creatinine-selective MIP layer was fabricated by using the functional monomer of, say, 2-acrylamido-2-methyl-1-propanesulfonic acid, itaconic acid, or methacrylic acid, with *N*, *N* ′-methylenebisacrylamide as a crosslinker [31]. Choice of the functional monomer was based on the results of computational modeling. Photo-initiated grafting polymerization was applied to develop the MIP film onto the surface of a microfiltration polyvinylidene fluoride membrane.

Subrahmanyam et al. proposed a method for the selective detection of creatinine, which was based on the reaction between polymerized hemithioacetal, formed by allyl mercaptan, *o*-phthalic aldehyde, and primary amine leading to the formation of a fluorescent isoindole complex [32]. This method was previously demonstrated for the detection of creatine using imprinted polymers. Other than the traditional methods, the use of a computer simulation to develop a rational design of a MIP selective for creatinine was developed. The functional monomers from the virtual library were assigned and screened against the target molecule, creatinine, using molecular modeling software. The monomers giving the highest binding score were tested by simulated annealing to mimic the complexation of the functional monomers with the template in the mixture. The simulation results gave an optimized MIP composition. The computationally designed polymer demonstrated superior selectivity. The "Bite-and-Switch" approach combined with molecular imprinting can be used for the design of assays and sensors that are selective for amino-containing substances.

Syu's lab also investigated the synthesis of imprinted polymer materials with different functionalities for the specific uptake of creatinine [33–38]. Additionally, they fabricated MIP film for AC impedance measurement of creatinine concentration via which the calibration curves of creatinine concentration both in a serum and urine environment could be established, respectively. Furthermore, serum creatinine and urine creatinine concentrations in clinical samples can thus be obtained by the impedance mode of detection. In the future, together with the AC impedance chip from the IC design, a creatinine-imprinted impedance chip can be assembled and developed.

13.1.3 Urinary Albumin

Up to this stage, there have been a lot of reports on the binding or sensing of serum albumin or the design of its microchip or with MEMS. Nevertheless, there have been very few reports on urinary albumin. There have been a few reports on the sensing of urine albumin. Up to the present, there has been no literature that really deals with the sensing of urine albumin.

Human serum albumin bound to TiO_2, CeO_2, Al_2O_3, and ZnO nanoparticles and SPR have been applied to measure the change after the binding of albumin [39].

Wang et al. [40] synthesized an ionic-liquid composite imprinted polymer by using 3-(3-aminopropyl)-1-vinylimidazolium tetrafluoroborate ionic liquid as the functional monomer, N,N′-methylenebisacrylamide as the crosslinker, and ammonium persulfate and N,N,N′,N′-tetramethylethylenediamine to initiate the polymerization with serum albumin as the template. Thus, using this ionic-liquid imprinted polymer-modified electrode, the current change could be measured against the uptake of serum albumin, and thus the calibration of albumin concentration could be achieved. Earlier than the proposed work of Wang, there were already albumin imprinted polymers being synthesized, and there had been discussion of the binding towards albumin as well.

A one-spot synthesis method was used in a study to load acridine orange onto a gold nanorod-based electrode [41]. The gold nanorods were coated with mesoporous silica. Bovine serum albumin could be detected by the quenching of fluorescence signal upon the uptake of the albumin to the acridine orange modified Au nanorod@SiO_2 electrode.

Park et al. [42] developed a biofield-effect-transistor (BioFET) for detection of albumin in a human urine/phosphate buffer saline (PBS) mixture of albumin. PBS was used to reduce the electrolyte effect of the urine on the FET. Anti-albumin was immobilized on the gate surface of the FET device to become the so-called BioFET. The drain current was modulated by the albumin bound to the immobilized anti-albumin on the BioFET. The current variation ratio was approximately proportional to the albumin concentration in the range 50–250 mg/L. This study shows the feasibility of the BioFET as a urinary albumin sensor.

Additionally, Park et al. designed and developed an albumin BioFET based on the potentiometric measurement of albumin concentration. The gate surface of the BioFET was chemically modified by a self-assembled monolayer (SAM), which was synthesized by thiazole benzo crown ether ethylamine (TBCEA)–thioctic acid to selectively immobilize anti-albumin [43]. Thus, the binding of various albumin concentrations to the anti-albumin/TBCEA-thioctic acid SAM was quantitatively measured by the output voltage changes in the BioFET. Furthermore, a molecularly imprinted chitosan-acrylamide, graphene, ferrocene composite cryogel biosensor was also fabricated to detect microalbumin [44].

Once the layered materials specifically for the sensing of urine creatinine and urine albumin are both fabricated onto a microchip, we are going to develop a biochip for ACR in a valid clinical range.

13.2 Urine Detection System

In the urine detection system, some of the biomedical signal is a potentiometric signal, such as pH. Therefore, we have to use the potential method to deal with this potential signal. In the following section, the basic potential method for a front-end readout circuit, ADC design, and mix-signal design are discussed.

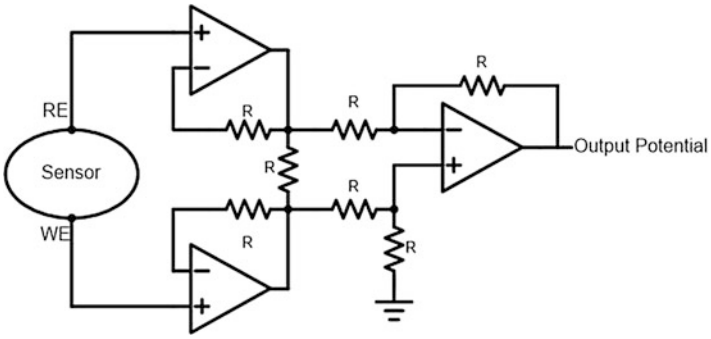


Fig. 13.2 Potentiometric circuit

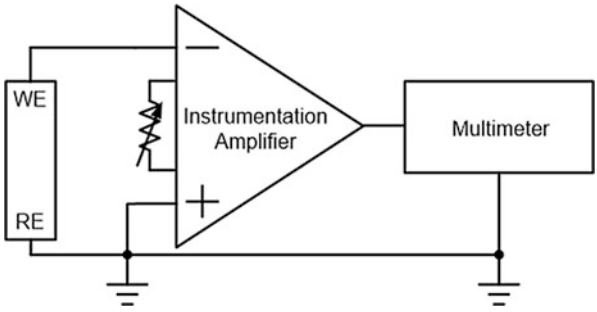


Fig. 13.3 Measurement structure of the pH biosensor

13.2.1 Front-End Readout Circuit

Figure 13.2 shows the general architecture of a potentiometric circuit [45]. The first part is a set of two voltage follower circuits. The input impedance of the operational amplifier is designed to be very high for good isolation of the output from the input signal source. It can avoid loading effects with very little power drawn from the signal source. The second stage of the circuit is an instrumentation amplifier. The advantage of the differential instrumentation amplifier (IA) is that the total gain can be varied by R_G. The other advantages are related to its high input impedance, high common mode rejection ratio, and noise elimination capability.

Figure 13.3 shows the measurement method for pH detection in which an instrumentation amplifier is used [46] and the reference electrode (RE) is a commercial Ag/AgCl electrode.

In the urine detection system, some of the biomedical signal is a current signal, such as nitrite, so we have to use the potentiostat method to deal with this current signal. In this section, the basic potentiostat method is discussed.

Figure 13.4 shows another structure of the potentiostat method. This structure includes a control amplifier, a buffer and an I–V converter [47]. It's worth noting that the current signal of the nitrite is an approximation of nA, so we use a high

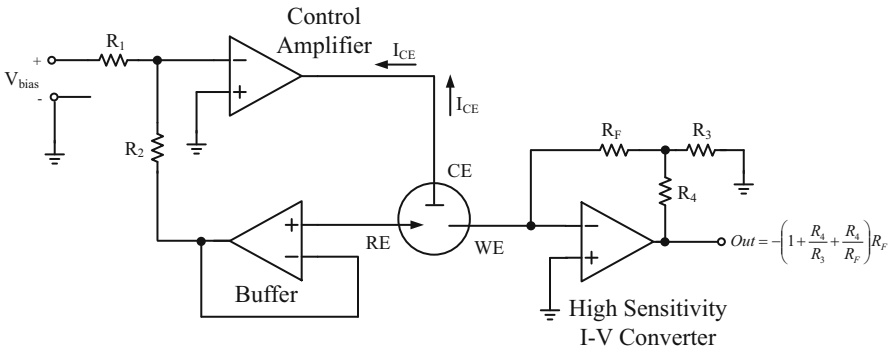


Fig. 13.4 Basic potentiostat structure [47]

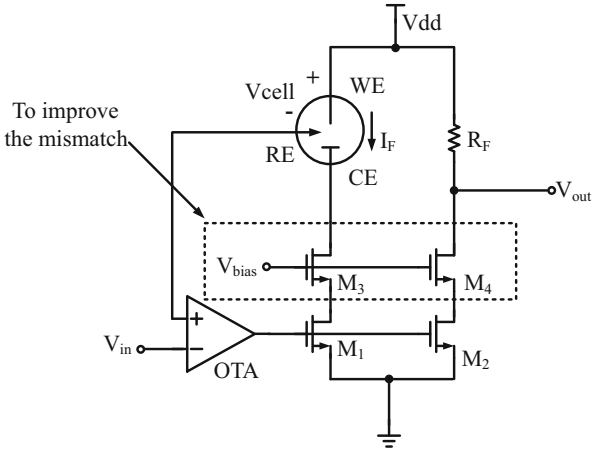


Fig. 13.5 Structure of the current readout circuit

sensitivity I–V converter to replace the traditional transimpedance amplifier. The control amplifier can provide a current path to the sensor, and the buffer is used to reduce the loading effect. In addition, the operation amplifier of the high sensitivity I–V converter offers a virtual ground at the working electrode and generates the output voltage simultaneously.

In Fig. 13.5, the readout circuit with the potentiostat [48] maintains a constant bias potential between the reference and the working electrodes using one operational trans-conductance amplifier with an MOS transistor and a resistor to perform current-to-current conversion. In addition, in order to reduce MOS transistor mismatch effects, the cascode technique is used to improve mismatch problems.

A current mirror [49] is placed in the sensor current path to generate a mirror of the sensing current. In addition, through the current-to-frequency converter, the proposed circuit improves the linearity with calibration data measured in advance. In order to increase the dynamic range and reduce the noise, a fully differential potentiostat [50] is used.

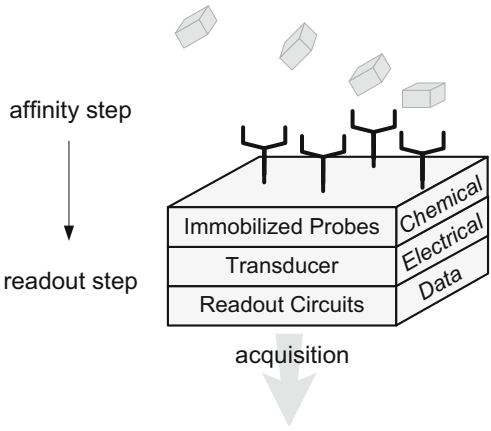


Fig. 13.6 Scheme showing concept of affinity-based biosensor

With the rapid progress of the CMOS process, miniaturization and heterogeneous integration have become the trend in biosensors. The main advantages of miniaturized biosensors include the fact that they are portable, disposable, low power, low cost, and high capacity. Based on these characteristics, biosensors are able to replace bulky instruments operated by professionals. Thus, ease-of-use measurement devices can spread among families to provide self-diagnosis records. Due to its high capacity, the multi-sensing technique integrates biosensors into an array to provide high throughput. Schienle [51] implemented a DNA sensor array chip with 128 positions and an in-pixel A/D conversion which provides a dynamic range of five decades for wide-range applications, and Hassibi [52] proposed a biosensor array for label-free biomolecular detection based on the electrochemical impedance spectroscopy (EIS) method.

To understand what really determines the performance of impedance biosensors, the concept of an affinity-based biosensor is introduced in Fig. 13.6.

Affinity-based biosensors divide the operation into two steps: (1) selecting and binding the target and excluding non-target biomolecules through a chemical reaction using selective probes, which is called the affinity step, and (2) sensing the change at the probe surface, then processing it in an electrical form, including amplification, filtering, and even post-processing (readout step). The affinity step, rather than the readout step, dominates the overall performance of impedance biosensors [53]. The impedance sensing technique is sufficient for many applications because of its low cost, compact, and simple features.

Basically, there are two major processes in the EIS method: the fast Fourier transform (FFT)-based and frequency response analyzer (FRA)-based EIS. These two categories were compared [54], and the latter one was chosen as the better way to achieve compact implementation at the cost of long interrogation time. As for the FFT-based method, the advantage is that all interested frequencies can be computed at the same time, but it requires a quite complicated computation to perform the discrete Fourier transform; thus, it is not suitable for microsystem applications. Previous works [54–57] used the FRA-based method in a bio-impedance circuit

design, tending to develop a microarray system. A 5-channel impedance-to-digital converter (IDC) array was successfully implemented [57] with an on-chip signal generator and miniaturized electrodes. However, these works required two identical reference currents with opposite polarities at the input of the integrator to reset whenever the output of the integrator exceeded the range between two appropriately set voltage levels. Unless a calibration unit is designed, these architectures suffer from severe current mismatch, which may cause a 20 % (or larger) error in the worst case PVT variations. This may cause different measurement results for the same biosensor.

EIS is one of the measurements for measuring the dielectric and transport properties of materials [58]. Using a FRA for measuring impedance has become the de facto standard [59]. FRA realization is based on sending different frequency (e.g., 1 mHz–1 MHz) stimulating signals (sine and cosine wave) into an impedance sensor, collecting frequency responses and analyzing the results. Through this analysis, a sensor impedance model can be built up [58].

Because of changes in lifestyle and dietary habits, Taiwan's CKD incidence rate has become the highest in the world. CKD medical treatment is included in Taiwan's National Health Insurance, which is a huge burden to the medical-care system and to the social welfare system. Therefore, developing methods by which to examine and monitor kidney system function and detect urinary infections is becoming important. Creatinine concentration in urine is an index chemical used to monitor human kidney function. However, it takes time and procedures to determine the creatinine concentration in human urine. In order to acquire the examination results faster and more convenient, an examining or monitoring system circuit is designed to meet the need for low power usage and small size. It is hoped that a research effort towards kidney system function evaluation and urinary infection detection will decrease the CKD incidence rate and contribute to public health and the quality of medical care.

To acquire the creatinine concentration in urine quickly and conveniently, the analog front-end with sigma-delta modulator impedance readout circuit is proposed, as shown in Fig. 13.7. This front-end readout circuit includes an interface circuit for the creatinine sensor, a read and procedure circuit and an analog-to-digital converter (A/D). This design provides a fast and convenient detection circuit.

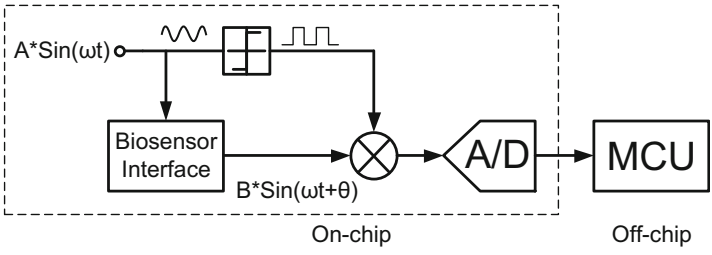


Fig. 13.7 Front-end with sigma-delta modulator impedance readout schematic diagram

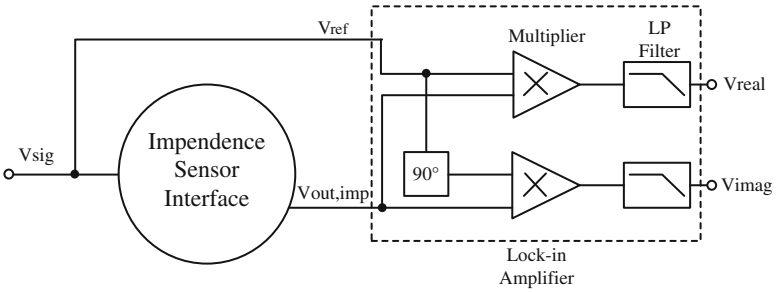


Fig. 13.8 Analog approach method based on lock-in amplifier

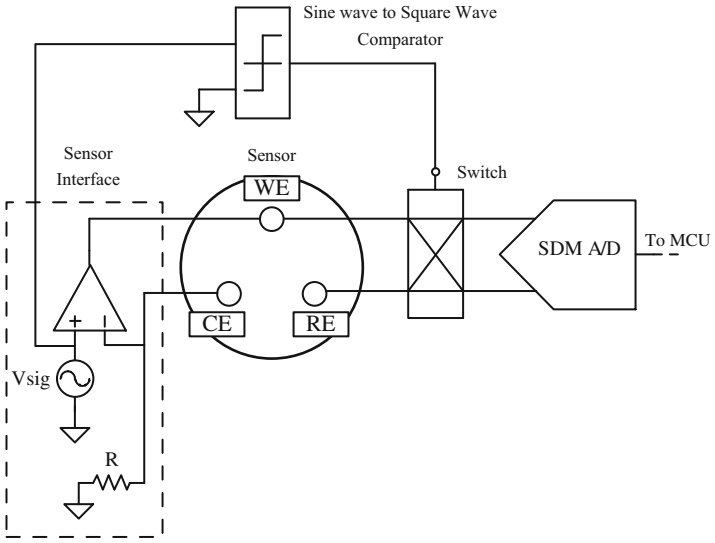


Fig. 13.9 Front-end with sigma-delta modulator impedance readout system implementation

To sense the creatinine concentration in urine, the well-known lock-in amplifier analog approach method is adopted in this design, as shown in Fig. 13.8. By sending a stimulating signal $V_{sig} = A \cdot \sin(\omega t)$ on an impedance sensor, changes will occur based on the creatinine concentration in solution. Because of impedance variations due to different creatinine concentrations, the output of interface circuit will be $V_{out,imp} = B \bullet \sin(\omega t + \theta)$. Multiplying this output $V_{out,imp}$ with V_{ref}, which is used to elucidate the 0° and 90° phases of V_{sig}. After filtering, the impedance (i.e., V_{real}) is extracted using the lock-in amplifier method.

However, some parts of the lock-in amplifier method are changed to propose a new version to extract impedance from this sensor. The system is shown in Fig. 13.9, and the output waveform is shown in Fig. 13.10. The system includes a sensor interface circuit, a sine wave to square wave comparator, a switch circuit,

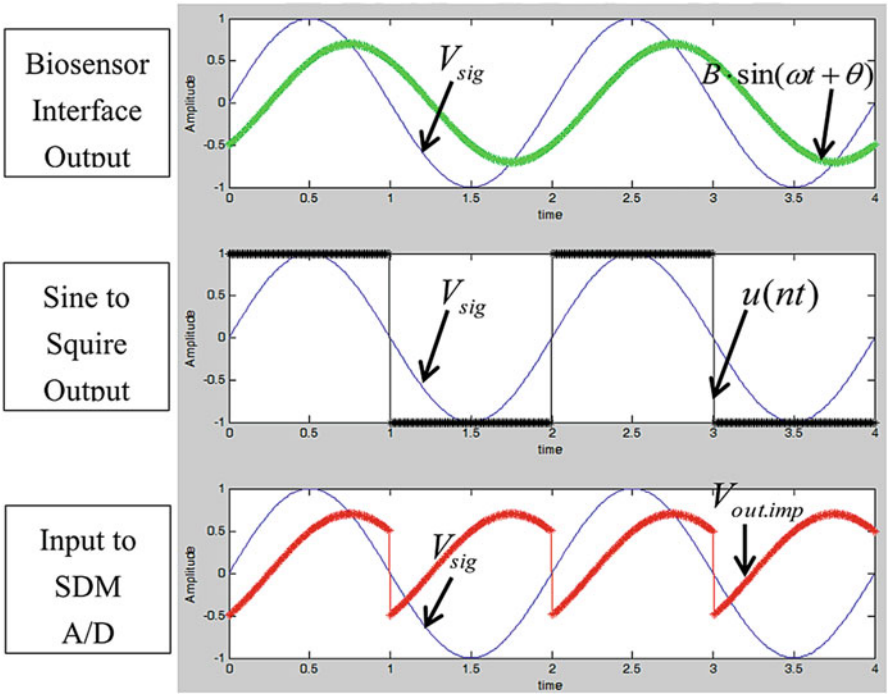


Fig. 13.10 Circuit output waveform from sub-circuits

and an SDM A/D. This new method kicks out the 90° phase of V_{sig} (i.e., V_{imag}) and moves the filter part onto a micro-processor or to a computer after the A/D digital output is complete. The filter algorithm is simplified to reduce hardware requests. This filter algorithm is intended to search for the highest and lowest value from a data waveform and average these two values to determine two parameters that can be calculated for sensor impedance.

13.2.2 Analog-to-Digital Converter

In this section, a successive approximation of analog-to-digital converters (SAR ADCs) is presented for the biomedical system design. Generally, most biomedical systems are characterized by low power consumption, low-to-medium speed (few kHz), and medium-to-high resolution (8–12 bits). For example, in the electrocardiogram application, the desired frequency range is 0–100 Hz with a 10 bit resolution for the ADC [60, 61]. Compared with other ADC structures, the SAR ADC is famous for its medium-to-high resolution and medium speed conversion, which makes it suitable for biomedical system applications, as shown in Fig. 13.11. In

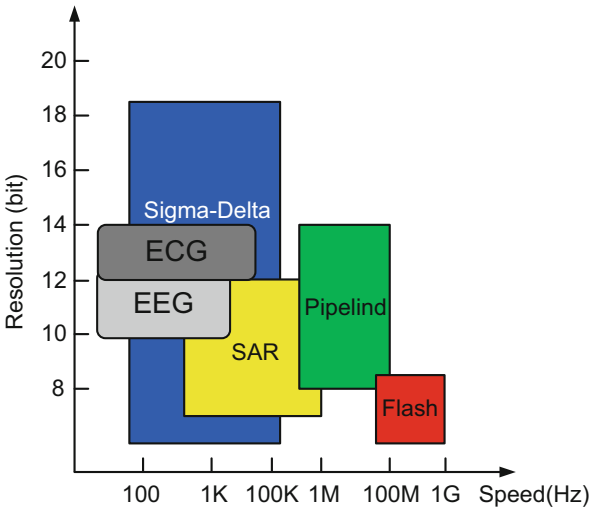


Fig. 13.11 Compared with different structure of the ADCs [64]

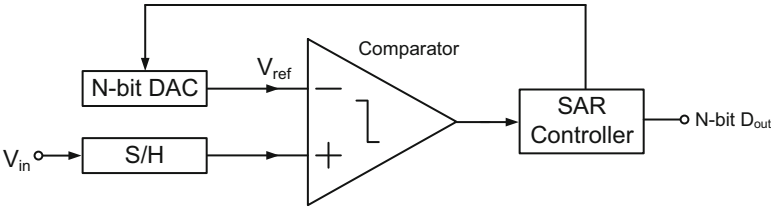


Fig. 13.12 Basic SAR ADC architecture

addition, the SAR ADC has the advantage of low power consumption and a low complexity circuit.

The basic blocks of the SAR ADC include a sample and hold circuit (S/H), an N-bit digital-to-analog (DAC) capacitor array, a comparator circuit and a successive approximation controller (SA), as shown in Fig. 13.12. It is important to note that a fully differential architecture is adopted, which is used to suppress supply noise and to obtain good common-mode noise rejection. In addition, in order to generate better performance in the input signal, a bootstrapped switch is adopted. Because the gate-source voltage of the bootstrapped switch is fixed at supply voltage, which creates small on-resistance value, the linearity of the switch will be improved [62].

To understand the basic operation of the SAR ADC, we have to know the binary search algorithm. For example, consider a random input signal from 1 to 8, where a signal of 4 is used as the basis of comparison. If the signal is greater than 4, the output signal is "1," and the second step is compared with 6. However, if the signal is not greater than 4, the output signal is "0," and the second step is compared with

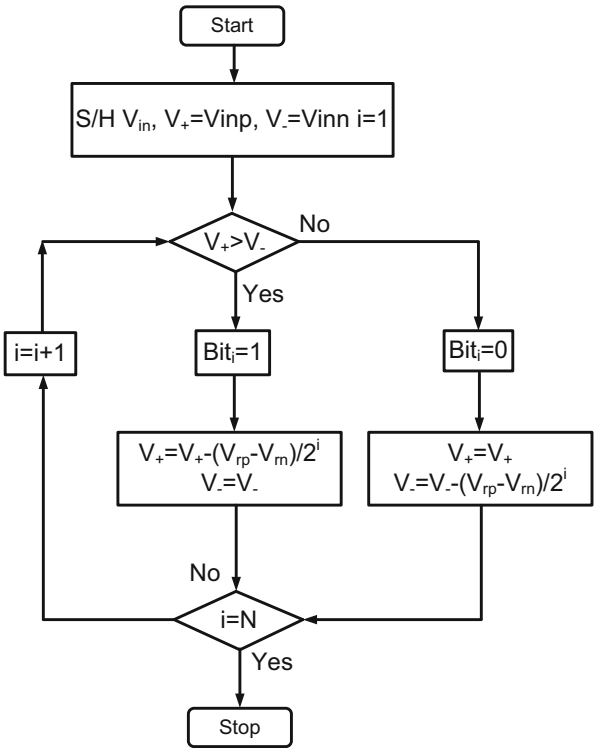


Fig. 13.13 Flow graph of the SAR ADC [64]

2. The third step divides the search space in two once again, and the procedures repeat until the input signal is determined [63].

Figure 13.13 shows the flow graph for the SAR ADC approach. The operation of the SAR ADC can be divided into two parts: sampling and compared modes. During the sampling mode, the SAR ADC samples the input signal on the top plates via the switch, and all the DAC array capacitors are switched to the reference voltage (V_{ref}). During the first compared mode, if V_+ is smaller than V_-, the most significant bit capacitor is switched to V_{refp}. Thus, the voltage of the V_+ is maintained at $V_+ = V_{refp}$, and the voltage of the V_- is Vinn-1/2 ($V_{refn} - V_{refp}$). Then the procedures repeat until the least significant bit (LSB) is decided [64]. In general, an *N*-bit SAR ADC utilizes only one comparator with *N* clock cycles to complete a full conversion and requires the number of unit capacitors in a DAC capacitor array to be 2^N. In addition, when the operation procedure is completed, then the output voltage of the SAR ADC will approximate the reference voltage, such as VDD/2. Another point worth nothing is the voltage result of the comparator is not within ± 0.5 LSB when the circuit completes the operation procedure.

For SAR ADCs, performance limitations are dominated by the DC offset of the comparator, so the switching procedure can be modified to make the output voltage of the DAC approach common-mode voltage (V_{cm}), as shown in Fig. 13.14 [64].

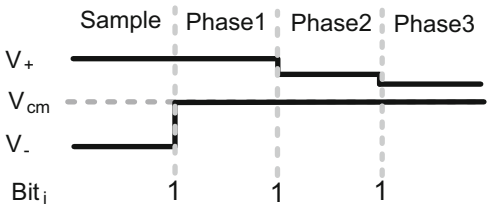


Fig. 13.14 Waveform of the SAR ADC switching procedure [64]

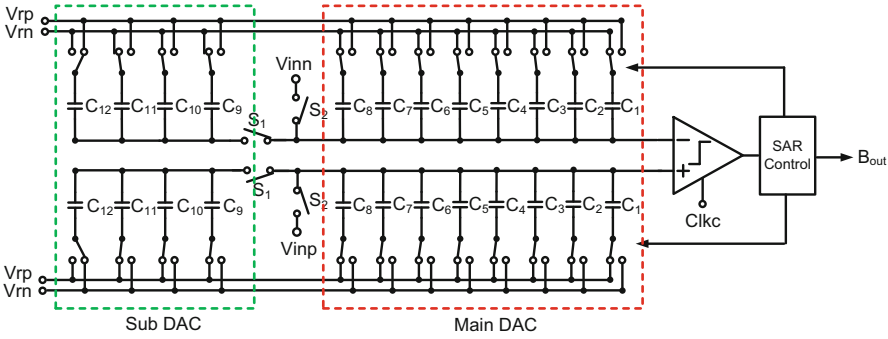


Fig. 13.15 Adaptive SAR ADC structure [64]

The advantage of this switching procedure is that it reduces the variation of V_{GS} and thus decreases the DC offset of the comparator.

Nowadays, for various biomedical signal applications, the operation of the SAR ADC is not only in 8 bit but also in 12 bit resolution. Figure 13.15 shows the adaptive SAR ADC for a biomedical acquisition system [64], where the DAC of this architecture can be separated into a main DAC and a sub DAC for SAR ADC operation in 12 and 8 bit resolution. In addition, energy-efficient switching is adopted to reduce the power consumption of this ADC, and the switching operation is modified to make the output voltage of the DAC to approximate the common-mode voltage in order to reduce the DC offset of the comparator.

13.2.3 Micro-Control Unit

Figure 13.16 is the block diagram of the microcontroller unit, which includes: (1) an analog-to-digital converter (ADC) controller; (2) a digital signal processor, (3) an asynchronous serial communication port, and (4) a central control unit. The ADC controller is designed to control ADC in order to ensure the detected data and provide a clock signal for the ADC. The DSP block is to locate an FIR filter in order to smooth any high-frequency noise. The asynchronous serial communication port can transmit detected data to a computer. The central control unit of the microcontroller is designed with a finite state machine (FSM) with four states.

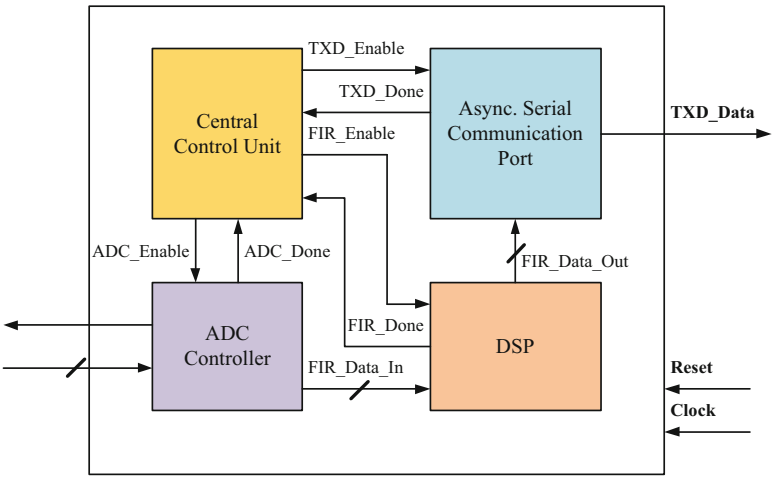


Fig. 13.16 The block diagram and I/O ports of the proposed microcontroller

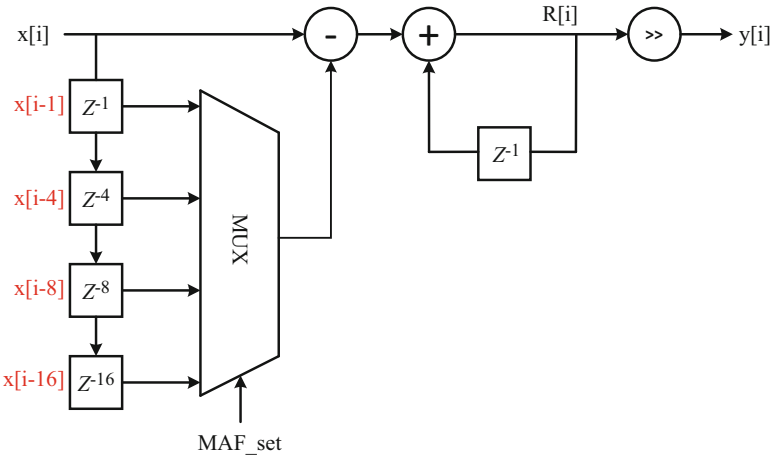


Fig. 13.17 Proposed recursive moving average filter design

The ADC control module is designed to control the SAR ADC, as mentioned above, for the purpose of ensuring the detected data and providing a clock signal for the SAR ADC. The DSP module implements a recursive moving average filter, as shown in Fig. 13.17, that smooths high-frequency noise. The moving average filter operates by averaging a number of points from the input signal to produce each point in the output signal. It only requires an adder and a subtractor. The average points of the moving average filter in this system can be 4 points, 8 points, or 16 points. The UART transmission module can transmit detected data to a computer. The band rate of the system can be 9600, 19200, 115200, or 921600. For the

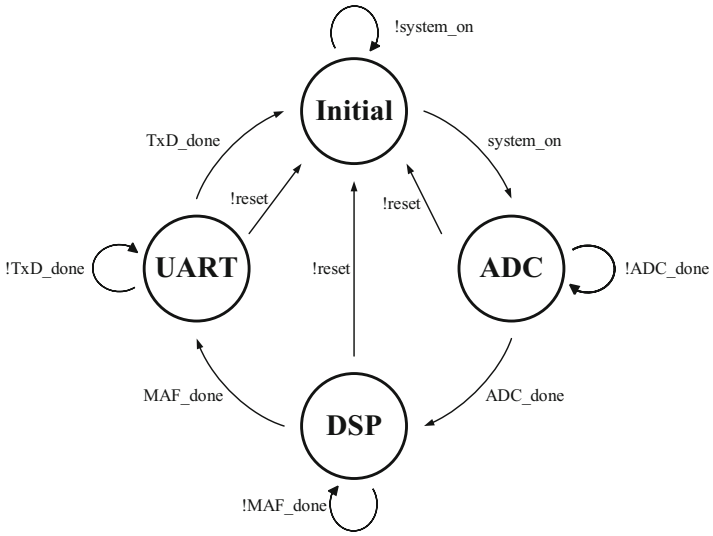


Fig. 13.18 FSM of the proposed controller

central control unit of the microcontroller, there are four states in the proposed FSM which include an initial state, an ADC state, a DSP state, and a UART state. The sampling rate of the whole system ranges from 80 to 40 KHz. The FSM is shown in Fig. 13.18.

In order to implement an SoC, we co-simulate the mixed-signal design on the SAR ADC and microcontroller using Nano Sim Integration with VCS (NIV) for this biomedical sensing system.

13.2.4 Mixed-Signal IC Design

This section illustrates the design flow and simulation methodology for the mixed-signal IC. A mixed-signal IC design implies that analog and digital circuits are all designed in a single chip [65]. In the case of this bio-sensing system, it includes a front-end readout circuit, an analog-to-digital converter, and a MCU. The general mixed-signal design flow can be briefly divided into the following five steps: (1) a cell-based design flow for the digital circuit, (2) a full-custom design flow for the analog circuit, (3) co-simulation for the digital and analog circuits, (4) chip layout, circuit integration, and system function verification for the digital and analog circuits, and (5) chip tape-out and chip measurements.

Full-custom design flow is for the analog circuit. The Netlist circuits after coding and describing assume the process of the pre-layout simulation until the circuit function and specification are satisfied. Then, the processes of chip layout, DRC, LVS, and LPE simulations are all required to occur before the post-layout

simulation. Because the circuit delay is greatly related to the size and connection of the basic components, these simulations are applied to make sure that the timing, power, and area specifications are all met. On the other hand, cell-based design flow is for a digital circuit. The programmer uses the Verilog HDL to code first and then verifies the circuit functions after adopting the synthesis tool. During the synthesis and verification phases, it can be determined whether the circuit delay and other information requirements meet the specifications. After the pre-simulation phase, if the delay and functions generated from the standard cell meet the desired specifications, we can further deal with the issues of placement and routing for the standard cell using the APR tool. Therefore, the timing, power, and area information can be more accurately estimated. Again, a final simulation check is also required to make sure whether or not the circuit function is correct.

For the mixed-signal chip design flow, an important step we call "co-simulation" is used to check the correctness of the interface between the analog and digital circuits in the early design stage and to modify the hidden errors caused by the mixed-signal interface. The common mixed-signal co-simulation and verification ways include: (1) the AMS Environment, (2) the Ultrsim Verilog Environment, and (3) the Nano Sim Integration with the VCS Environment (NIV). In the following introduction, we only focus on the NIV methodology [66, 67]. The method adopts Nano Sim and VCS, which are provided by the Synopsys company, to co-simulate the proposed design. The simulation results are displayed via the Waveform Viewer tool.

In the NIV co-simulation concept, most of the designers will use hierarchical architecture to design their chip. Here, we can regard the analog circuits as high-level modules, where the designer calls the low-level sub-modules such as digital circuits to process the co-simulation (called SPICE-Top), or we can regard the digital circuits as high-level modules, and then call the low-level analog sub-modules to process the co-simulation (called the Verilog-Top). In addition, we use different syntaxes to describe the test pattern for the analog signal when the mixed-signal circuits are co-simulated. Moreover, we adopt the extension syntax of the Verilog hardware description, i.e., Verilog-A [68]. Similar to digital circuit design by Verilog, Verilog-A as a high-level hardware description language is specially proposed for analog circuit design in a Verilog digital design way. The modeling descriptions of analog circuit in Verilog-A include structural description, behavioral description, and hybrid description. In addition, the key parameter values will affect the model accuracy. In the following paragraph, we take an instance of the co-simulation of an analog circuit (ex: ADC) and digital circuit (e.g., MCU) to make a brief illustration.

Figure 13.19 shows the co-simulation procedure with a mixed-signal for the biomedical sensing system using the NIV method, where the digital background circuit for the MCU is defined as an "MCU.v" file, and the foreground analog-to-digital circuit is defined as an "adc.spi" file. The sub-module, adc.spi, requires an extra test pattern while it is in the simulation phase. The simulation tool, i.e., Nano Sim, will link these two Verilog-A files to provide the data for the test pattern and to verify the circuit described by adc.spi in the co-simulation phase. When the simulation is finished, there are two waveform files generated: one is the "adc.

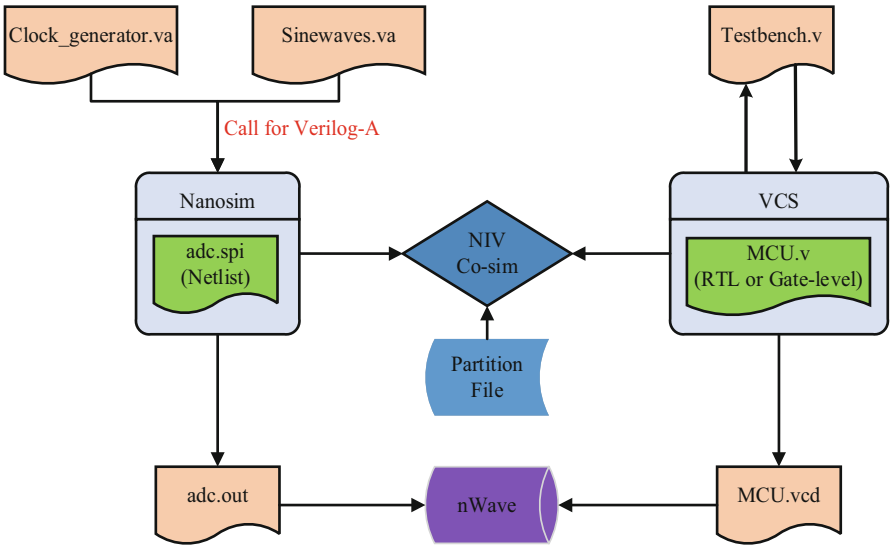


Fig. 13.19 Co-simulation flow for the medical sensing system

out" file and the other is the "MCU.vcd" file. Therefore, we can also use the Waveform Viewer to observe the simulation results and to debug any errors. In addition, we can also command the "Partition File" to merge these two waveform files into a single file, and the file extension is auto set to be labeled "*.uod". This method allows a designer to watch the behavioral results of the analog and digital circuits in the same waveform file.

As mentioned above, the front-end analog circuit and digital MCU can be implemented in a single chip, which is very useful for small size and low power bio-sensing system development.

13.3 Clinical Validation

To validate the accuracy of the proposed homecare urine sensing system close to the hospital level of detection, the clinic validation plays a key role to ensure the correct diagnostic information provided by the proposed system in the commercial market.

13.3.1 Chronic Kidney Disease

CKD has now emerged as a public health burden due to the high risk of progression to end stage renal disease (ESRD). The global prevalence of CKD is more than

10 % and increases with age, exceeding 20 % in subjects older than 60 years old and 35 % in those 70 years or older [69, 70]. CKD is associated with a wide range of causes of increased mortality and morbidity. Therefore, it is critical to have objective measures of kidney damage and function for early identification and treatment of affected patients [71–75]. The current best markers of renal function and staging of renal disease are mainly serum creatinine for the eGFR and urine albumin. In the new KDIGO guidelines [75], urine albumin is as important as eGFR in evaluating the severity of kidney disease. CKD is diagnosed based on low eGFR (<60 mL/min/1.73 m^2) and/or UACR greater than 30 mg albumin/g creatinine. These abnormalities are related to increased risk of kidney failure or to the development of complications, especially cardiovascular disease. Therefore, it is of particular importance to have early detection and management to prevent or delay kidney failure, and ultimately to reduce the health burden associated with renal or extra-renal complications.

13.3.2 Urine Microchip Sensing System

Albuminuria, an increased excretion of urinary albumin, is currently the primary laboratory indicator for kidney damage and also is a risk stratification of kidney disease and cardiovascular disease [74, 76, 77]. Albuminuria is defined in terms of urinary albumin excretion per 24 h. Current guidelines recommend the use of UACR (or ACR, urine albumin (mg/dL) divided by urine creatinine (g/dL) = ACR in mg/g) as a surrogate for the error-prone collection of 24-h urine samples. Testing of ACR is recommended to people with the following risk factors: diabetes, hypertension, acute kidney injury, cardiovascular disease, multi-system disease with potential kidney involvement, family history of ESRD, and opportunistic detection of hematuria. Increased ACR is associated with increased risk of adverse outcomes [74, 76, 77]. In addition, patients with increased albuminuria may benefit from lower blood pressure with medications blocking the renin-angiotensin system.

For the detection of albuminuria, it is recommended to detect and identify proteinuria using urine ACR in preference to protein to **c**reatinine **r**atio (PCR) because ACR has greater sensitivity than PCR for low levels of proteinuria. ACR is the recommended method for people with diabetes to monitor nephropathy complications. However, ACR results may be affected by patient preparation and time of day of sample collection [78]. Currently, the first morning urine is suggested for the testing of ACR. Using a urine microchip sensing system to quantitate albumin and creatinine simultaneously is of great help to identify albuminuria and is also beneficial for prevention and early intervention.

Validation of the analytical and clinical performance characteristics of biosensors is critical for their use in clinical practice [79]. Validation is the assessment of measurement error, including imprecision (within-run, run-to-run and total) and bias by method comparison (in comparison with a reference method or a known bias routine method). Within-run imprecision is determined by quantifying a target

molecule using reference materials and patient urine samples (two-level, with one in normal range and one abnormal). In addition, recovery, interferences, measuring range (linearity), detection limit, and establishment of reference intervals should be performed using the established sensors. Measurement range (or calibration curve) should be carried out using the same matrix as clinical samples, such as urine, to minimize measurement error.

References

1. Harta JP, Crew A, Crouch E, Honeychurch KC, Pemberton RM (2004) Mini-review: some recent designs and developments of screen-printed carbon electrochemical sensors/biosensors for biomedical, environmental, and industrial analyses. Anal Lett 37(5):789–830
2. Mattix HJ, Hsu CY, Shaykevich S, Curhan G (2002) Use of the albumin/creatinine ratio to detect microalbuminuria: implications of sex and race. J Am Soc Nephrol 13:1034–1039
3. Sevillano-cabeza A, Herráez-Hernández R, Campíns-Falcó P (1991) Evaluation and elimination of the interference effects of three cephalosporins on the kinetic-spectrophotometric determination of creatinine in serum using the Jaffé reaction. Anal Lett 24:1741–1766
4. Dsakai T, Ohta H, Ohno N, Imai J (1995) Routine assay of creatinine in newborn baby urine by spectrophotometric flow-injection analysis. Anal Chim Acta 308:446–450
5. McClatchey KD (2002) Clinical laboratory medicine, 2nd edn. Lippincott Williams & Wilkins, Philadelphia, PA
6. Mohabbati-Kalejahi E, Azimirad V, Bahrami M, Ganbari A (2012) A review on creatinine measurement techniques. Talanta 97:1–8
7. Tsuchida T, Yoda K (1983) Multi-enzyme membrane electrodes for determination of creatinine and creatine in serum. Clin Chem 29:51–55
8. Km EJ, Haruyama T, Yanagida Y, Kobatake E, Aizawa M (1999) Disposable creatinine sensor based on thick-film hydrogen peroxide electrode system. Anal Chim Acta 394:225–231
9. Meyerhoff M, Rechnitz GA (1976) An activated enzyme electrode for creatinine. Anal Chim Acta 85:277–285
10. Shih YT, Huang HJ (1999) A creatinine deiminase modified polyaniline electrode for creatinine analysis. Anal Chim Acta 392:143–150
11. Huang CJ, Lin JL, Chen PH, Syu MJ, Lee GB (2011) A multi-functional electrochemical sensing system using microfluid technology for detection of urea and creatinine. Electrophoresis 32(8):931–938
12. Jurkiewicz M, Alegret S, Almirall J, García M, Fàbregas E (1998) Development of a biparametric bioanalyser for creatinine and urea. Validation of the determination of biochemical parameters associated with hemodialysis. Analyst 123:1321–1327
13. Wen TT, Zhu WY, Xue C, Wu JH, Han Q, Wang X, Zhou XM, Jiang HJ (2014) Novel electrochemical sensing platform based on magnetic field-induced self-assembly of Fe_3O_4@polyaniline nanoparticles for clinical detection of creatinine. Biosens Bioelectron 56:180–185
14. Wulff G, Sharhan A, Zabrocki K (1973) Enzyme analogue built polymers and their use for the resolution of racements. Tetrahedron Lett 14:4329–4332
15. Mosbach K (1994) Molecular imprinting. Trends Biochem Sci 19:9–14
16. Mosbach K, Haupt K (2000) Molecularly imprinted polymers and their use in biomimetic sensors. Chem Rev 100(7):2495–2504
17. Masque N, Marce RM, Borrull F, Cormack PAG, Sherrington DC (2000) Synthesis and evaluation of a molecularly imprinted polymer for selective on-line solid-phase extraction of 4-nitrophenol from environmental water. Anal Chem 72:4122–4126

18. Sellergren B (2001) Molecularly imprinted polymers: man-made mimics of antibodies and their applications in analytical chemistry. Elsevier, Amsterdam
19. Delaney TP, Mirsky VM, Wolfbeis OS (2002) Capacitive creatinine sensor based on a photografted molecularly imprinted polymer. Electroanalysis 14:221–224
20. Komiyama M, Takeuchi T, Mukawa T, Asanuma H (2003) Molecular imprinting: from fundamentals to applications. Weinheim, Wiley-VCH
21. Sergeyeva TA, Piletsky SA, Piletska EV, Brovko OO, Karabanova LV, Sergeeva LM, El'skaya AV, Turner APF (2003) In-situ formation of porous molecularly imprinted polymer membranes. Macromolecules 36:7352–7357
22. Yan M, Ramström O (eds) (2005) Molecularly imprinted materials: science and technology. Marcel Dekker, New York
23. Turner NW, Jeans CW, Brain KR, Allender CJ, Hlady V, Britt DW (2006) From 3D to 2D: a review of the molecular imprinting of proteins. Biotechnol Prog 22(6):1474–1489
24. Li S, Ge Y, Piletsky SA, Lunec J (eds) (2012) Molecularly imprinted sensors: overview and applications. Elsevier, Oxford
25. Uchida A, Kitayama Y, Takano E, Ooya T, Takeuchi T (2013) Supraparticles comprised of molecularly imprinted nanoparticles and modified gold nanoparticles as a nanosensor platform. RSC Adv 3:25306–25311
26. Takeuchi T, Sunayama H (2014) Molecularly imprinted polymers. Encyclopedia of polymeric nanomaterials. Springer, Heidelberg, pp 1–5
27. Whitcombe MJ, Kirsch N, Nicholls IA (2014) Molecular imprinting science and technology: a survey of the literature for the years 2004–2011. J Mol Recognit 27(6):297–401
28. Lakshmi D, Prasad BB, Sharma PS (2006) Creatinine sensor based on a molecularly imprinted polymer-modified hanging mercury drop electrode. Talanta 70(2):272–280
29. Sharma PS, Lakshmi D, Prasad BB (2007) Highly sensitive and selective detection of creatinine by combined use of MISPE and a complementary MIP-sensor. Chromatographia 65(7–8):419–427
30. Lakshmi D, Sharma PS, Prasad BB (2007) Imprinted polymer-modified hanging mercury drop electrode for differential pulse cathodic stripping voltammetric analysis of creatine. Biosens Bioelectron 22(12):3302–3308
31. Sergeyeva TA, Gorbach LA, Piletska EV, Piletsky SA, Brovko OO, Honcharova LA, Lutsyk OD, Sergeeva LM, Zinchenko OA, El'skaya AV (2013) Colorimetric test-systems for creatinine detection based on composite molecularly imprinted polymer membranes. Anal Chim Acta 770:161–168
32. Subrahmanyam S, Piletsky SA, Piletska EV, Chen BN, Karim K, Turner APF (2001) 'Bite-and-Switch' approach using computationally designed molecularly imprinted polymers for sensing of creatinine. Biosens Bioelectron 16(9–12):631–637
33. Tsai HA, Syu MJ (2005) Synthesis and characterization of creatinine imprinted poly (4-vinylpyridine-*co*-divinylbenzene) as a specific recognition receptor. Anal Chim Acta 539:107–116
34. Tsai HA, Syu MJ (2005) Synthesis of creatinine imprinted poly(β-cyclodextrin) for the specific binding of creatinine. Biomaterials 26:2759–2766
35. Hsieh RY, Tsai HA, Syu MJ (2006) Designing a molecularly imprinted polymer as an artificial receptor for the specific recognition of creatinine in serums. Biomaterials 27(9):2083–2089
36. Chang YS, Ko TH, Hsu TR, Syu MJ (2009) Synthesis of an imprinted hybrid organic-inorganic polymeric sol-gel matrix toward the specific binding and isotherm kinetics investigation of creatinine. Anal Chem 81(6):2098–2105
37. Syu MJ, Hsu TR, Lin ZK (2010) Synthesis of recognition matrix from 4-methylamino-*N*-allylnaphthalimide with fluorescent effect for the imprinting of creatinine. Anal Chem 82 (21):8821–8829
38. Tsai HA, Syu MJ (2011) Preparation of imprinted poly(tetraethoxysilanol) sol-gel for the specific uptake of creatinine. Chem Eng J 168:1369–1376

39. Canoa P, Simón-Vázquez R, Popplewell J, González-Fernández Á (2015) A quantitative binding study of fibrinogen and human serum albumin to metal oxide nanoparticles by surface plasmon resonance. Biosens Bioelectron 74:376–383
40. Wang YY, Han MA, Liu GS, Hou XD, Huang YN, Wu KB, Li CY (2015) Molecularly imprinted electrochemical sensing interface based on in-situ-polymerization of amino-functionalized ionic liquid for specific recognition of bovine serum albumin. Biosens Bioelectron 74:792–798
41. Zhu WH, Xuan CL, Liu GL, Chen Z, Wang W (2015) A label-free fluorescent biosensor for determination of bovine serum albumin and calf thymus DNA based on gold nanorods coated with acridine orange-loaded mesoporous silica. Sens Actuators B Chem 220:302–308
42. Park KM, Lee SK, Sohn YS, Choi SY (2008) BioFET sensor for detection of albumin in urine. Electron Lett 44(3):185–186
43. Park KY, Sohn YS, Kim CK, Kim HS, Bae YS, Choi SY (2008) Development of FET-type albumin sensor for diagnosing nephritis. Biosens Bioelectron 23(12):1904–1907
44. Fatoni A, Numnuam A, Kanatharana P, Limbut W, Thavarungkul P (2014) A novel molecularly imprinted chitosan-acrylamide, graphene, ferrocene composite cryogel biosensor used to detect microalbumin. Analyst 139(23):6160–6167
45. Das A, Bhadri P, Beyette FR, Am J, Bishop P, Timmons W (2006) A potentiometric sensor system with integrated circuitry for in situ environmental monitoring. Sixth IEEE conference on nanotechnology, Cincinnati, Ohio, USA, 17–20 June 2006, pp 917–920
46. Nien-Hsuan C, Jung-Chuan C, Tai-Ping S, Shen-Kan H (2006) Study on the disposable urea biosensors based on PVC-COOH membrane ammonium ion-selective electrodes. IEEE Sensors J 6:262–268
47. Ahmadi MM, Jullien GA (2005) A very low power CMOS potentiostat for bioimplantable applications. Proceedings of the fifth international workshop on system-on-chip for real-time applications, Banff, Alberta, Canada, 20–24 July 2005, pp 184–189
48. Wen-Yaw C, Paglinawan AC, Ying-Hsiang W, Tsai-Tseng K (2007) A 600 μW readout circuit with potentiostat for amperometric chemical sensors and glucose meter applications. IEEE conference on electron devices and solid-state circuits, Tainan, 20–22 Dec 2007, pp 1087 – 1090
49. Ahmadi MM, Jullien GA (2009) Current-mirror-based potentiostats for three-electrode amperometric electrochemical sensors. IEEE Trans Circuits Syst I Regul Pap 56:1339–1348
50. Martin SM, Gebara FH, Strong TD, Brown RB (2009) A fully differential potentiostat. IEEE Sensors J 9:135–142
51. Schienle M, Paulus C, Frey A et al (2004) A fully electronic DNA sensor with 128 positions in-pixel ADC. IEEE J Solid-State Circuits 39(12):2438–2445
52. Hassibi A, Vikalo H, Riechmann JL et al (2009) Real-time DNA microarray analysis. Nucleic Acids Res 37(20), e132
53. Daniels JS, Pourmand N (2007) Label-free impedance biosensors: opportunities and challenges. Electroanalysis 19(12):1239–1257
54. Yang C, Rairigh D, Mason A (2007) Fully integrated impedance spectroscopy systems for biochemical sensor array. Biomedical circuits and systems conference, Montreal, QC, 27–30 Nov 2007, pp 21–24
55. Jafari HM, Genov R (2011) CMOS impedance spectrum analyzer with dual-slope multiplying ADC. Biomedical circuits and systems conference, San Diego, CA, 10–12 Nov 2011, pp 361–364
56. Yang C, Jadhav SR, Worden RM et al (2009) Compact low-power impedance-to-digital converter for sensor array microsystems. IEEE J Solid-State Circuits 44(10):2844–2855
57. Liu X, Rairigh D, Mason A (2010) A fully integrated multi-channel impedance extraction circuit for biosensor arrays. Proceedings of 2010 I.E. international symposium on circuits and systems, Paris, 30 May 2010–2 June 2010, pp 3140–3143

58. Rahal M, Demosthenous A, Bayford R (2009) An integrated common-mode feedback topology for multi-frequency bioimpedance imaging. European solid-state circuits conference, Athens, 14–18 Sept 2009, pp 416–419
59. Aguirre J, Medrano N, Calvo B, Celma S, Azcona C (2011) An analog lock-in amplifier for embedded sensor electronic interfaces. European conference on circuit theory and design (ECCTD), Linkoping, 29–31 Aug 2011, pp 425–428
60. Zou X, Xu X, Yao L, Lian Y (2009) A 1-V 450-nW fully integrated programmable biomedical sensor interface chip. IEEE J Solid-State Circuits 44(4):1067–1077
61. Yazicioglu RF, Merken P, Puers R, Hoof CV (2007) A 60 μW 60 nV/$\sqrt{}$Hz readout frond-end for portable biopotential acquisition system. IEEE J Solid-State Circuits 42(5):1100–1110
62. Liu CC, Chang SJ, Huang GY, Lin YZ (2010) A 10-bit 50-MS/s SAR ADC with a monotonic capacitor switching procedure. IEEE J Solid-State Circuits 45(4):731–740
63. Johns D, Martin K (1997) Analog integrated circuit design. John Wiley & Sons, New York
64. Chang HW, Huang HY, Juan YH, Wang WS, Luo CH (2013) Adaptive successive approximation ADC for biomedical acquisition system. Microelectron J 44(9):729–736
65. Chip Implementation Center (2010) CIC referenced flow for mixed-signal IC design (Versio1.0). Taiwan
66. Lin C-P (2014) Post-layout simulation verification with nanosim. CIC Training Courses, http://www.synopsys.com/support/li/installation/documents/archive/iuguxu2003-09.pdf
67. Synopsys (2003) Nanosim integration with VCS manual. Version U-2003.03, http://www.siue.edu/~gengel/ece585WebStuff/OVI_VerilogA.pdf
68. Open Verilog International (1996) Verilog-A language reference manual. Version 1.0, http://www.siue.edu/~gengel/ece585WebStuff/OVI_VerilogA.pdf
69. Eckardt KU, Coresh J, Devuyst O et al (2013) Evolving importance of kidney disease: from subspecialty to global health burden. Lancet 382:158–169
70. Coresh J, Selvin E, Stevens LA et al (2007) Prevalence of chronic kidney disease in the United States. JAMA 298:2038–2047
71. Radhakrishnan J, Remuzzi G, Saran R et al (2014) Taming the chronic kidney disease epidemic: a global view of surveillance efforts. Kidney Int 86:246–250
72. KDIGO (2012) Clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Suppl 3(2013):1–150
73. Levey AS, de Jong PE, Coresh J et al (2011) The definition, classification and prognosis of chronic kidney disease: a KDIGO controversies conference report. Kidney Int 80:17–28
74. National Institute for Health and Clinical Excellence (NICE) (2014) Chronic kidney disease: early identification and management of chronic kidney disease in adults in primary and secondary care. Clin Guide 182
75. National Kidney Foundation (2015) KDOQI clinical practice guideline for hemodialysis adequacy: 2015 update. Am J Kidney Dis 66:884–930
76. Hillege HL, Fidler V, Diercks GFH et al (2002) Urinary albumin excretion predicts cardiovascular and noncardiovascular mortality in general population. Circulation 106:1777–1782
77. Cote AM, Brown MA, Lam E et al (2008) Diagnostic accuarcy of urinary spot protein: creatinine ratio for proteinuria in hypertensive pregnant women: systematic review. BMJ 336:1003–1006
78. Miller WG, Bruns DE, Hortin GL et al (2009) Cuurent issues in management and reporting of urinary albumin excretion. Clin Chem 55:24–38
79. McTaggart MP, Newall RG, Hirst JA et al (2014) Diagnostic accuracy of point-of-care tests for detecting albuminuria. Ann Intern Med 160:550–557

Part V
Big Data as Sensor Applications

Chapter 14
Building a Practical Global Indoor Positioning System

Dongsoo Han and Sukhoon Jung

Abstract A global indoor positioning system (GIPS) is a positioning system that provides positioning services in most buildings in villages and cities globally. Among the various wireless signals, the Wi-Fi signal has become one of the most feasible signals to realize GIPS because of its proliferation. This study introduces methods and tools to construct a GIPS by using Wi-Fi fingerprinting. An unsupervised learning-based radio map construction method is adopted to label locations of crowdsourced fingerprints, and a probabilistic indoor positioning algorithm is developed for the radio maps constructed with the crowdsourced fingerprints. Along with these techniques, collecting indoor and radio maps of buildings in villages and cities is essential for a GIPS. This study aims to collect indoor and radio maps from volunteers who are interested in deploying indoor positioning systems for their buildings. The methods and tools for the volunteers are also described in the process of developing an indoor positioning system within the larger GIPS. An experimental GIPS, named KAist Indoor LOcating System (KAILOS), was developed integrating the methods and tools. Then the COEX-mall indoor navigation system and KAIST campus indoor/outdoor integrated navigation system were developed on KAILOS, revealing the effectiveness of KAILOS in developing indoor positioning systems. The more volunteers who participate in developing indoor positioning systems on KAILOS-like systems, the sooner GIPS will be realized.

Keywords Global indoor positioning system • Crowdsourcing • Wi-Fi fingerprinting • Radio map

D. Han (✉) • S. Jung
Department of Computer Science, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
e-mail: dshan@kaist.ac.kr

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_14

14.1 Introduction

A global indoor positioning system (GIPS) is a positioning system that can provide indoor positioning services in most buildings in villages and cities globally. The goals of wide availability and high resolution should be accomplished for a positioning system to be a GIPS. While the global positioning system (GPS) has been dominantly used outdoors since its completion in the early 1990s, no GIPS with such a wide availability and high resolution has yet appeared.

Various signals such as cell-tower signals, radio frequency (RF), Wi-Fi, Bluetooth, magnetic fields, and ultra-sounds can be used for indoor positioning. Among these signals, the Wi-Fi signal is one of the best candidates to construct a GIPS because of its widespread availability of Wi-Fi hot spots all over the world. However, the wide availability of Wi-Fi signals does not guarantee a high resolution of Wi-Fi-based positioning systems. A radio map, which is a collection of fingerprints along with their collected location information, should be constructed at each building to provide WLAN-based positioning service. Here, the fingerprint is the WLAN signal characteristics represented by a set of signal strength and access point ID pairs.

However, constructing precise radio maps covering most buildings in cities globally requires tremendous time and effort. Consequently, reducing calibration efforts to construct radio maps has long been a critical issue in this research area [1]. Google has been collecting indoor floor plans and radio maps by crowdsourcing since the end of 2011 [2]. Thousands of floor plans have been collected, but until now, they have been mainly from large-scale buildings such as airport terminals, shopping malls, and exhibition centers. However precise positioning services are not yet available in most buildings because of lack of radio maps which require manual calibration efforts.

In fact, constructing a GIPS that integrates indoor maps, radio maps, and positioning algorithms is a large, complex project. This study introduces methods and tools to construct a GIPS by using Wi-Fi signals. The key idea is to realize a GIPS by collecting indoor and radio maps from volunteers who are interested in developing or deploying indoor positioning systems for their buildings. A GIPS provides methods and tools to support the volunteers, and the indoor and radio maps collected from the volunteers while they are developing their indoor positioning systems are shared by the users and applications of the GIPS.

The reason we employ a crowdsourcing approach is that it is the only way to collect indoor maps and construct radio maps globally at a very low cost in a short period of time. We developed an unsupervised learning-based radio map construction method to construct radio maps with crowdsourced fingerprints. However, all the radio maps for the GIPS need not be constructed this way. We also developed tools and web interfaces to collect radio maps from volunteers on the Internet. This radio map collection strategy inevitably results in diverse types of radio maps because collected radio maps may be constructed in various ways.

The GIPS has to deal with such various types of radio maps. It is known that one positioning algorithm cannot always outperform other positioning algorithms for all types of radio maps. The GIPS should be equipped with multiple positioning algorithms to cope with the diversity of radio map types. It switches from one positioning algorithm to another depending on the types of radio maps for better positioning performance. Mapping radio maps into appropriate positioning algorithms is an essential function of the GIPS. In this study, we propose a method to map radio maps into positioning algorithms. The diversity of radio maps cannot be handled only by the mapping of radio maps and positioning algorithms, because most existing positioning algorithms are not suitable, especially for radio maps constructed with crowdsourced fingerprints. This study also proposes a new probabilistic positioning algorithm adapted for radio maps constructed with crowdsourced fingerprints.

Using the methods and tools needed for the GIPS, an experimental GIPS, KAist Indoor LOcating System (KAILOS) was developed. KAILOS allows anyone to contribute indoor and radio maps of buildings by using its methods and tools. In return, it provides indoor positioning and navigation services for the buildings. KAILOS has a long way to go to cover most of the buildings in villages and cities globally. Nevertheless, it was used very effectively for developing indoor positioning and navigation systems in some confined areas such as COEX mall, Seoul, and the KAIST Daejeon campus. In addition, it has been shown that the unsupervised leaning-based radio map construction method and the proposed probabilistic positioning algorithm can be effectively used in reducing the cost of radio map construction and improving the accuracy of positioning in crowdsourced radio maps.

This chapter is organized as follow: Sect. 14.2 introduces a positioning technology stack; Sect. 14.3 describes the methods, techniques, and tools to construct a GIPS; Sect. 14.4 describes an experimental GIPS, KAILOS; Sect. 14.5 describes the performance evaluation of radio map construction methods and the proposed probabilistic positioning algorithm and introduces examples of using KAILOS; finally, we draw our conclusion in Sect. 14.6.

14.2 Location Technology Stack

A positioning technology stack is a layered structure of technical and environmental elements required to implement positioning systems and services. Signals, maps, positioning systems, and location-based applications are the major components constituting a positioning technology stack. Figure 14.1 shows the layered architecture of such technical and environmental elements in the positioning technology stack. As illustrated in the figure, there is only a slight difference between outdoor and indoor positioning environments from the technology point of view.

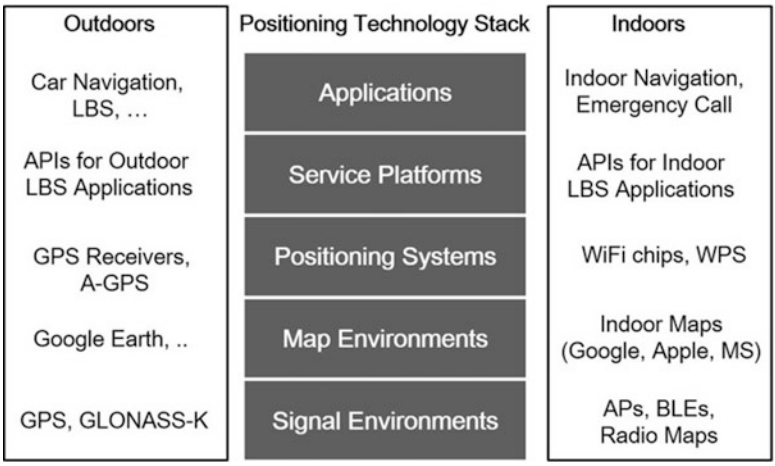


Fig. 14.1 Positioning technology stack

Signal environments are the fundamental basis of a positioning service. Without the presence of appropriate signals, positioning service is not possible in both indoor and outdoor environments. GPS signals are mainly used outdoors, whereas 3G, 4G, Wi-Fi, Bluetooth signals, and magnetic fields can be used indoors. From an availability point of view, 3G, 4G, and Wi-Fi signals can be used for the GIPS because they are available in most buildings in cities. However, from an accuracy point of view, there is no other way but to use Wi-Fi signals incorporated with fingerprint-based positioning techniques [3]. Thus, radio maps, which are collections of fingerprints with their collected locations, should be constructed for most of the buildings globally.

Map environments are another fundamental basis of positioning systems and services. Without a map, the positioning service can hardly be provided to users. In addition, many advanced positioning techniques, such as map matching and automatic radio map construction, are developed on models constructed on a map. Partitioning and drawing road networks for an area are typical examples of the modeling. A more complex modeling can be done with a state machine such as a Hidden Markov Model (HMM). We leave the details to Sect. 14.3.1.

Various location-based applications, such as navigation systems, can be developed using positioning systems. Although the positioning systems can be integrated with the applications, a positioning service platform is usually placed between the positioning systems and the applications. Hence, the positioning service platform layer is placed on top of the positioning system layer. Lastly, the location-based applications layer is placed on the top of the stack. The details of each layer of the positioning technology stacks are described in the following sections.

14.3 Methods and Tools to Construct a GIPS

14.3.1 Deployment Process of Indoor Positioning System

Prior to describing the construction of a GIPS, we describe the process of installing an indoor positioning system in a building. The techniques and tools for a GIPS will be explained along the process of installing an indoor positioning system. Various activities should be performed to deploy a fingerprint-based indoor positioning system in a building. Indoor maps should be prepared, and a model of the area is required for radio map construction and more advanced techniques. Radio maps are then constructed for the modeled indoor area by using one of the radio map construction methods. Once the construction of a radio map is completed, an indoor positioning system is installed on top of the radio maps. Testing and evaluation must be performed to ensure the quality of the deployed indoor positioning systems. In this section, we describe the detailed activities required at each step along with the methods and tools provided by the GIPS for these activities.

14.3.2 Indoor Map Registration and Modeling

14.3.2.1 Indoor Map Drawing and Registration

An indoor map is the basis for developing a fingerprint-based indoor positioning system. The collection of fingerprints, installation of positioning systems, and provision of positioning services can hardly be performed without an indoor map. However, indoor maps of a majority of the buildings in cities are not yet available. Crowdsourcing seems to be the only possible way to address this problem. To support the crowdsourcing of indoor maps, the GIPS provides tools and interfaces to collect indoor maps from volunteers all over the world. Building registration should be the first step in the collection process. This is done by drawing a polygon on a Google outdoor map. The GIPS highlights the specified building by changing the color of the polygon. Floor map registration is performed after the building registration. The GIPS also provides interfaces for volunteers to register points of interests (POIs), such as room numbers and store names, on the registered indoor map for users to locate their destinations.

Many positioning techniques were developed using various kinds of indoor maps. However, if an indoor map is specified in an image file format, it lacks information needed for some advanced positioning techniques. In addition, to use an advanced learning-based method to label the locations of crowdsourced unlabeled fingerprints, indoor areas need to be modeled with a state machine such as an HMM. We describe the details of modeling in the next section.

14.3.2.2 Modeling of Indoor Areas

To collect fingerprints, simplify the movement of a user, or represent the characteristics of signals in relation to locations, modeling of the indoor area is required. Partitioning is a basic modeling technique to collect fingerprints in an area. Fingerprints are collected at each partitioned area. Road networks are usually used to support the optimization of map-matching filters. A state machine is often used to represent the characteristics of signals observed in an indoor area [4]. An HMM, which is a variation of the state machine, is used to model the movement of users and signal characteristics in an indoor area. The transition and observation probabilities of the HMM match well to the movement of users and signal characteristics observed at each location, respectively [5].

Supporting partitioning and road network modeling in the GIPS is rather simple and straightforward. It is only necessary to provide tools to draw road networks on indoor maps. Unlike the modeling of road networks, HMM modeling involves a rather complex process. The first requirement for HMM modeling is to partition each floor into parts. The GIPS automatically divides an indoor map into parts after the specification of walls and doors. Because each partition of an indoor area is mapped onto a state of the HMM, the partitions may be considered as the states of the HMM.

14.3.3 Radio Map Construction

Radio map construction is usually performed on the models just described. In fact, radio map construction is one of the key features distinguishing the GIPS from other ordinary indoor positioning systems. It aims to support all kinds of radio map construction methods, including a novel unsupervised learning-based location labeling method (ULM) for crowdsourced fingerprints collected without location information. The ULM is expected to greatly reduce the time and effort needed to construct radio maps [6]. In this section, we present the existing radio map construction methods and ULM for the GIPS.

The first proposed radio map construction method was point-by-point manual calibration (PMC). The primary goal of this method was to achieve a high accuracy without much regard for the cost. In this method, an indoor area is partitioned into numerous parts, and then dedicated collectors collect Wi-Fi fingerprints point-by-point (see Fig. 14.2a). Because PMC requires considerable time and effort, a walking survey was developed to reduce the cost [4]. In the walking survey, only some points, such as corners and the start and end points of the survey paths, are specified to guide the collectors (see Fig. 14.2b). The fingerprints are collected while the collectors walk along the path carrying collection devices.

The semi-supervised learning method utilizes references to label the locations of fingerprints (see Fig. 14.2c). The location of access points (APs), Bluetooth signals,

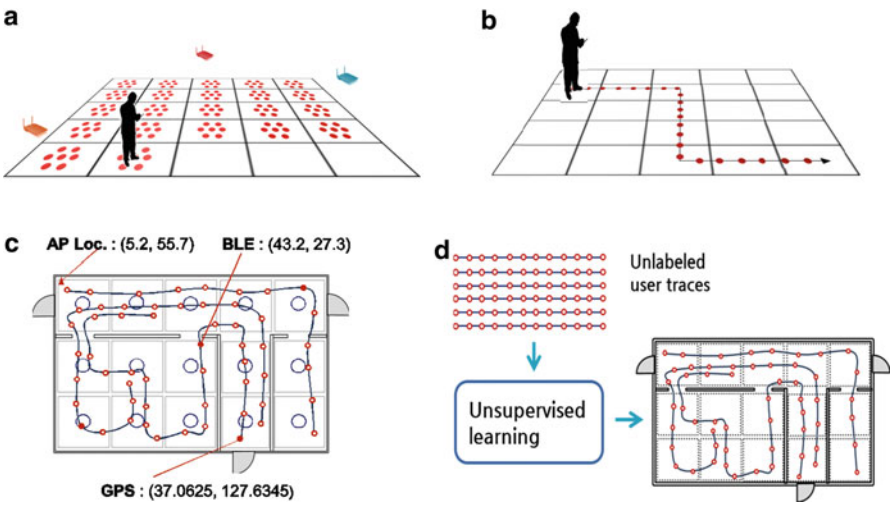


Fig. 14.2 Radio map construction methods: (**a**) PMC, (**b**) walking survey, (**c**) semi-supervised learning, and (**d**) unsupervised learning

or GPS signals are often used for the references [7]. Manifold learning [8–10] and expectation maximization [11] techniques have been developed to construct radio maps by using unlabeled fingerprints with only a few labeled fingerprints. The goal of semi-supervised learning is to further reduce the manual calibration cost, although it still requires some effort to acquire references.

A recent research approach uses inertial sensors such as a three-axis accelerometer, gyroscope, and compass. Microsoft Zee [12], UnLoc [13], WILL [14], LiFS [15], and other methods have tried to incorporate dead-reckoning techniques for labeling the locations of crowdsourced fingerprints by using inertial sensors in smartphones along with the location of stairs, elevators, and other features. These so-called sensor-based methods can further reduce the collection effort. However, the involvement of additional sensors impedes the contribution of fingerprints from numerous smartphones because sensor operation consumes additional power.

The ULM labels the location of crowdsourced fingerprints from numerous smartphones (see Fig. 14.2d). This method is distinguished from the sensor-based methods because it does not require any explicit labeling effort or sensing data for reference. The ULM, which has been implemented in the GIPS for the first time, allows almost all crowdsourced fingerprints to be used for the construction of radio maps.

Figure 14.3 shows an overview of the method. Here, the target area is assumed to have already been modeled with an HMM. It integrates location optimization and global search in a hybrid learning framework to determine an optimized placement of collected fingerprint sequences in an area. When a set of unlabeled fingerprint sequences has been collected in a building, the ULM repeats the local optimization and global search in turn until it finds an optimized placement of the unlabeled fingerprint sequences in the hybrid learning framework.

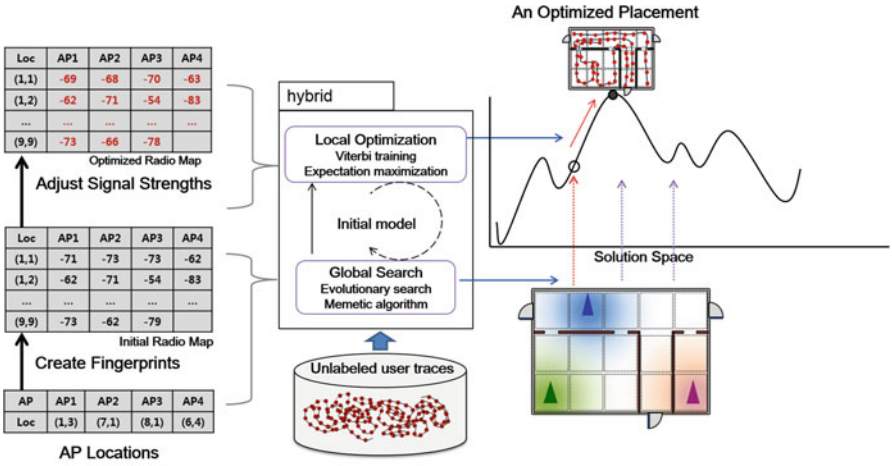


Fig. 14.3 ULM to label the location of crowdsourced fingerprint sequences

Because the local optimization often becomes stuck in local optima, the global search provides new inputs to prevent this. When a genetic algorithm is used for learning, the genotype of the leaning should be created by combining access point (AP) locations and path loss exponents, and the phenotype of the learning is represented with a set of fingerprints. The genotype, which is an abstraction of phenotypes, is used for the global search, and the phenotype is used for the local optimization. The search space of the global search can be drastically reduced using the genotype, and the local optimization can be effectively performed using the phenotype. Although the ULM requires some computational time for labeling, it can drastically reduce the manual effort needed to construct radio maps.

14.3.4 Mapping of Radio Maps and Positioning Algorithms

Once a radio map is constructed, a positioning system equipped with various positioning algorithms can be installed on the radio map to provide a positioning service. Installing a positioning system on top of the constructed radio maps is one of critical steps in deploying an indoor positioning system in a building. If the radio maps are assumed to be collected by crowdsourcing, the types of radio maps for the GIPS would be diverse in terms of collection methods and density of collected fingerprints. A positioning algorithm appropriate to each radio map type should be used to achieve high accuracy.

In general, it is known that probabilistic positioning algorithms such as the Gaussian and histogram methods can achieve a relatively good performance on radio maps with a high fingerprint density, whereas discrete positioning algorithms

do better on radio maps with a low fingerprint density. A single positioning algorithm can hardly be expected to achieve a good performance on both low- and high-density radio maps. Hence, the GIPS must be equipped with several positioning algorithms and select the most appropriate one depending on the characteristics of the respective crowdsourced radio maps. In fact, the only major difference between the positioning system for the GIPS and ordinary positioning systems is the diversity of the radio maps to be handled.

In this section, we describe the mapping between radio maps and positioning algorithms for the GIPS. We start this with a brief introduction to existing positioning algorithms.

14.3.4.1 Positioning Algorithms

Positioning algorithms can be classified into two groups: probabilistic and discrete. Positioning algorithms that estimate locations by searching the nearest fingerprints for an online measurement in discrete radio maps, such as kNN and weighted kNN [16], belong to the discrete positioning algorithm category. On the other hand, positioning algorithms that estimate locations based on the distributions of signal strengths, such as Gaussian, histogram, kernel methods, and Viterbi tracking algorithm, belong to the probabilistic positioning algorithm category.

Figure 14.4 is a rough sketch of the expected accuracy of positioning algorithms with respect to the precision of radio maps and radio map construction methods.

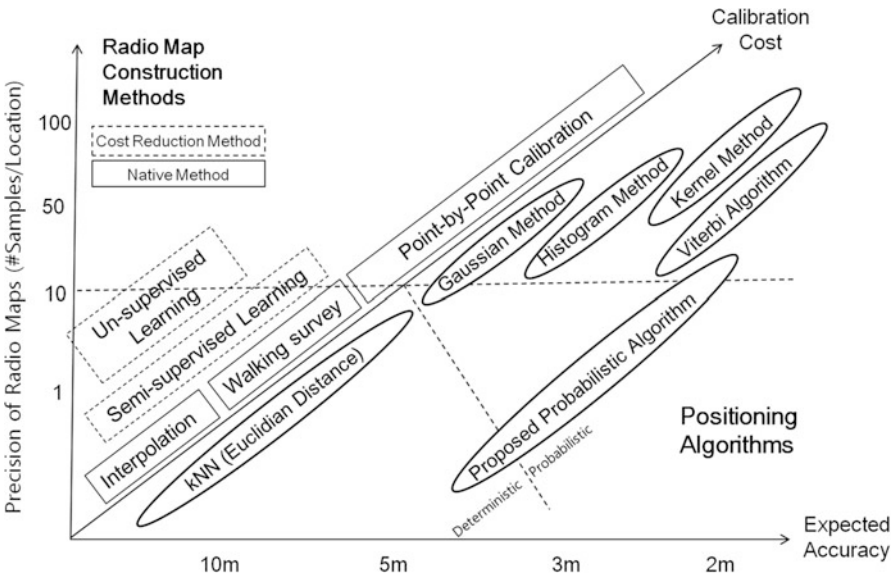


Fig. 14.4 Radio map construction methods and positioning algorithms with respect to cost and accuracy

As shown in the figure, the accuracy of the positioning algorithms is usually expected to improve with the increment of the precision of radio maps, i.e., the number of fingerprints collected at each measurement point or partitioned area. The radio map construction methods and appropriate positioning algorithms are also illustrated with the cost expectation. The information in the graph should not be taken as a definitive mapping between the radio map models and the positioning algorithms, however, because it is only a rough sketch of their matches.

14.3.4.2 Mapping Radio Maps into Positioning Algorithms

The GIPS classifies the crowdsourced radio maps into four radio map types depending on the fingerprint collection method and number of fingerprints collected at each measurement point. The types of radio maps given their collection method and number n of collected fingerprints at each measurement point or partitioned area are as follows:

- *Regular-highly-dense radio map*: PMC and $n > 20$
- *Regular-dense radio map*: PMC and $20 \geq n > 10$
- *Regular-sparse radio map*: walking survey and $n = 1$ or 2
- *Irregular radio map*: crowdsourcing (*Irregular-dense* if $n \geq 15$ and *Irregular-sparse* if $n < 15$).

Figure 14.5 shows the mapping of radio map types into radio map models and then positioning algorithms. As shown in the mapping, if a radio map is constructed

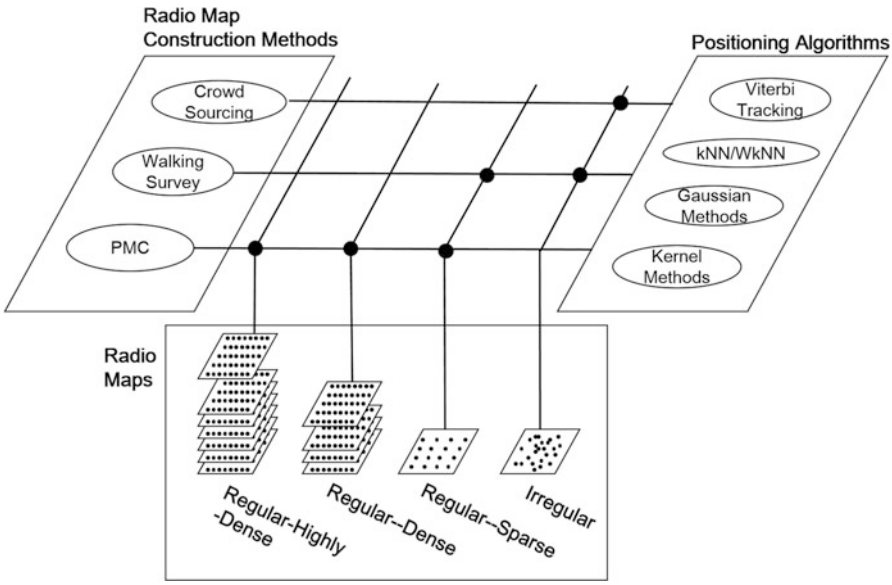


Fig. 14.5 Mapping of radio map construction methods, radio maps, and positioning algorithms

by a PMC, and more than 20 fingerprints have been collected at each measurement point, it is classified as a regular-highly-dense radio map. Any kind of radio map model can be used to represent the characteristics of a regular-highly-dense radio map, and any kind of positioning algorithms can be used for regular-highly-dense radio maps.

If a radio map is constructed by PMC with 10–20 fingerprints at each measurement point, it is classified as a regular-dense radio map. The discrete and Gaussian radio map models can be used to represent the characteristics of a regular-dense radio map. However the histogram and kernel models are not suitable because of the lack of collected fingerprints.

In contrast, if a radio map is constructed by a walking survey and only one or two fingerprints have been collected at each measurement point, it is classified as a regular-sparse radio map. Only a discrete radio map model can be used to represent the characteristics of a regular-sparse radio map. There is no other option but to use the discrete positioning algorithms, kNN or WkNN for a regular-sparse radio map.

When fingerprints have been collected by crowdsourcing, an irregular radio map is constructed. Irregular radio maps are divided into irregular-dense and irregular-sparse radio maps, depending on the number of fingerprints collected at each partitioned area. The discrete and Gaussian radio map models can be used for an irregular radio map. The positioning algorithms corresponding to the selected radio map models should be used.

14.3.5 Probabilistic Positioning Algorithm for Radio Maps with Crowdsourced Fingerprints

In the previous section, the diversity of radio maps was assumed to be handled by the GIPS with the mapping of radio maps and positioning algorithms. In reality, however, most of the existing positioning algorithms are not suitable, especially for the radio maps constructed with crowdsourced fingerprints. As a result, the GIPS can hardly be expected to achieve a good performance on radio maps constructed with crowdsourced fingerprints only by the mapping. In this section, we propose a new probabilistic positioning algorithm adapted for radio maps constructed with crowdsourced fingerprints.

The proposed positioning algorithm extends Viterbi tracking algorithm (VA) in the framework of the HMM because the VA is a probabilistic positioning algorithm utilizing historical trajectories of users [17, 18], and probabilistic tracking of dynamic movement of users can be effectively modeled in the HMM. The Extended VA (EVA) can take advantage of the probabilistic framework for tracking the dynamic movement of a user—even when only a few samples have been collected at each measurement point—without incorporation of inertial sensors. These are the common conditions of fingerprints collected from crowdsourcing.

14.3.5.1 Emission Probability

The modification of the emission probability calculation of the HMM is the first extension of the VA. The emission probability is usually calculated based on received signal strength (RSS) distributions, which require multiple fingerprint samples. An RSS distribution is known to be multimodal and often represented in the form of a histogram, log normal distribution, and Gaussian distribution, or even by a single mean RSS value [19]. The more complex distributions, such as a histogram, require more samples, but it is known that performance improvement is not significant by the choice of the distributions unless samples are enough [20]. We derive emission probabilities by using only mean RSSs that can be obtained from a few samples collected at each location, because not so many samples are assumed to be available at each location.

For the derivation of emission probabilities from a mean RSS of a location, we assume that RSS distribution follows the Gaussian for an AP at a location. Another hidden assumption of Euclidean distance (ED)-based methods is that standard deviations of RSSs are the same for all APs at any locations. If enough samples are provided, the standard deviation is worth being referred to in positioning. Otherwise, it would be better to ignore it because the standard deviation calculated from few samples is usually unreliable.

The assumptions of Gaussian distribution and uniform standard deviation may not hold for real-world RSS distributions. However, by taking these assumptions, EVA inherits the robustness of the ED-based methods despite the lack of training samples.

14.3.5.2 Dynamic Transition Probability

The second extension of the algorithm is the calculation of dynamic transition probabilities. Transition probabilities can be dynamically calculated based on the moving distance of a user at each time. In order to obtain the moving distance without the incorporation of inertial sensors, we exploit the fact that the change of user positions is reflected in Wi-Fi fingerprint. The moving distance indicated by signal changes can be estimated by the distance between the positions of two successive online fingerprints.

In the distance estimation, we consider the topology of routes induced by inner structures including doors, walls, and other barriers in an indoor area. We also consider the k most probable locations of a fingerprint for a more reliable estimation. Let kL_{t-1} be the set of the k most probable locations for an online fingerprint o_{t-1} given at time $t-1$, and kL_t be that for the next fingerprint o_t. The moving distance $d_{t-1,t}$ of the two fingerprints is then calculated as the average distance between the pairs of locations in the two sets,

$$d_{t-1,t} = \frac{1}{\left|kL_{t-1}\right| \cdot \left|kL_t\right|} \sum_{u \in kL_{t-1}} \sum_{v \in kL_t} TD(u,v), \tag{14.1}$$

where, $TD(u, v)$ returns the topological distance between two locations u and v.

Given a distance estimate $d_{t-1,t}$ at time t, a transition probability $P\left(l_j \middle| l_i\right)$ is extended to $P\left(l_j \middle| l_i, d_{t-1,t}\right)$.. This probability can be computed using PDF under Gaussian assumption if the standard deviation of estimation errors σ is given as follows:

$$P\left(l_j \middle| l_i, d_{t-1,t}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\left(TD\left(l_i,l_j\right)-d_{t-1,t}\right)^2}{2\sigma^2}}.$$

The standard deviation of estimation errors , σ is unknown. In the implementation, from an arbitrary value, σ was adjusted based on the accumulated results of the tracking; the difference between $d_{t-1,t}$ and its corresponding distance in the tracking result was considered as the error of the moving distance estimation. A maximum moving distance of a human was also set in order to remove unrealistic transitions; transitions longer than a maximum distance setting were regarded to have a zero transition probability.

14.3.5.3 Extension of Viterbi Algorithm

The VA finds a location sequence $L_t^{seq} = <l_o, \ldots, l_t>$ that maximizes $P(L_t^{seq}, O_t^{seq})$ for a given fingerprint sequence $O_t^{seq} = <o_0, \ldots, o_t>$ observed from time 0 to time t. Here, we extend the probability function to $P\left(L_t^{seq}, O_t^{seq} \middle| D_t^{seq}\right)$ in order to reflect dynamic user movements represented by a moving distance vector $D_t^{seq} = <d_{0,1}, \ldots, d_{t-1,t}>$.

With the extended probability function, the standard dynamic programming technique of VA determines the most probable trajectory of a user by simply replacing emission and transition probabilities with the extended ones. Once the most probable trajectory is found, the end point of the trajectory can be considered as the user position at current time t.

While the basic tracking considers only a single best trajectory, the best k trajectories are considered by the proposed algorithm. It estimates the position of a user by averaging the final locations of the k most probable trajectories. That strategy was known to be effective in decoding a data sequence observed through a noisy channel [21].

14.3.6 Testing and Evaluation

Testing and evaluation should be performed to measure the positioning accuracy after deploying a positioning system on the radio maps of a building. Thus the GIPS should be equipped with methods and tools for evaluation and testing. For example,

it is helpful to visualize the signal distributions of collected fingerprints before testing and evaluation. The areas in which the collection of fingerprints has been incompletely performed can be easily identified through visualization. This visualization is especially important in the GIPS because the collection activity is usually performed by volunteers who cannot easily communicate with the developers or operators of the GIPS.

The GIPS visualizes the signal characteristics of collected fingerprints at a location with a light green circle when only weak Wi-Fi signals are available, a bright green circle for a few strong signals, and a red circle when many strong Wi-Fi signals are available. When an area's set of fingerprints is incomplete by mistake, the collection of fingerprints should be performed once more at the area.

Once the test results are obtained, any faults and errors should be examined and fixed if possible. For example, floor errors are often detected during testing. The GIPS can mitigate the floor errors to some extent by using the sensing data from a pressure sensor. We do not delve into the details because of space limitations. The volunteers can improve the signal environment by installing additional APs or Bluetooth beacons.

14.4 KAIST Indoor Locating System

Most of the aforementioned methods, tools, and algorithms have been integrated into an experimental GIPS named KAILOS [22, 23]. KAILOS supports the entire deployment process of an indoor positioning system in a building. KAILOS consists of three parts: KAI-Map, KAI-Pos, and KAI-Navi. Kai-Map includes indoor and radio maps. KAI-Pos, installed on top of KAI-Map, is an indoor positioning system equipped with several positioning algorithms, such as discrete and probabilistic positioning algorithms, from which it selects an appropriate one for the underlying radio map. It provides positioning services for the buildings whose indoor and radio maps have been contributed to KAILOS. KAI-Navi is an indoor/outdoor integrated navigation system. It directs users to the destinations in both indoor and outdoor environments. The following describes the details of the KAILOS components. Figure 14.6 shows the architecture of KAILOS.

14.4.1 KAI-Map

14.4.1.1 KAI-Radio Map

KAILOS is equipped with methods and tools to support the PMC, walking survey, and unsupervised-learning-based radio map construction methods. To support the PMC, KAILOS provides tools to plan collection points on indoor maps, and collect Wi-Fi fingerprints on the planned points. The collectors are assumed to collect 10–20 fingerprints on the marked collection points. The match of the marked

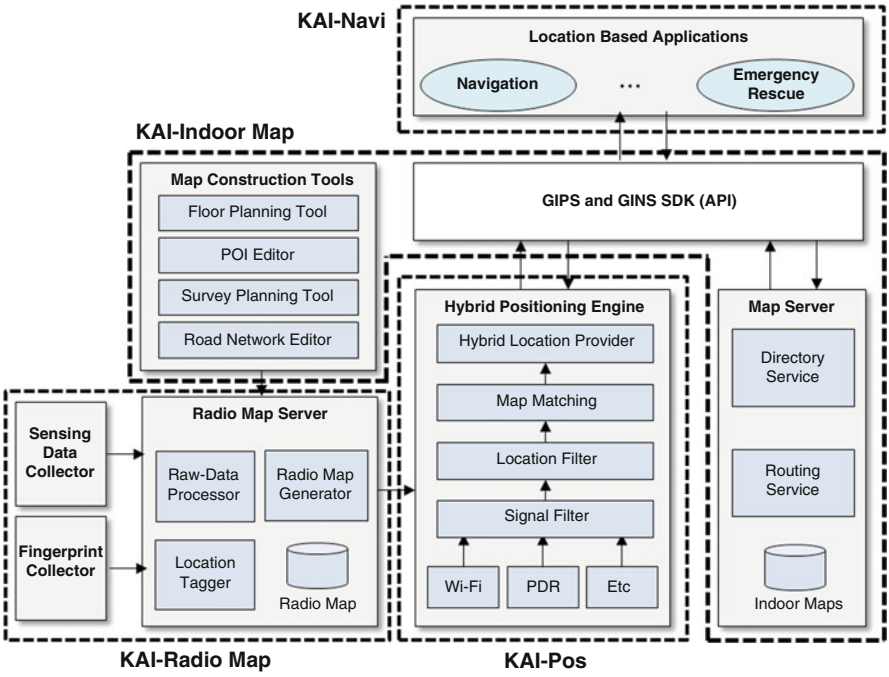


Fig. 14.6 The architecture of KAILOS

collection points and the real collection points should be confirmed by the collectors themselves during the collection activity.

KAILOS also supports the walking survey by providing tools to plan survey lines, collect fingerprints, and label the locations of collected fingerprints. In the waking survey, the survey lines should be planned in advance without the specification of collection points. Only the start and end points of survey lines are specified to guide the collectors.

KAILOS also provides tools for the proposed unsupervised learning-based method. The fingerprints are assumed to be collected by collectors who just walk around all the locations of the floor carrying the smartphones after installing the collection tool. Then KAILOS labels the locations of collected fingerprints. The time and effort of collectors can be drastically reduced by the learning-based method.

14.4.1.2 KAI-Indoor Map

In Sect. 14.3, indoor maps were assumed to be given. In reality, however, most of the indoor maps of buildings are still unavailable. To address this problem, KAILOS provides tools and interfaces to contribute indoor maps. KAILOS assumes that the contributors already have the indoor maps of their buildings in an image file format, and radio maps will be collected by walking survey or

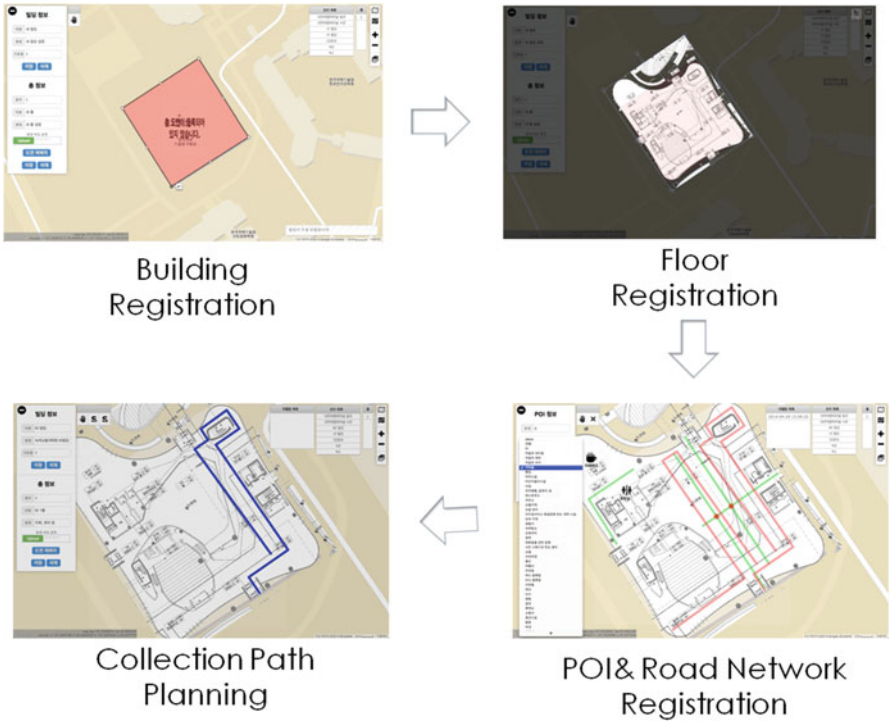


Fig. 14.7 Registration of an indoor map and designing survey lines and road networks

crowdsourcing. Figure 14.7 shows the schematic view of the contribution process. As shown in the figure, building registration is the first step of the contribution process. The floor map registration step follows after the registration of a building. POIs are registered and specified on the floor map for users to find or search their destinations. KAILOS supports volunteers by providing interfaces to register POIs for the registered indoor maps.

KAILOS also provides tools to model indoor areas. Partitioning, designing road networks, and constructing a state machine such as an HMM are typical modeling methods. Currently, KAILOS is equipped with a tool to design road networks. This is because the path finding for indoor navigation cannot be realized without the road networks.

14.4.2 KAI-Pos

The probabilistic positioning algorithm introduced in Sect. 14.3 is the basis of KAI-Pos. It also incorporates various filters for more accurate positioning. KAI-Pos adopts an adaptive hybrid filter to utilize the dynamics of user movements, and a

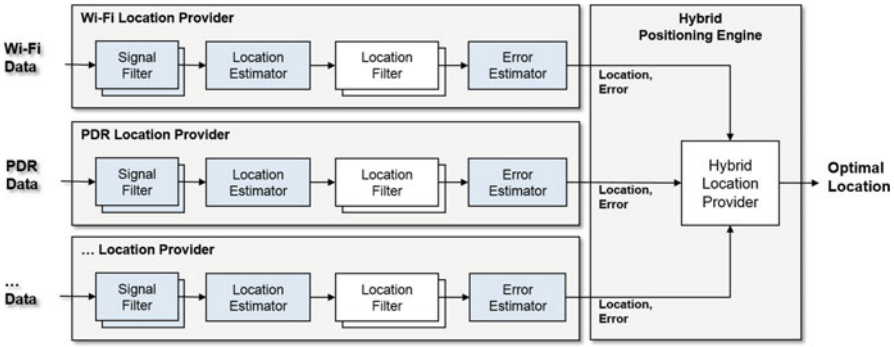


Fig. 14.8 Hybrid architecture of KAILOS positioning system

map-matching filter to utilize the road network model. The effect of the filters was apparent when they were applied for developing indoor positioning systems at the COEX mall, Seoul, Korea in 2010 and 2014.

Though Wi-Fi signals have been mainly used, KAI-Pos incorporates other wireless signals and sensing data from various sensors for more accurate positioning. Bluetooth signals are often used to cover areas in which Wi-Fi signals are not available or only weak signals are available. A pedestrian dead-reckoning PDR using smartphone sensors, such as a three-axis accelerometer and a gyroscope, has been incorporated to compensate for the gap incurred by the time interval of consecutive scans. The time interval of consecutive Wi-Fi scans is known to be approximately 3–4 s. The floor errors can also be mitigated with the incorporation of a barometer sensor. Figure 14.8 shows the architecture of KAI-Pos. In the near future, it will incorporate the additional signals and sensors in a very tightly coupled manner using the so-called sensor-fusion technique [24].

14.4.3 KAI-Navi

KAI-Navi is an indoor/outdoor integrated navigation system. It directs users in both indoor and outdoor environments. KAI-Pos operates indoors, and GPS outdoors, and the estimated current location is displayed on KAI-Indoor Map and Google Maps, respectively. Any outdoor navigation system can be integrated with KAI-Navi if it can provide open APIs to be connected to. Currently, the SK Telecom T-map outdoor navigation system, which is the most popular outdoor navigation system in Korea, is connected with KAI-Navi. The outdoor paths have been connected to indoor paths through special points designated as entrances of buildings. KAI-Navi, named *Campus Atlas* in Google Play, was deployed on KAIST campus, Daejeon, Korea, accommodating around 40 four- to five-story buildings.

14.5 Evaluation and Examples

14.5.1 Performance Evaluation of Radio Map Construction Methods

Experiments were conducted to evaluate the performance of the radio map construction methods in the N5 building, and on the seventh floor, N1 building, KAIST, Daejeon. Four kinds of radio maps were constructed at the experiment buildings by using the PMC, walking survey, semi-supervised learning, and unsupervised learning methods. A Samsung Galaxy S3 was used to collect fingerprints with a sampling rate of 1 Hz. A simple kNN method ($k = 3$) was used for the localization test in the evaluation. All programs for the evaluation were implemented in Java and run on a 3.40-GHz Intel® Core™ i7 CPU with 8GB of memory.

Approximately 2000 fingerprints were collected on the seventh floor, N1 building, and approximately 4400 fingerprints in the N5 building for each method. Figure 14.9 shows the evaluation results. When 400 fingerprints were collected

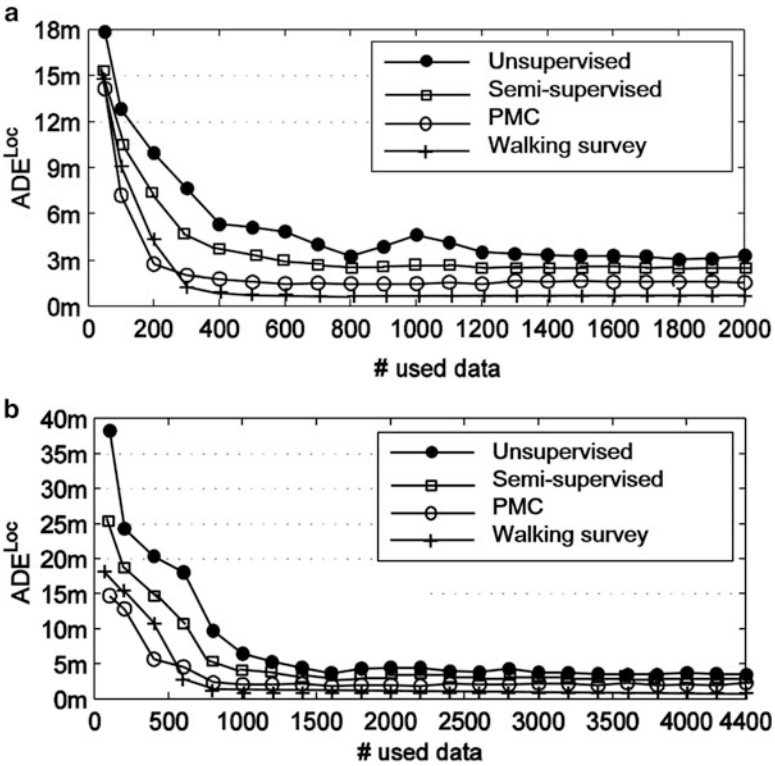


Fig. 14.9 Accuracy evaluation of radio map construction methods. (**a**) Accuracy achieved on the seventh floor, N1 building, KAIST. (**b**) Accuracy achieved in N5 building, KAIST

for the learnings, the walking survey achieved an accuracy of 1.7 m; PMC, 2.3 m; semi-supervised learning, 4.5 m; and unsupervised learning, 5.3 m. The accuracy gradually improved and then the improvement was saturated with the increase in the amount of learning data for all methods. The radio map types also converted from a sparse, to a dense, and then to a highly dense radio map with the increment in the number of fingerprints. When 2000 fingerprints were used, the walking survey achieved an accuracy of 1.2 m; PMC, 1.7 m; semi-supervised learning, 3.2 m; and unsupervised learning, 3.7 m. The walking survey outperformed the PMC in accuracy when the learning data were collected one or more times at every 1 m. This trend continues with the increment of the amount of learning data. This is because the distance between measurement points is reduced with the increase of learning data point. To collect 2000 fingerprints on the seventh floor, N1 building, the walking survey must collect fingerprints at every 4 cm, whereas the PMC was assumed to collect fingerprints at every 3 m just by varying the number of fingerprints collected at each measurement point.

Similar results were obtained in the N5 building. When 1500 fingerprints were used, the walking survey achieved an accuracy of 1.7 m; PMC, 2.5 m; semi-supervised learning, 3.3 m; and unsupervised learning, 4.7 m. When 4400 fingerprints were used, the walking survey achieved an accuracy of 1.5 m; PMC, 2.1 m; semi-supervised learning, 2.9 m; and unsupervised learning, 4.3 m.

Although the semi-supervised and unsupervised learning methods achieved accuracies slightly worse than the manual calibrations, the results were promising because they were within the accuracy range that can be used by practical positioning systems.

14.5.2 Performance Evaluation of Proposed Probabilistic Tracking Algorithm

The evaluation of EVA was conducted on the seventh floor, N1, KAIST. As three extensions were proposed, the extensions were applied to VA in stages. First, ED-based emission probabilities, dynamic transition probabilities, and their combination were applied to VA, separately; we denote them by EVA (ED), EVA (dynamic transition), and EVA (ED + dynamic transition), respectively. Then, the three best results of EVA (ED + dynamic transition) were used to evaluate the effect of considering multiple best trajectories on positioning accuracy; this configuration is denoted by 3-EVA. A simple kNN method ($k = 3$) was also compared in the evaluation. The maximum moving distance was set to 6 m for VA and EVAs in the experiments.

All of the extensions and their combinations were revealed to be effective for accurate positioning. EVA(ED) was more effective than VA, especially when using a few training samples. It provided 11 % improved positioning accuracy against VA when the number of samples was two at a location, whereas the improvement of

using 20 samples was 5 %. Using fewer samples results in less-reliable RSS distributions being constructed. Since ED is calculated based on mean RSSs irrespective of RSS distributions, EVA(ED) was less sensitive to the amount of samples.

The improvement made by dynamic transition probabilities was also significant. The strategy of considering multiple trajectories was also revealed to be effective. The improvements of 3-EVA ranged from 8 to 14 % against EVA (ED + dynamic transition), and from 20 to 28 % against VA.

The cost of training to label locations of unlabeled fingerprints is a major hurdle in building a crowdsourcing-based indoor positioning system. It is closely related to the amount of data required for the training. Because the proposed positioning algorithm can achieve high positioning accuracy even if only a few training samples are available at a location, it can be used for practical crowdsourcing-based positioning systems to reduce the cost of constructing radio maps with crowdsourced fingerprints.

14.5.3 Examples of Using KAILOS

Since the release of KAILOS in the middle of 2014, the indoor positioning systems for subway stations, indoor shopping malls, and buildings on university campuses have been developed on KAILOS. Table 14.1 is the summary of the areas, buildings, and stores in which KAILOS has been used. The indoor positioning systems of COEX and KAIST were released integrated with mycoex 3.0 and Campus Atlas apps in the Google Play store to direct pedestrians to their destinations. The indoor positioning systems for shopping malls and game rooms on KAILOS were developed to issue coupons or game items for the promotion of particular stores, events, and games. The coupons or game items are notified to users when they are detected at specific areas.

In addition, a company started registering indoor maps and constructing radio maps of subway stations on KAILOS to provide an indoor positioning service at subway stations in Seoul. The indoor and radio maps of five subway stations were collected along with their POIs. The collection activity should be conducted at as many as 600 subway stations for the complete installation of the subway indoor positioning system in Seoul. The indoor positioning systems of sea vessels were also developed on KAILOS for the safety of crew members. The working locations of crew members are displayed on a monitoring panel in an upper deck control room, and when someone is working or staying in dangerous working areas, an alarm is activated on the panel to watch the situation.

Among the examples, we describe the details of two examples: the COEX indoor positioning and navigation system deployed at the COEX mall in 2014, and an indoor/outdoor integrated campus navigation system deployed at KAIST in 2015. The KAIST campus, Daejeon, Korea, accommodates around 40 four-, five-, and ten-story buildings in an area of 1 km^2 (see Fig. 14.10a). In contrast, the COEX area

Table 14.1 Summary of indoor positioning systems developed using KAILOS

	Buildings (area size)	Main services	Radio map const. methods	Samples (train/ test)	Collection time	# of POIs	Average error distance	Collectors	Year
Subway stations	5 stations (220 × 230 m)	IPIN	Walking survey	15,000/ 2400	1 week, 3 collectors	~1250	3–15 m	Third party comp.	2015
Shopping malls	5 ten-story shopping malls (150 × 130 m)	IPIN	Walking survey	240,000/ 3500	2 weeks, 10 collectors	~10,500	2–7 m	Third party comp.	2016
Sea vessels	2 sea vessels (87 × 125 m)	Location-based safety	Walking survey	1700/210	1 day, 3 collectors	53	2–10 m	Third party comp.	2015
COEX	9-story building (304 × 652 m)	IPIN	Walking survey (2014), PMC (2010)	10,000/ 1800	1 week, 5 collectors	380	3–8 m	KAILOS team, volunteers	2010, 2014
KAIST campus	40 buildings (870 × 1250 m)	In-and-outdoor integrated navigation	Walking survey and learning	8000/ 1500	5 days, 5 collectors	4000	2–10 m	KAILOS team, volunteers	2015

Indoor positioning and indoor navigation

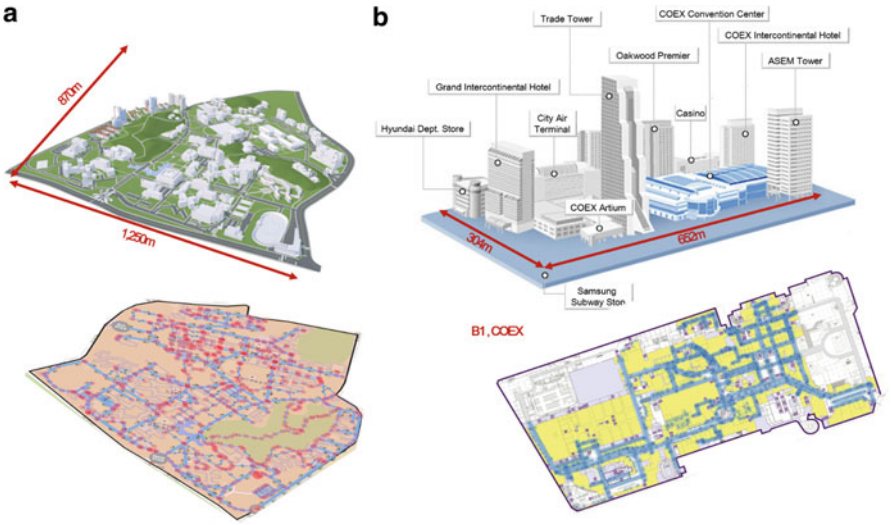


Fig. 14.10 The bird's eye view on KAIST, Daejeon, and the COEX, Seoul. (**a**) KAIST Campus, Daejeon. (**b**) COEX mall, Seoul

comprises Asia's largest underground shopping mall area, three five-star hotels, one 55-story and one 41-story premier office tower, a big department store, a subway station, a city airport terminal, and other buildings (see Fig. 14.10b).

In both cases, a walking survey was used for the collection of fingerprints. It took 5 days for five collectors to collect the fingerprints at the COEX, and 4 days for five collectors at KAIST. The accuracy differed depending on the conditions of the Wi-Fi signal environments. The areas with many strong AP signals usually showed a good accuracy. At the COEX, the accuracy on the first floor with a big open space was worse than that on the B1 floor. This result is similar to that obtained in 2010. The PMC was used to collect fingerprints at that time. Note that only one or two fingerprints were collected at each location by the walking survey, whereas approximately 20 fingerprints were collected by the PMC. On the B1 floor of the COEX (304×652 m in size), an accuracy of approximately 3–8 m was achieved by the walking survey. A similar accuracy was achieved by the PMC in 2010. A more detailed description on the PMC can be found in [23].

The man-month (MM) estimation to deploy an indoor positioning system in a building can be reduced to a tenth by using KAILOS. Approximately 20 MM were required for the development of the COEX indoor positioning and navigation system in 2010, whereas only 2 MM were required in 2014. Indoor map drawing, POI registration, modeling of road networks, survey planning, and fingerprint collection activities have been included in the MM estimation. Approximately 3 MM were required for deploying an indoor positioning system at the KAIST campus. The COEX and KAIST examples confirmed the benefits of using KAILOS for maintaining accuracy and reducing the time and effort needed to deploy indoor positioning systems.

14.6 Conclusion

Wi-Fi zones are still expanding, and the density of Wi-Fi signals is ever increasing globally. Despite the hurdles in implementing GIPS utilizing Wi-Fi signals, the Wi-Fi signals will be mainly used for the GIPS because of already-available Wi-Fi infrastructures.

This study presents the essential methods and tools to develop a GIPS by using Wi-Fi signals. An unsupervised learning-based fingerprint labeling technique was developed to construct radio maps by using crowdsourced fingerprints. It allows us to build the radio maps for most of the buildings in cities and villages at a very low cost. The practical probabilistic positioning algorithm turned out to be applicable for crowdsourced radio maps because it showed relatively good performance on the radio maps with randomly and sparsely collected fingerprints. Finally, the idea of mapping radio maps into positioning algorithms for positioning systems will allow the GIPS to accommodate new positioning algorithms.

However the techniques are applicable only for buildings whose indoor maps are available. Since the indoor maps of most of the building are currently unavailable, it will take some time until we have a complete GIPS. If the radio maps can be constructed for the buildings without indoor maps, we can shorten the time to realize the GIPS. Now, we are planning to develop the crowdsourcing radio map construction techniques for the buildings without indoor maps.

Acknowledgements This work is supported partly by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT and Future Planning as the Global Frontier Project, and partly by National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT and Future Planning (No. 2015R1A2A1A10052224).

References

1. Hossain AM, Soh W (2015) A survey of calibration-free indoor positioning systems. Comput Commun 66:1–13
2. https://www.google.com/intl/en/maps/about/partners/indoormaps/. Accessed 25 Jan 2016
3. Liu H, Darabi H, Banerjee P, Liu J (2015) Survey of wireless indoor positioning techniques and systems. IEEE Tran Syst 37(6):1067–1080
4. Kontkanen P, Myllymaki P, Roos T, Tirri H, Valtonen K, Wettig H (2004) Topics in probabilistic location estimation in wireless networks. In: Proceedings of the IEEE symposium personal, indoor and mobile radio communications (PIMRC'04), Sept 2004, Barcelona, vol 2, pp 1052–1056
5. Ladd AM, Bekris KE, Rudys AP, Wallach DS, Kavraki EE (2004) On the feasibility of using wireless ethernet for indoor localization. IEEE Trans Robot Autom 20(3):555–559
6. Jung S, Moon B, Han D (2015) Unsupervised learning for crowdsourced indoor localization in wireless networks. IEEE Trans Mob Comput PP:1–15
7. Chintalapudi K, Padmanabha Iyer A, Padmanabhan VN (2010) Indoor localization without the pain. In: Proceedings of the international conference on mobile computing and networking, ACM, New York, NY, pp 173–184

8. Ferris B, Fox D, Neil LD (2007) WiFi-SLAM using Gaussian process latent variable models. In: Proceedings of the international joint conference on artificial intelligence, vol 7. Morgan Kaufmann Publishers Inc, San Francisco, CA, pp 2480–2485
9. Pulkkinen T, Roos T, Myllymäki P (2011) Semi-supervised learning for WLAN positioning. In: Proceedings of the 21st international conference on artificial neural networks (ICANN'11), Espoo, pp 355–362
10. Pan JJ, Pan SJ, Yin J, Ni LM, Yang Q (2012) Tracking mobile users in wireless networks via semi-supervised colocalization. IEEE Trans Pattern Anal Mach Intell 34(3):587–600
11. Chai X, Yang Q (2007) Reducing the calibration effort for probabilistic indoor location estimation. IEEE Trans Mob Comput 6(6):649–662
12. Rai A, Chintalapudi KK, Padmanabhan VN, Sen R (2010) Zee: zero-effort crowdsourcing for indoor localization. In: Proceedings of the 18th annual international conference on mobile computing and networking (MobiCom'10), ACM, New York, NY, pp 293–304
13. Wang H, Sen S, Elgohary A, Farid M, Youssef M, Choudhury RR (2012) No need to war-drive: unsupervised indoor localization. In: Proceedings of the 10th international conference on mobile systems, applications, and services (MobiSys'12), ACM, New York, NY, pp 197–210
14. Wu C, Yang Z, Liu Y, Xi W (2013) WILL: wireless indoor localization without site survey. IEEE Trans Parallel Distrib Syst 24(4):839–848
15. Wu C, Yang Z, Liu Y (2015) Smartphones based crowdsourcing for indoor localization. IEEE Trans Mob Comput 14(2):444–457
16. Bahl P, Padmanabhan VN (2000) RADAR: an in-building RF-based user location and tracking system. In: Proceedings of the IEEE conference on computer communications (INFOCOM'00), vol 2. Tel Aviv, 2000, pp 775–784
17. Dugad R, Desai UB (1996) A tutorial on hidden Markov models. Technical report SPANN-96-1, Signal Processing and Artificial Neural Networks Laboratory, Department of Electrical Engineering, Indian Institute of Technology, Bombay, doi:10.1109/5.18626
18. Liu J, Chen R, Pei L, Guinness R, Kuusniemi H (2012) A hybrid smartphone indoor positioning solution for mobile lbs. Sensors 12:17208–17233
19. Kaemarungsi K (2006) Distribution of WLAN received signal strength indication for indoor location determination. In: Proceedings of the IEEE international symposium on wireless pervasive computing, 16–18 Jan 2006, pp 1–6
20. Lin T, Lin P (2005) Performance comparison of indoor positioning techniques based on location fingerprinting in wireless networks. In: Proceedings international conference wireless communications (WiCOM'05), 13–16 June 2005, vol. 2, pp 1569–1574
21. Brown DG, Golod D (2010) Decoding HMMs using the k best paths: algorithms and applications. BMC Bioinf 11(1):S28
22. Han D, Lee S, Kim S (2014) KAILOS: KAIST indoor locating system. In: Proceedings of international conference indoor positioning and indoor navigation (IPIN'14), Busan, 27–30 Oct 2014, pp 615–619
23. Han D, Jung S, Lee M, Yoon G (2014) Building a practical wi-fi-based indoor navigation system. IEEE Pervasive Comput 13(2):72–79
24. Strömback P, Rantakokko J, Wirkander SL, Alexandersson M, Fors K, Skog I, Händel P (2010) Foot-mounted inertial navigation and cooperative sensor fusion for indoor positioning. In: Proceedings of the 2010 institute of navigation-international technical meeting, San Diego, CA, 25–27 Jan 2010, pp 89–98

Chapter 15
Proximity-Based Federation of Smart Objects and Their Application Framework

Yuzuru Tanaka

Abstract This chapter focuses on the formal modeling of complex application scenarios using autonomic proximity-based federation among smart objects with wireless network connectivity, and proposes the middleware application framework to develop such complex application scenarios. Our modeling consists of three different levels. In the first-level modeling, each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object. The federation between a pair of smart objects having a pair of ports of the same type and opposite polarities is modeled as the port matching between these two ports. The second-level modeling describes the dynamic change of the federation structure among smart objects as a graph rewriting system, where each node and each link, respectively, represents a smart object and a connection between two smart objects. The third-level modeling uses a catalytic reaction network to describe each complex federation scenario in which more than one federation are involved, and an output federation of a reaction may work either as an input federation of another reaction and/or a catalyst to activate another composition or decomposition reaction.

Keywords Proximity-based federation • Smart object • Port matching • Graph rewriting • Catalytic reaction network • Autocatalytic reaction network • Pervasive computing • Ubiquitous computing • Regulation switch • Middleware framework

15.1 Introduction

During the last couple of decades, information system environments have been rapidly expanding their scope of subject resources, their geographical distribution, their reorganization, and their advanced utilization. Currently, this expansion is understood only through its several similar but different aspects, and referred to by several different stereotyped terms such as ubiquitous computing, pervasive

Y. Tanaka (✉)
Meme Media Laboratory, Hokkaido University, N13, W8, Sapporo 060-8628, Japan
e-mail: tanaka@meme.hokudai.ac.jp

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_15

computing, mobile computing, sensor networks, and IoT. No one has clearly defined this expansion as a whole.

Recently it is often pointed out that the lack of a formal computation model capable of context modeling to cover this diversity as a whole is the main reason why most applications of ubiquitous computing and/or pervasive computing are still within the scope of the two stereotyped scenarios [1, 2], i.e., the location transparent service continuation and the location- and/or situation-aware service provision. The former one focuses on the ubiquity of services, while the latter focuses on the context-dependent services. In addition to these two scenarios, IoT focuses on the dynamic federation among smart objects through the Internet, i.e., the web-based federation of smart objects.

Some researchers have been trying to extend the application target of formal computation models of process calculi, which were originally proposed to describe dynamically changing structures and behaviors of interoperating objects, from sets of software process objects to sets of mobile physical computing objects [1]. Such formal computation models of process calculi include Chemical Abstract Machine [3], Mobile Ambients [4], P-Systems [5], Bigraphical Reactive System [6], Seal Calculus [7], Kell Calculus [8], and LMNtal [9]. These trials mainly focus on mathematical description and inference of the behavior of a set of mobile objects, but not those of the dynamically changing interconnection structures among mobile physical objects based on the abstract description of their interfaces. For this reason, their formal computation models are not sufficient to develop any innovative application framework for the dynamic and complex interoperability application scenarios of smart objects.

On the other hand, as to the modeling and analysis of dynamically changing topology of ad hoc networks, there have been lots of mathematical studies on network reconfiguration and rerouting for energy saving, for improving quality of service, and/or for maintaining connectivity against mobility [10, 11]. They focus on physical connectivity among nodes, but not on their logical or functional connectivity. These models cannot describe application frameworks.

Some studies on mobile ad hoc networks are inspired by biological systems that share such similar features as complexity, heterogeneity, autonomy, self-organization, and context-awareness. These approaches are sometimes categorized as studies on bio-inspired networking [12]. The middleware framework to be proposed in this chapter is also bio-inspired, especially by DNA self-replication and RNA transcription mechanisms.

In expanding information environments of ubiquitous and/or pervasive computing, some resources are accessible through the Web, while others are accessible only through peer-to-peer ad hoc networks. IoT focuses only on the former case. Any advanced utilization of some of these resources needs a way to select them, and a way to make them interoperable with each other to perform a desired function. Here we use the term “federation” to denote the definition and execution of interoperation among resources that are accessible either through the Internet or through peer-to-peer ad hoc communication. This term was probably first introduced to IT areas by Dennis Heimbigner in the context of a federated database

architecture [13], and then secondarily in late 1990s, by Bill Joy in a different context, namely federation of services [14]. Federation is different from integration in which member resource objects involved are assumed to have previously designed standard interoperation interface. The current author has already proposed federation architectures for resources over the Web [15–18], and extended their targets to cover sensor networks using ZigBee Protocol and mobile phone applications. These architectures are, however, still within the framework of Web-based federation.

This chapter will focus on the proximity-based federation of intellectual resources on smart objects. Proximity-based federation denotes federation that is autonomously activated by the proximity among smart objects. Smart objects denote computing devices such as RFID tag chips, smart chips with sensors and/or actuators, mobile phones, mobile PDAs, intelligent electronic appliances, embedded computers, and access points with network servers.

This chapter will propose a generic middleware framework to develop complex applications of smart objects in which more than one federation are involved. Our framework is based on the three formal models we proposed in [19, 20] to describe three different levels of autonomic proximity-based federation among smart objects including both physical smart objects with wireless network connectivity and software smart objects such as services on the Web. These three formal models focus on different levels, i.e., federation and interoperation mechanisms between a pair of smart objects, dynamic change of federation structures among smart objects, and complex application scenarios with mutually related more than one federation.

Our first-level formal modeling focuses on the federation interface of smart objects, hiding any details on how functions of each smart object are implemented. Each smart object is modeled as a set of ports, each of which represents an I/O interface of a service provided by this smart object or by another smart object. We consider the matching of a service-requesting query and a service-providing capability as the matching of a service-requesting port and a service-providing port. In the preceding research studies, federation mechanisms were based on the matching of a service-requesting message with a service-providing message, and used either a centralized repository-and-lookup service as in the case of Linda [21] or multiple distributed repository-and-lookup services each of which is provided by some mobile smart object as in the case of Lime [22]. Java Space [23] and Jini [24] are Java versions of Linda and Lime middleware architectures. A survey on such middleware architectures can be found in [25]. In these architectures, messages to be matched are issued by program codes, and therefore the dynamic change of federation structures by message matching cannot be discussed independently from the codes defining the behavior of the smart objects. Our first-level formal modeling allows us to discuss applications from the view point of their federation structures. This enables us to extract a common substructure from similar applications as an application framework.

Our second-level formal modeling based on graph rewriting rules focuses on developing application frameworks each of which uses a dynamically changing single federation, while our third-level formal modeling focuses on developing

complex application scenarios in which many smart objects are involved in mutually related more than one federation, and describes them as collectively autocatalytic sets proposed by Stuart Kauffman in the context of complex systems [26].

Based on these three formal models of smart object federations, this chapter proposes a generic middleware framework for the development of applications with complex federation scenarios in which more than one federation are involved, and a result federation may work as a component of another federation and/or a promoter of another federation.

15.2 Smart Object and Its Three Levels of Formal Modeling

Here we first briefly review our three different levels of formal modeling [19, 20].

15.2.1 Smart Object and Its Port Matching (The First-Level Formal Modeling)

Each smart object communicates with another through a peer-to-peer communication facility, which is either a direct cable connection or a wireless connection. Some smart objects may have WiFi communication and/or cellular phone communication facilities for their Internet connection. These different types of wireless connections are all proximity-based connections, i.e., each of them has a distance range of wireless communication. We model this by a function *scope*(o), which denotes a set of smart objects that are currently accessible by a smart object o.

For a smart object to request a service running on another smart object, it needs to know the id and the interface of the service. We assume that each service is uniquely identified by its service type in its providing smart object. Therefore, each service can be identified by the concatenation of the object id of its providing smart object and its service type. The interface of a service can be modeled as a set of attribute-value pairs without any duplicates of the same attribute. We call each attribute and its value, respectively, a signal name and a signal value.

Pluggable smart objects cannot specify its access to another or its service request by explicitly specifying the object id or the service id. Instead, they need to specify the object by its name or the service by its type. The conversion from each of these two different types of reference to the object id or the service id is called "resolution." These are, respectively, called object-name resolution and service-type resolution.

When a smart object can access the Internet, it can ask a central repository-and-lookup service to perform each resolution. When a smart object can access others only through peer-to-peer network, it must be able to ask each of them to perform

each resolution. Here we assume that every smart object performs required resolution for its own services.

When a service-requesting smart object o requests a service, it sends an object id *oid*', an object name *oname*, or a service type *stype* to each smart object with *oid* as its object id in its proximity represented by *scope*(o). Each recipient smart object with *oid*", when receiving *oid*', *oname*, or *stype*, respectively, performs object-id resolution, object-name resolution, or service-type resolution. Object-id resolution returns the input *oid*' if it is equal to *oid*", or "nil" otherwise. Object-name resolution returns *oid*" if the recipient has *oname* as its name, or "nil" otherwise. Service-type resolution returns *oid*" if the recipient provides a service of type *stype*, or "nil" otherwise. After obtaining *oid*", the service-requesting smart object can directly request the object with *oid*" for a service of type *stype*. The object with *oid*" can acknowledge this request if it provides such a service. Otherwise it returns "nil."

Our model represents each resolution mechanism as well as the corresponding resolution request (namely, the corresponding access request) as a port. Each smart object is modeled as a set of ports. Each port consists of a port type and its polarity, i.e., either a positive polarity "+" or a negative polarity "−". Each resolution mechanism is represented by a positive port, while each resolution request is represented by a negative port. A smart object with *oid* as its identifier has a port +*oid*. If it exposes its name *oname*, namely if it allows its reference by its name, then it has a port +*oname*. A smart object has ports −*oid* and/or −*oname* if it requests another smart object identified by *oid* and/or *oname*. A smart object that provides a service of type *stype* has a port +*stype*. A smart object has a port -*stype* if it requests a service of type *stype*. A smart object with *oid* and *oname* may have +*oid* and +*oname* as its ports, but neither of them is mandatory. Some smart objects may hide one or both of them.

Federation of a smart object o with another smart object o' in its scope *scope* (o) is initiated by a program running on o or on some other activating smart object that can access both of these objects. This program detects either a specific user operation on o or the activating object, a change of *scope*(o), or some other event on o or the activating object as a trigger to initiate federation. The initiation of federation with o' by a smart object o or by some other activating object first checks if o' exists in *scope*(o), and, if yes, it performs the port matching between the ports of o and the ports of o'. As its result, every port –p in o is connected with a port +p in o' by a channel identified by their shared port type p. We assume that ports are not internally matched with each other to set any channel within a single object.

The same smart object may be involved in more than one different channel. The maximum number of channels in which the same port can be involved is called the arity of this port. Unless otherwise specified, we assume in all the examples described in this chapter that the arity of each port is one.

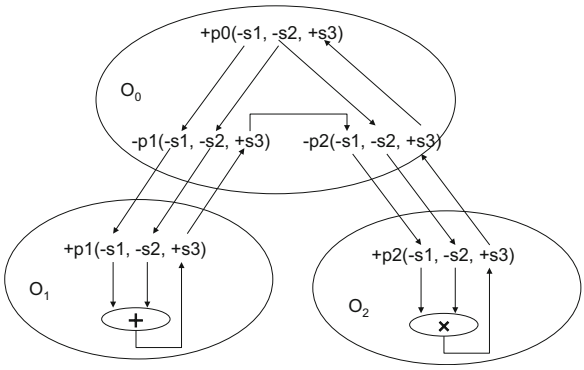


Fig. 15.1 Example federation among three smart objects

15.2.2 Semantics of Federation

This chapter does not deal with the semantics of each federation in detail. Here we briefly show how the semantics of federation can be defined. Let us consider a federation among three smart objects O_0, O_1, and O_2, as shown in Fig. 15.1, respectively, having the following port sets, {+p0, −p1, −p2}, {+p1}, and {+p2}. Each of these three ports has three IO signals represented by s1, s2, and s3. The polarity −s or +s of the signal s denotes that it works, respectively, as an input or as an output. The smart objects O_1 and O_2, respectively, perform the addition s3:=s1 + s2 and the multiplication s3:=s1 × s2, while the service p0 provided by O_0 combines these two functions provided by O_1 and O_2 to calculate (s1 + s2) × s2 as the value of s3.

The semantics of this federation can be described in a Prolog-like language as follows:

$$\begin{aligned} &O_0 : \mathrm{p0}(x, y, z) \leftarrow \mathrm{ext}(\mathrm{p1}(x, y, w)),\ \mathrm{ext}(\mathrm{p2}(w, y, z)) \\ &O_1 : \mathrm{p1}(x, y, z) \leftarrow [z := x + y] \\ &O_2 : \mathrm{p2}(x, y, z) \leftarrow [z := x \times y] \end{aligned}$$

Each literal on the left-hand side of a rule corresponds to a service-providing port, while each literal on the right-hand side corresponds to either a program code or a service-requesting port. Each literal ext(L) denotes that L is evaluated externally by some other accessible smart objects. When the service-providing port p0 of O_0 is accessed with two signal values "a" and "b," the smart object O_0 begins to evaluate the goal

$$? \leftarrow \mathrm{p0}(a, b, z).$$

The evaluation of this goal finally obtains the value of z as $z = (a + b) \times b$.

15.2.3 Software Smart Object

Our model treats each WiFi access point as a smart object. Once a smart object federates with some access point *apoint*, it can access whatever Web services this access point is given permission to access. Such an access point *apoint* provides the following service:

$$\text{ResDelegation}(\text{x}, \text{y}) \leftarrow \text{isURL}(\text{x}), \text{permitted}(\textit{apoint}, \text{x}), [\text{WebEval}(\text{x}, \text{y})].$$

The procedure WebEval(x, y) invokes the web service x with signals y. In the proximity of *apoint*, any smart object o with a port −ResDelegation can request this service with a Web service URL *url* and a list of parameters **v**. The access point *apoint* delegates this access request to the target Web service at *url* together with the parameter list **v**, which enables the smart object o to utilize this Web service through *apoint*.

A smart object may presume a standard API to request a service of another smart object that does not provide the compatible API but provides a downloadable driver to access this service through the presumed standard API. Such a mechanism is described as follows. Suppose that a smart object has a service to download a software smart object y satisfying the query x. This smart object has a port +SOdownload defined as follows:

$$\text{SOdownload}(\text{x}, \text{y}) \leftarrow [\text{find}(\text{x}, \text{y})]$$

Here the procedure find(x, y) finds the software smart object y satisfying the query x specified in the list format of attribute-value pairs $((attr_1, \text{v}_1), \ldots, (attr_k, \text{v}_k))$. If there are more than one such smart object, it returns an arbitrary one of them. A requesting smart object with a port −SOdownload can ask this smart object to download a software smart object y that satisfies the query x, and install this software smart object y on itself. The evaluation of the following by the requesting smart object performs both the downloading and the installation.

$$\text{SOdloadInstall}(\text{x}) \leftarrow \text{ext}(\text{SOdownload}(\text{x}, \text{y})), [\text{install}(\text{y})]$$

The installation of a software smart object y by a requesting smart object o also adds y in the scope of o, and initiates a federation between o and y.

If a downloaded software smart object is a proxy object to some Web service or some other smart object, the recipient smart object can access this remote service through this proxy software smart object.

One of the stereotyped application scenarios of ubiquitous computing, i.e., the location transparent service continuation, can be easily described using a software smart object. Let us consider the following example. Suppose that a personal data adapter PDA1 has federated with an access point Office with a server. Suppose that Office provides a DB service and a print service that are available at the user's

office with this access point. Let Home be another access point with a server providing a print service available at his or her home. The location transparent continuation of services means that he or she can continue the job that was started at the office using PDA1 even after he or she goes back home carrying PDA1. This is realized by the following mechanism. When he or she carries out PDA1 from Office environment, he or she just needs to make PDA1 download the proxy software smart object of Office. This operation is called federation suspension. When arriving at home, PDA1 is WiFi connected to Home, and then federates with Home. At the same time, the proxy software smart object installed in it resumes the access to Office via this proxy software smart object. Therefore, they set up two channels for print services, to the one available at home and to the one at the office, and one more channel for the DB service to access the Office DB service from home. Now PDA1 can access the database service of Office, and the two printing services of Office and Home. When PDA1 requests a print service, it asks its user or the application which of these services to choose. This enables the PDA1 user to restart his or her work with the same service accessibility after moving from the Office to Home.

This downloadable software proxy smart object can be defined as follows:

$$\begin{aligned}
&\mathrm{DB}(\mathrm{x}) \leftarrow \mathrm{remoteEval}(\mathrm{DB}(\mathrm{x}), \mathrm{Office}) \\
&\mathrm{Print}(\mathrm{x}) \leftarrow \mathrm{remoteEval}(\mathrm{Print}(\mathrm{x}), \mathrm{Office}) \\
&\mathrm{suspend}(\cdot) \leftarrow [\mathrm{disconnect}(\mathrm{Office})] \\
&\mathrm{resume}(\cdot) \leftarrow [\mathrm{connect}(\mathrm{Office})]
\end{aligned}$$

Here, remoteEval(p(x), y) denotes an evaluation of p(x) in a remote object y for which this proxy object works as the proxy. The last two rules define the disconnection and connection of this proxy object from and to the smart object Office.

15.2.4 Graph Rewriting System (The Second-Level Formal Modeling)

The second-level formal modeling focuses on the dynamic change of federation structures among smart objects. It describes a system of smart objects as a directed graph in which each node represents either a smart object or a port, and each directed edge represents either a channel or a proximity relationship. A smart object node and a port node are, respectively, represented by a bigger white or gray circle and a smaller black circle. A channel and a proximity relationship are represented, respectively, by a black arrow and a gray arrow. Each gray arrow denotes that the pointed smart object is in the scope of the pointing smart object. A port node with an outgoing (or incoming) channel edge p to (from) an object node o denotes that this object o has a service-providing (service-requesting) port +p (−p), and that it is not involved in any federation yet. Each object node has its state and its type. Smart

objects of the same type share the same port set and the same functions. The formalization with graph rewriting rules aims to describe the dynamic change of the channel connections among smart objects through the activation of federation rules, and to hide all the details about the execution of service functions.

Each rewriting rule is specified as a combination of the following four types of rules, i.e., port activation/deactivation rules, state setting rules, channeling rules, and channel dependency rules. In each of the following rules, there always exists only one gray smart object node. This gray smart object node called the rule-activation node denotes that this rule is stored in this smart object node and executed by this node. Each type of rules is carefully designed to satisfy reasonable hardware and performance constraints of the smart objects of our concern so that it can be executed locally without assuming the accessibility to any global information about the current overall federation structure. The left-hand side of each rule specifies the condition for this rule to be executed. The rule-activation node should be able to check all the specification conditions in this condition part such as those on smart object states, smart object types, port types, port availabilities, channel types, channel connections, and proximity relations of this activation node itself and of all the other nodes specified in the condition part. This implies that the rule-activation node should be able to access all these nodes through channels. The right-hand side of each rule specifies the actions to be executed by the rule-activation node, which should be able to perform these actions directly or to instruct other nodes through channels to perform these actions.

Port activation and deactivation rules have the forms in Fig. 15.2a, b. A dotted arrow σ denotes a channel path, i.e., a sequence of channels in the same direction whose length may be zero. The gray smart object node of type t at its state S can activate or deactivate a specified port of the right smart object node through the channel σ, and change the state of itself to S'. The four black smaller nodes in (a) and (b) denote port nodes. If a port is inactive, it cannot be seen from the outside of its owner smart object. If it becomes active, it can be seen from the outside.

State setting rules have the form in Fig. 15.3, where $S' = T$ if the length of σ is zero.

Figure 15.4 shows the form of channeling rules for setting channels. In each rule in Fig. 15.4a, the rule-activation smart object node (i.e., the gray node) can activate or deactivate a specified port of the left smart object node through the channel σ to establish or to break the corresponding channel with its neighboring smart object node. The length of σ may be zero. In the first rule in Fig. 15.4b, the rule-activation smart object node (i.e., the gray node) first reads the oid of the left smart object node through the channel $\sigma 1$, and ask the right smart object node to create the corresponding oid-requesting port in itself, and establishes an oid channel between these two smart objects. In the second rule in Fig. 15.4b, the rule-activation smart object uses $-p$ and $+p$ ports to establish a channel of type p between the two smart objects. The length of either $\sigma 1$ or $\sigma 2$ may be zero.

Figure 15.5 shows the form of channeling rules for breaking channels. The activation smart object node (i.e., the gray node) can break a specified channel

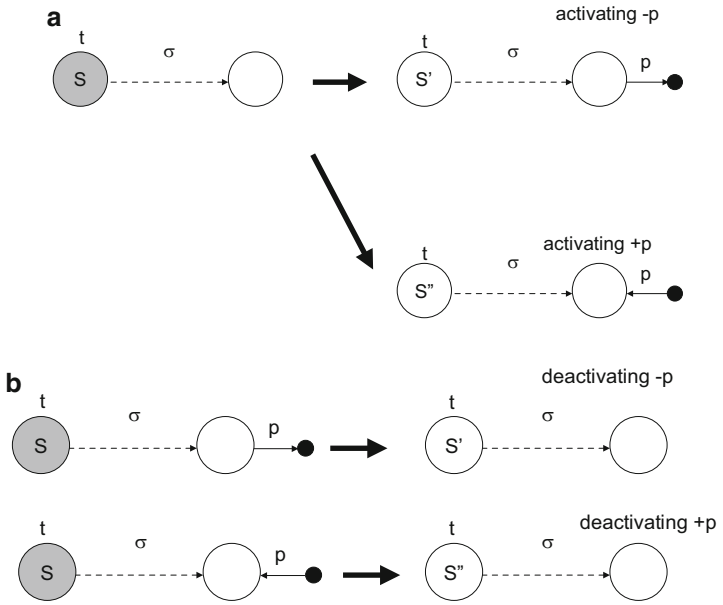


Fig. 15.2 Port activation/deactivation rules. (**a**) Port activation rules. (**b**) Port deactivation rules

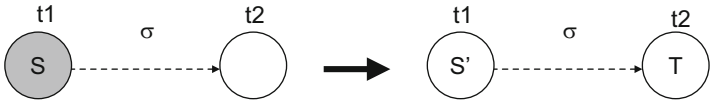


Fig. 15.3 State setting rules

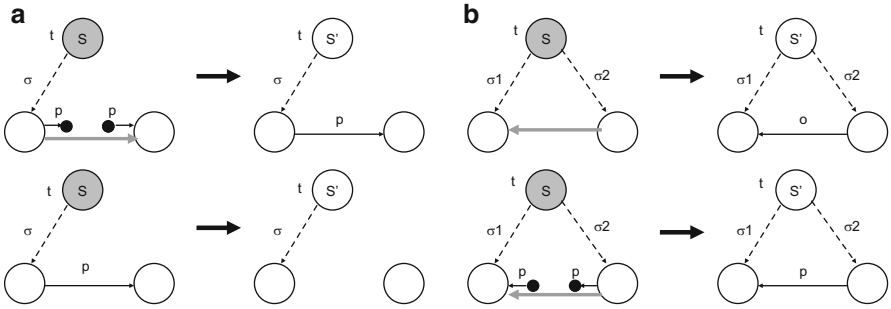


Fig. 15.4 Channeling rules for setting channels. (**a**) Rules with one path reference. (**b**) Rules with two path references

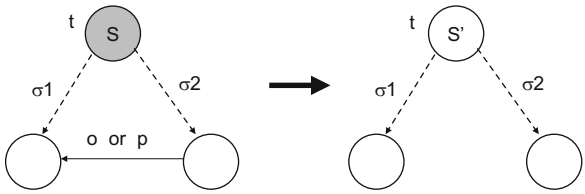


Fig. 15.5 Channeling rules for breaking channels

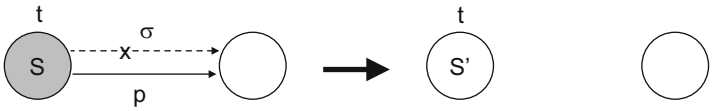


Fig. 15.6 Channel dependency rule

between the left smart object node accessible through σ1 and the right smart object node accessible through σ2. The length of either σ1 or σ2 may be zero.

Channel dependency rules have the form in Fig. 15.6, where the channel p is assumed to depend on the channel path σ. Whenever p is to be used, the existence of the channel path σ is checked. If σ is broken, the channel p is immediately broken.

15.2.5 Catalytic Reaction Network (The Third-Level Modeling)

A linear federation $o_1o_2...o_n$ of n smart objects $o_1, o_2, ..., o_n$ denotes a sequence of smart objects in which, for each $i = 1, ..., n-1$, there is an L channel from o_{i+1} to o_i. For two linear federations or smart objects X and Y, their linear federation XY denotes a liner federation in which the first smart object in Y spans an L channel to the last smart object in X. The type of a linear federation $o_1o_2...o_n$ is defined as the concatenation of the object types of $o_1, o_2, ..., o_n$. Once we have a linear federation, it is easy to span an additional channel between any pair of nodes by making the object o_n to execute such a rewriting rule in Fig. 15.4b since o_n can access any objects in this linear federation through L channels. Such additional rules can be packaged into a software smart object, and downloaded later into o_n for execution. Therefore, in the sequel, we focus our discussion only on linear federations.

For the modeling of complex application scenarios of autonomic linear federation of smart objects, we use catalytic reaction networks. A catalytic composition reaction accepts more than one input, and combines them to produce one output, while a catalytic decomposition reaction accepts one input, and decomposes it into two outputs. Each reaction may or may not require a catalyst. Some catalyst works as the context of a reaction and is immobile, while some other catalyst is mobile. Mobile catalysts are called stimuli. Each stimulus works as either a promoter or an inhibitor of the reaction, while each context always works as a promoter of the reaction. Different from inputs, catalysts are not consumed nor modified by their reactions.

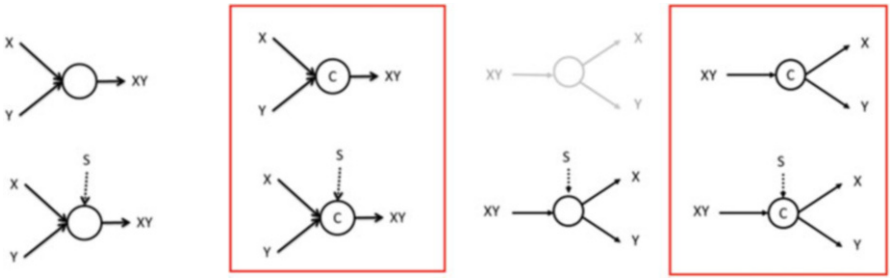


Fig. 15.7 Composition and decomposition reactions with and without contexts and/or stimuli

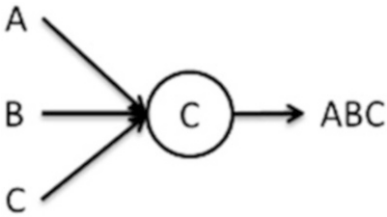


Fig. 15.8 A confederation catalytic reaction with three inputs and their linear federation as its output

Our third-level modeling uses a catalytic composition reaction and a catalytic decomposition reaction to, respectively, represent federation and defederation of linear federations of smart objects. Figure 15.7 lists up all kinds of composition reactions and decomposition reactions, where X, Y, C, and S denote smart object types or linear federation types, and XY denotes the type of a linear federation of two linear federations of types X and Y, i.e., the type of their concatenation. Each stimulus may take either a positive or a negative sign depending on whether it works as a promoter or an inhibitor. Here in this chapter we only focus on promoter stimuli. The catalysts C and S are, respectively, called the context and the stimulus of the corresponding reaction. Any decomposition reaction without any context or stimulus means autonomous decomposition, which indicates that its input linear federation is instable. Therefore, we do not consider such decomposition reactions in our catalytic reaction network modeling. Furthermore, this chapter mainly focuses on those reactions with contexts, which allows us to design the graph rewriting rules of each context to execute the corresponding reaction.

Figure 15.8 shows a catalytic reaction with three inputs of type A, B, and C and one output of type ABC, and a context of type C. This reaction represents the federation of three different types of objects under the support of a context smart object. This context object federates an input triple consisting of three smart objects a, b, c of type A, B, and C to compose a linear federation abc of type ABC. Whenever a new triple comes within the scope of the context object, a new linear federation of type ABC is immediately composed.

Let us consider the following example. When a rescue center receives an emergency call, it mobilizes a rescue team. Each rescue worker of the team needs to pick up some rescue equipment modules necessary for the mission. In a near

future, those equipment modules may include a wearable computer, a GPS module, a hands-free mobile phone, a head-mount display, a small reconnaissance camera robot, and its remote controller. The rescue center has a sufficient number of equipment modules of each type. Depending on each mission, each worker picks up one from the stock of each different type of equipment. The set of picked-up equipment modules may differ for different missions. These equipment modules are advanced IT devices, and can interoperate with each other. It is necessary to set up instantaneously all the necessary federation channels among those equipment modules picked-up by each worker. These federations should be able to avoid cross-talks between equipment modules picked up by different workers. Suppose each equipment A of a worker P needs to interoperate with his another equipment B, A and B should not interoperate with B and A of another worker P′ even if P and P′ work within their proximities. We need a new mechanism for the instantaneous setting-up of such a federation among a set of equipment modules for each of more than one worker. We call such a mechanism a "confederation" mechanism. Such a confederation mechanism can be embedded in a gate. It works as the context object of this confederation reaction. This context object is called the confederator. Whenever each worker passes through this gate, carrying the set of these required modules with him or her, this confederator in the gate immediately establishes the federation among these modules.

Here we give a general definition of confederation as follows. Consider an n-tuple of smart objects o_1, o_2, …, o_n of n different types T_1, T_2, …, and T_n. Initially, they are independent from each other. The type of a tuple of these smart objects (o_1, o_2, …, o_n) is defined as (T_1, T_2, …, T_n). A confederator of type (T_1, T_2, …, T_n) is a smart object that sets up a federation to each n-tuple of smart objects having the tuple type (T_1, T_2, …, T_n). Such a federation can be set up by the graph rewriting rules in Fig. 15.9 that are executed by this confederator. Here, for simplicity, we show the case for $n = 3$.

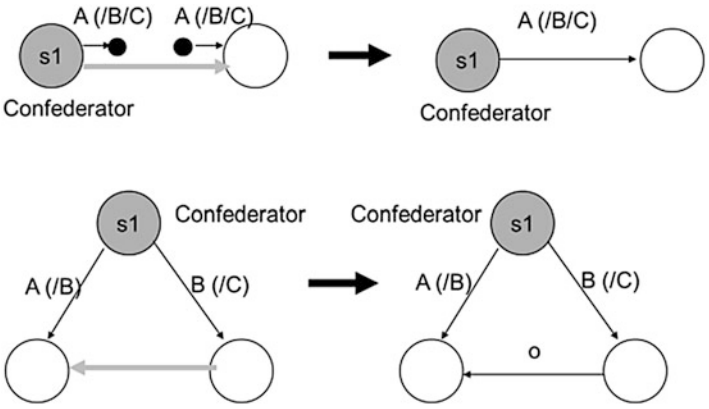


Fig. 15.9 The rewriting rules of the confederator for the linear connection

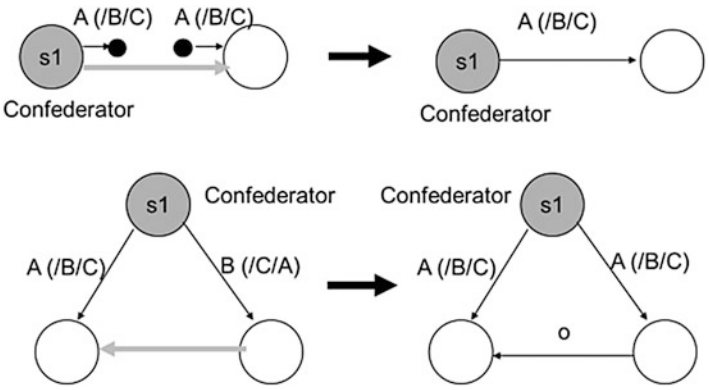


Fig. 15.10 The rewriting rules of the confederator for the cyclic connection

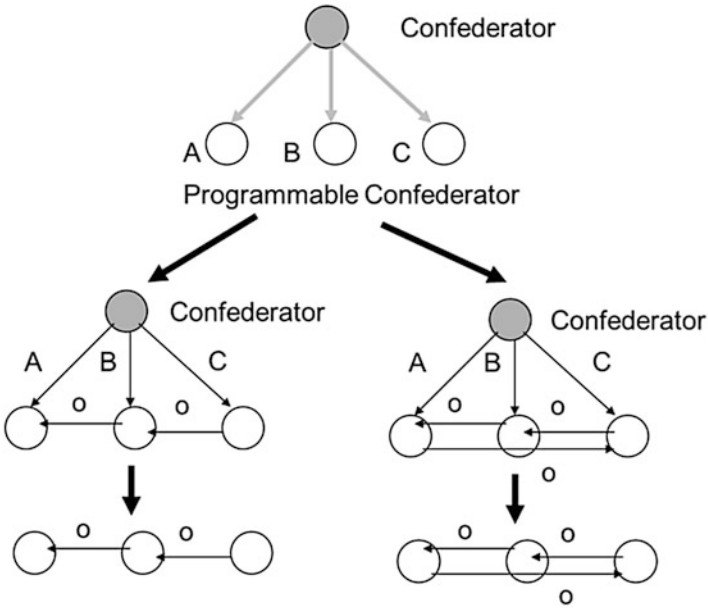


Fig. 15.11 Different rule sets in the confederator can set up different connections among module objects, for example, a linear connection and a cyclic connection

You may make the same confederator to execute a different set of rules given in Fig. 15.10 to set up a cyclic connection among the same types of module objects.

All these rules are executed by the confederator to set up either a linear connection from o_n to o_{n-1}, . . ., and from o_2 to o_1 (in case of the first set of rules) or a cyclic connection from o_n to o_{n-1}, . . ., from o_2 to o_1, and from o_1 to o_n (in case of the second set of rules) as shown in Fig. 15.11. The n channels between the

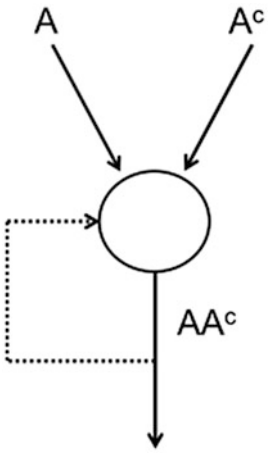


Fig. 15.12 A simple example of autocatalytic reaction networks

confederator and the n objects will be naturally broken when all the objects leave the proximity of the confederator.

Now we use catalytic reaction networks to describe two more complex application scenarios of linear federations among smart objects. A catalytic reaction network is a set of catalytic reactions in which some output of a reaction may work as an input or a stimulus of another reaction. Figure 15.12 shows a catalytic reaction network including a single reaction with two types of smart objects A and A^c as inputs, and their linear federation output AA^c working also as its stimulus. As observed later, this reaction requires no context. Such a catalytic reaction network in which some output of a reaction may work as a stimulus of another reaction is specially called an autocatalytic reaction network. The example in Fig. 15.12 shows the simplest form of autocatalytic reaction networks with a single reaction.

Suppose that A and A^c, respectively, represent a car and a parking space, strictly speaking, the type of cars and the type of parking spaces. These object types are complementary to each other. Figure 15.13 shows the six rewriting rules we use for these two types of smart objects.

Each smart object of type A (or A^c) has the proximity-based federation capability only with A^c (or A) to set up a temporal B (or B^c) channel. The service type B^c denotes the complementary service type of B. Once A and A^c are federated by the reaction, they are connected by an L channel from A^c to A using a mobile phone WiFi connection to form a stable linear federation AA^c, and their states are set to s_1. This system initially requires a single linear federation AA^c with the state of each object at s_1 as a seed, i.e., the first stimulus, to trigger further reactions. The A^c object of this seed federation AA^c can be embedded in the entrance gate of the parking lot, while its A object can be embedded somewhere in the parking space. Since they use L channel communication using the mobile phone WiFi, their distance can be arbitrary. The B and B^c connections are proximity based.

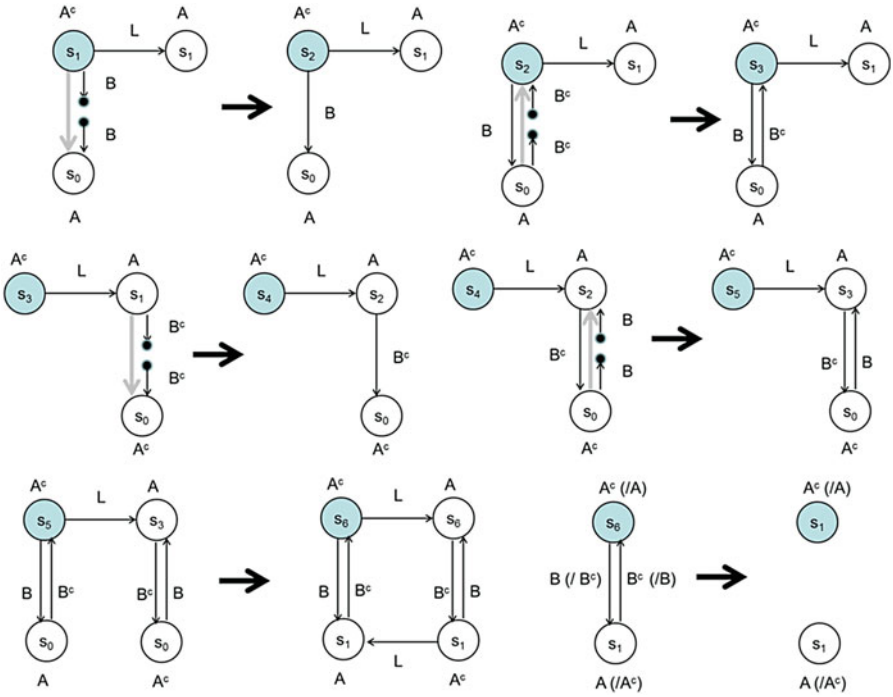


Fig. 15.13 Six graph rewriting rules for the smart objects A and A^c involved in the autocatalytic reaction network in Fig. 15.12

It is always difficult to find out an unoccupied parking space in a huge parking lot. Once a car with a smart object of type A enters a parking lot, it passes nearby the A^c object of the seed AA^c or some other A^c object in an already established federation AA^c. Then the A^c object in this AA^c establishes both B and B^c channels to and from the A object in the car. Through these connections, the A object in the car can ask the A^c object in this partner AA^c through a B^c channel to make the A object in this AA^c find out, in its proximity, some A^c object that is embedded in a unoccupied parking space, and establish both B^c and B channels to and from the found A^c. Then the A object in the car can ask the A^c object in the partner AA^c to set up an L channel from the found A^c object to the A object in the requesting car, then to reset the states of A and A^c in this new federation AA^c to s_1, and finally to break all the B and B^c channels between the two AA^c federations. This process establishes a new linear federation AA^c between the car and the found parking space. Now the A object can get the location of the available space, and guide the driver to park there. This new federation also works as a stimulus for further federations. When a car leaves from the parking space, its A object breaks the L channel to itself. This autocatalytic reaction utilizes the fact that $(AA^c)^c = AA^c$.

Figure 15.14 shows a catalytic reaction network with three composition reactions and one decomposition reaction. In this chapter, we focus on the cases in

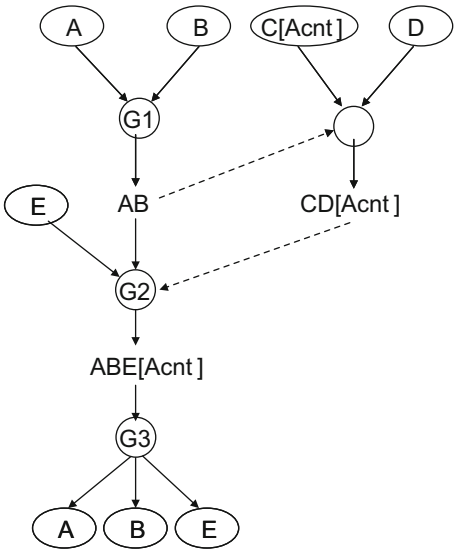


Fig. 15.14 Modeling a smart object federation scenario in a 3D interactive movie theater as a catalytic reaction network

which all the reactions have their contexts, and assume that each reaction can be associated with some physical location. This means that we can design the context of each reaction as some smart object or some linear federation of smart objects and put it at the physical location of this reaction to trigger this reaction.

Figure 15.14 shows a catalytic reaction network that describes the following complex application scenario of smart object federations. A user with a smart social ID card A and a smart member card B passes a check-in gate G1 of an interactive 3D movie theater. This gate sets up a federation between A and B to check if he or she is a registered member of this theater. He or she can pick up a stereoscopic pair of glasses D with an overlaid information display function, and an interactive controller C with an accounting software smart object Acnt. These two smart objects D and C[Acnt] are automatically federated to compose a compound smart object C[Acnt]D with the help of the federation AB as a security key. No user operation is necessary to set up the necessary connection between C[Acnt] and D. While viewing a movie, he or she can interactively issue a purchase order of items appearing in the movie. The software smart object Acnt records these orders. After the movie, he or she passes through the gate G2 with C[Acnt]D, which federates his or her mobile phone E with AB and downloads Acnt from C[Acnt] D to E in this compound object to change it to ABE[Acnt], which enables him or her to send all the purchase orders recorded in Acnt as well as the payment information just by a single click after checking the information. Then the exit gate G3 decomposes the federation into A, B, and E, and deletes Acnt.

The catalytic reaction network modeling enables us to describe complex application scenarios using more than one proximity-based federation of smart objects. Some of these smart objects are mobile devices, and moving from one place to another to be dynamically involved in various federation reactions, and federated

with other smart objects. Some of these reactions are performed only in the existence of another smart object or some smart object federation that works as a catalyst of this reaction. An output federation of some reaction may also works as an input or the stimulus of another reaction.

15.3 Middleware Framework for implementing Catalytic Reaction Networks

Now we need to discuss how to implement each catalytic reaction network using a generic mapping of each reaction to the graph rewriting system modeling of the proximity-based federation of smart objects. Such a generic mapping, if exists, works as a generic middleware framework for the development of complex application scenarios of smart object federations. For simplicity, here in this chapter, we consider only linear federation reactions with contexts. As observed in the confederation example, the use of a context allows us to make it establish any required linear federation among input federations. You can define a set of graph rewriting rules to be executed by the context. Our goal here is the simplification of this programming of the context. We would like to construct any required context as a linear federation of smart objects just by combining generic special types of standard smart objects without writing any new rewriting rules or any codes. This feature allows the easy field setting of contexts and hence complex catalytic reaction networks.

We first consider a composition reaction in Fig. 15.15 with AB and CD as its input linear federation types, ABCD as its output linear federation type, and EF as its stimulus linear federation type. Our basic idea for the generic mapping of such a catalytic reaction to the graph rewriting rule system modeling can be described as follows.

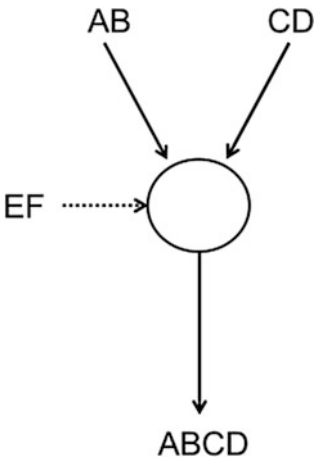


Fig. 15.15 A catalytic reaction with inputs of types AB and CD, and a stimulus of type EF

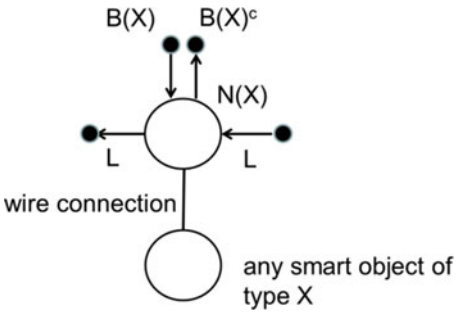


Fig. 15.16 A nucleotide object of type N(X) works as a tag for any object of type X

We assume that each object of type X is tagged, i.e., wire-connected, with a special type smart object called a nucleotide smart object of type N(X), i.e., the nucleotide type for object type X, as shown in Fig. 15.16. This nucleotide object is called the tag object. For each type X, we define its complementary type X^c. The complement of X^c is X, i.e., $(X^c)^c = X$. Each nucleotide smart object of type N(X) has +B(X) and $-B(X)^c$ ports, and −L and +L ports for L channel connections. From the definition, a nucleotide smart object of type N(X) can federate with another nucleotide smart object of type $N(Y)^c$ in its proximity only in the case of $Y = X$. Their federation uses both a B(X) channel from the $N(X)^c$ type to the N(X) type and a $B(X)^c$ channel in the opposite direction.

Nucleotide smart objects can form a linear federation using L channels. Such a linear federation is called a strand. They may form a single strand or a double strand with B(X) and $B(X)^c$ channels between each pair of mutually complementary objects in different strands (Fig. 15.17). In each double strand, two strands have mutually complementary types. For the first strand linear federation of type $X = T_1T_2...T_n$, the type of its second strand linear federation is always $X^c = (T_1T_2...T_n)^c = T_n{}^cT_{n-1}{}^c...T_2{}^cT_1{}^c$. These formations are similar to DNA and RNA single/double strand structures.

For the linear federation of AB and CD under the support of a catalyst EF, we can basically simulate the DNA replication mechanism with a regulation switch to control the replication with a catalyst. The input linear federations AB and CD and stimulus linear federation EF are implemented by the linear federations of their tag objects, i.e., N(A)N(B), N(C)N(D), and N(E)N(F), as shown in Fig. 15.18.

The DNA replication uses the first DNA strand as a template, and the stimulus and inputs dock to the appropriate parts of this first strand to form a concatenation of inputs as an output. The two strands are mutually complementary to each other. Each nucleotide in one strand is paired with its complementary nucleotide in the other strand, and the directions of linear connections are opposite with each other. The stimulus docks to the leftmost part of the first strand to activate this composition. Based on this idea, we design the first strand as a linear federation of nucleotide smart objects as shown in Fig. 15.19, and use it as the context of the reaction. This strand has a stimulus docking part of type $(N(E)N(F))^c = N(F)^cN(E)^c$, and input docking parts of types $(N(A)N(B))^c = N(B)^cN(A)^c$ and $(N(C)N(D))^c = N(D)^cN(C)^c$ with two

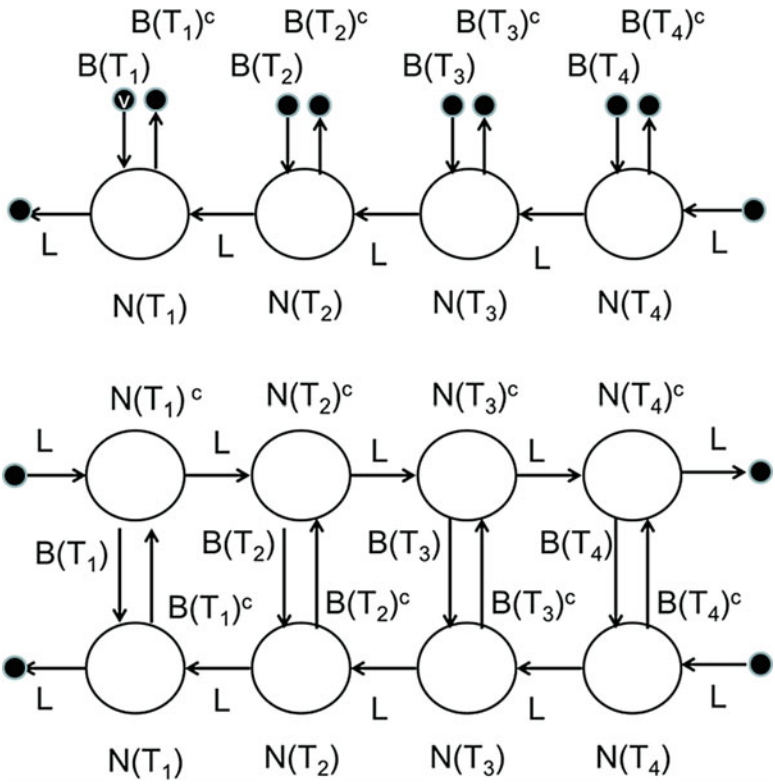


Fig. 15.17 A single strand and a double strand of nucleotide smart objects

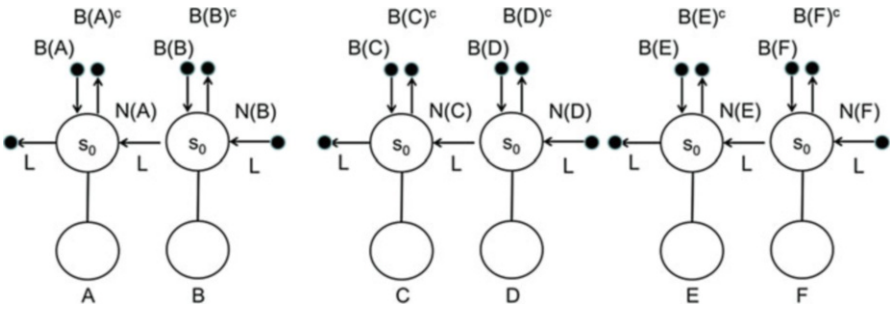


Fig. 15.18 Tagged representations of the inputs AB, CD, and the stimulus EF

kinds of separator smart objects, i.e., a stimulus separator of type N(Ssti) to separate the stimulus docking part from input docking parts, and an input separator of type N (Sinp) to separate consecutive input docking parts of types $(N(A)N(B))^c = N(B)^cN(A)^c$ and $(N(C)N(D))^c = N(D)^cN(C)^c$ from left to right. Therefore, the type of the first strand is designed to become $N(D)^cN(C)^cN(Sinp)\ N(B)^cN(A)^c\ N(Ssti)N(F)^cN(E)^c$.

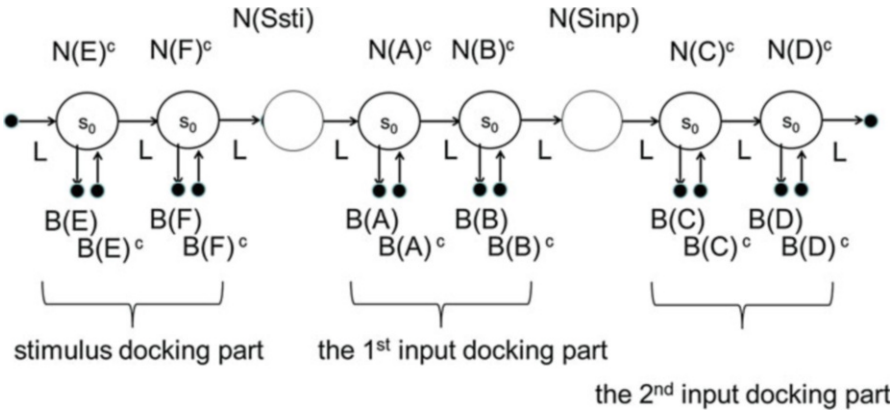


Fig. 15.19 The design of the first strand as the context of the composition reaction in Fig. 15.15

Each type of separator objects has only +L and −L ports. As shown in Fig. 15.19, the directions of L channels in the first strand are all from left to right, while their directions in the second strand to be formed later through docking will be all from right to left. In general, each docking part for a linear federation of type $T_1T_2 \ldots T_n$ is designed as a federation of nucleotide objects of types $N(T_n)^c$, $N(T_{n-1})^c$, …, $N(T_1)^c$ in this order. Namely, its federation type becomes $N(T_n)^cN(T_{n-1})^c \ldots N(T_1)^c$.

The reaction mechanism proceeds as follows. First, the stimulus' tag federation of type N(E)N(F) docks to the stimulus docking part of type $N(F)^cN(E)^c$ (=$(N(E)N(F))^c$) in the context federation of type $N(D)^cN(C)^cN(Sinp)N(B)^cN(A)^c$ N(Ssti)$N(F)^cN(E)^c$ to enable the following process. Then the first input's linear federation N(A)N(B) of tags docks to the first input docking part $N(B)^cN(A)^c$ (=$(N(A)N(B))^c$) of the context federation. This is followed by the docking of the second, the third, and finally the last input linear federations, i.e., the docking of the tag federation of type N(C)N(D) to the docking part of type $N(D)^cN(C)^c$ (=$(N(C)N(D))^c$) in this example. Then the stimulus and all the input linear tag federations are concatenated from right to left to form the second strand, i.e., the strand of type N(E)N(F)N(A)N(B)N(C)N(D) in this example, and then the L channel between the stimulus of type N(E)N(F) and the input concatenation of type N(A)N(B)N(C)N(D) is broken. Finally, both the stimulus tag federation of type N(E)N(F) and the output tag federation of type N(A)N(B)N(C) N(D) are cut off from the context, and all the nucleotide objects are reset to their initial states.

The next step is to design a set of graph rewriting rules to implement this process. For simplicity, here we first show the basic rule set for the docking of stimulus tag federation and input tag federations, and the composition of the output tag federation in Figs. 15.20, 15.21, 15.22, and 15.23.

Figure 15.20 deals with the bridging with B(X) and $B(X)^c$ channels between a pair of compatible nucleotide objects in two strands.

The different states of nucleotide objects denote the following meanings. The state s_0 is the initial state. The states s_1 and s_2 are used in the first strand to denote,

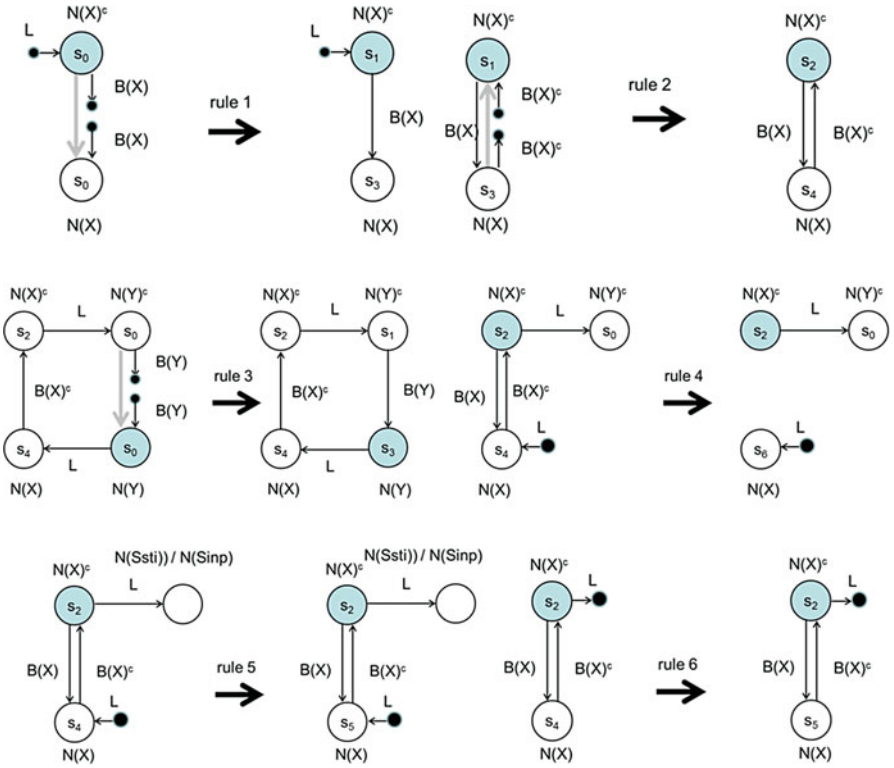


Fig. 15.20 Rewriting rules for the bridging with B(X) and $B(X)^c$ channels between a pair of compatible nucleotide objects in two strands

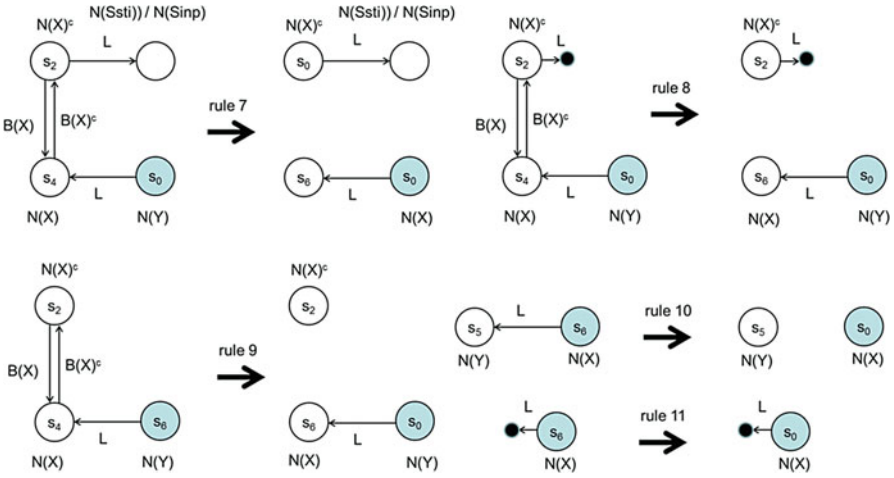


Fig. 15.21 Rewriting rules for completely disconnecting longer or shorter strands to dock

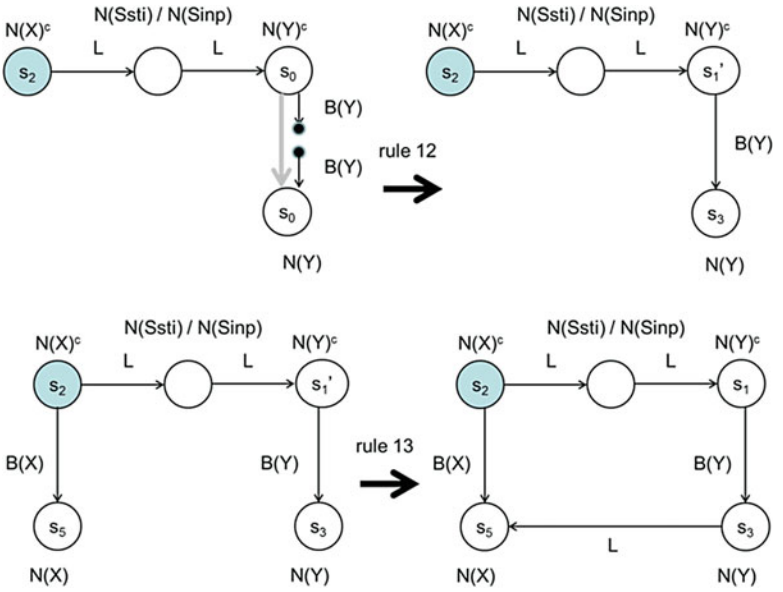


Fig. 15.22 Rewriting rules for concatenating the stimulus and the inputs

respectively, the bridging to the second strand with a single B(x) channel or with double channels B(X) and $B(X)^c$. The state s_3 and s_4 are used only in the second strand and, respectively, correspond to the state s_1 and s_2 in the first strand. The state s_5 in rules 5 and 6 is used in the rightmost nucleotide object in each docking federation to indicate that this federation is completely docked to a docking part. Rule 4 deals with an incompatible docking federation that is shorter than the docking part. The state s_6 indicates that this docking federation needs to be completely disconnected from the docking part.

Figure 15.21 shows the rewriting rules for completely disconnecting a docking federation that is longer than its docking part.

Figure 15.22 shows the rules for concatenating the stimulus and the first input in the second strand with L channels. The state s_1' is used to indicate a transition state between s_0 and s_1 of the leftmost object in each input docking part.

Figure 15.23 shows the rules for undocking the second strand (i.e., rule 14, rule 15, and rule 16), and the rule for breaking the L channel between the stimulus and the output federation (i.e., rule 17). These rules together with rule 18 also initialize all the objects to their initial state s_0. The state s_7 is used to propagate the state initialization from left to right in the second strand.

We need to consider the cases in which an incompatible linear federation tries to dock one of the docking parts. There are three possible cases. The federation may be shorter or longer than the docking part. Some nucleotide object in the federation may not be type compatible with the corresponding nucleotide object in its docking part, i.e., they are not type complementary to each other. The complete undocking

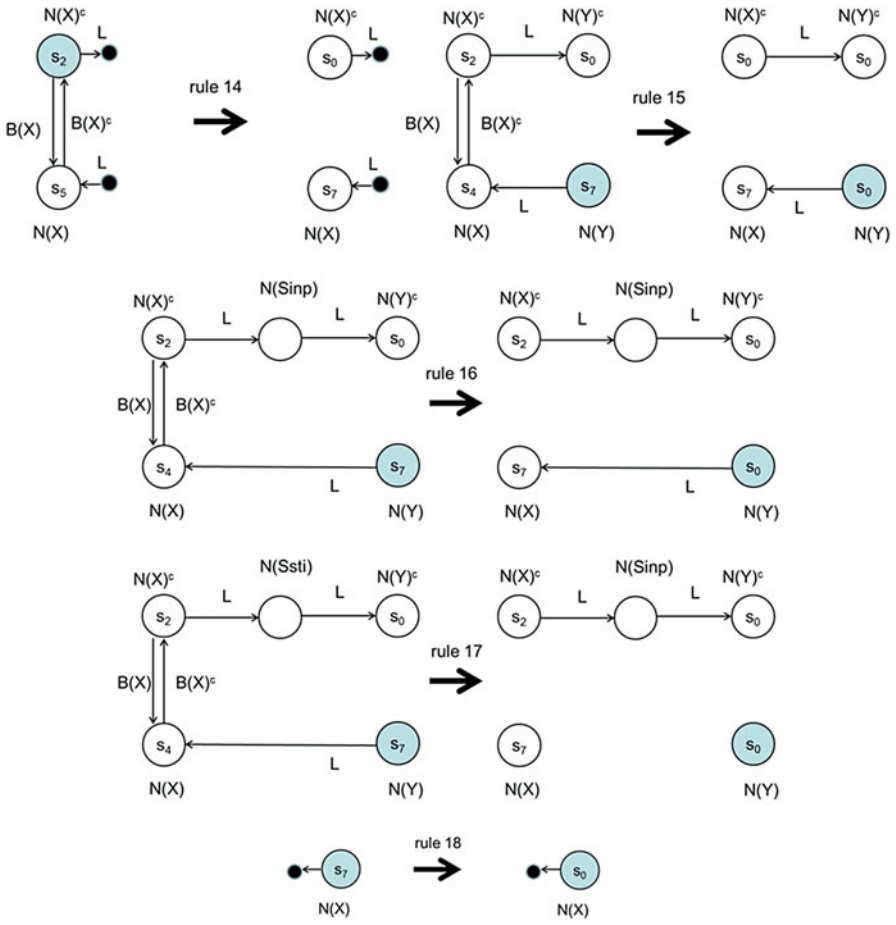


Fig. 15.23 Rewriting rules for undocking the second strand and breaking the L channel between the stimulus and the output federation

of any shorter linear federation is triggered by rule 4, while that of any longer linear federation is triggered by rule 7 and rule 8. Figure 15.24 also shows the rule, i.e., rule 19, to trigger the complete undocking of any docking federation with a type incompatible nucleotide object from its docking part.

You can easily check that no pair of rules out of these 19 rules may conflict with each other, i.e., no pair have their rule conditions be satisfied by the same subgraph.

We also need to consider the instability of wireless connections during the process. We assume that each L channel uses a mobile phone connection and is not broken unless it is explicitly broken by some rule. Both B(X) and $B(X)^c$ channels, however, use proximity relationship to span themselves, and may be unexpectedly broken because of the change of the proximity relationship during the reaction process. This requires the reconnection of the broken channel during the

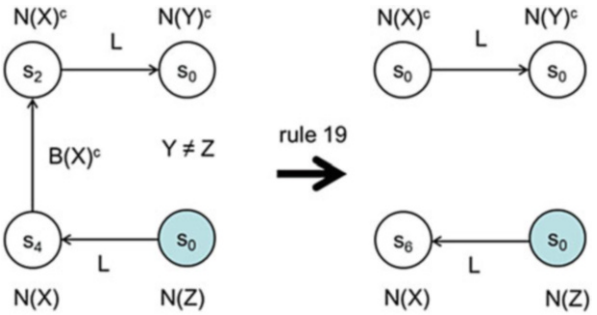


Fig. 15.24 The rule for triggering the complete undocking of any docking federation with a type incompatible nucleotide object from its docking part

reaction process. Such reconnection also requires additional rewriting rules. When a B(X) channel is broken at some paired nucleotide objects, the states of $N(X)^c$ and N(X) objects are both immediately reset to s_0. This makes this pair to reestablish both B(X) and $B(X)^c$ channels. When a $B(X)^c$ channel is broken at some paired nucleotide objects, the states of $N(X)^c$ and N(X) objects are, respectively, reset to s_1 and s_3. This makes this pair to reestablish the $B(X)^c$ channel. However, if one of the paired nucleotide objects becomes completely broken, the process of recovering the connection will not terminate. In such a case, we need to completely initialize the reaction process, and retry the reaction from the very beginning. The rules for the timeout of the connection recovery process are not shown in this chapter.

Now we need to consider the way to construct the context of any given decomposition reaction with a stimulus, an input, and outputs. This context can be easily defined as a linear federation of nucleotide objects in a similar way as the construction of the context of a composition reaction. Figures 15.25 and 15.26, respectively, show a decomposition reaction and its context as a linear federation of nucleotide objects. A special nucleotide smart object N(Sdec) is used to specify the location of the division of the input linear federation.

An additional set of rewriting rules necessary for decomposition reactions can be given, as shown in Fig. 15.27. Rule 20 is the modification of rule 3 to deal with N (Sdec), while rule 21 deals with the triggering of the complete undocking of a shorter input federation that ends at the decomposition separator. Rule 22 deals with the breaking of an L channel for decomposing the input federation into the output federations.

As described above, our middleware framework consists of the way to construct the context linear federation using nucleotide smart objects and the same set of rewriting rules stored in each nucleotide smart object. This set of rewriting rules includes only the 22 rules from rule 1 to rule 22. Separator nucleotide objects have no rewriting rules and never work as rule-activation object. This framework enables us to easily construct any context of a composition/decomposition reaction by linearly federating nucleotide smart objects. For each nucleotide smart object N (X) or $N(X)^c$ of a specific smart object type X, we can use the same prototype nucleotide smart object N or N^c that has the above 22 rules and allows us to manually set its type to N(X) or $N(X)^c$.

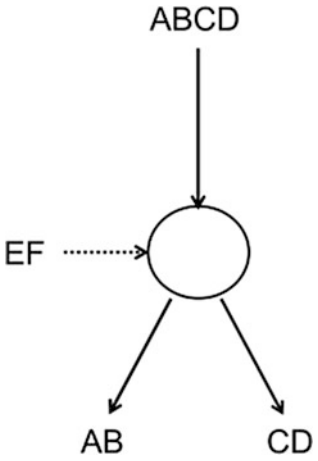


Fig. 15.25 A decomposition reaction with a stimulus

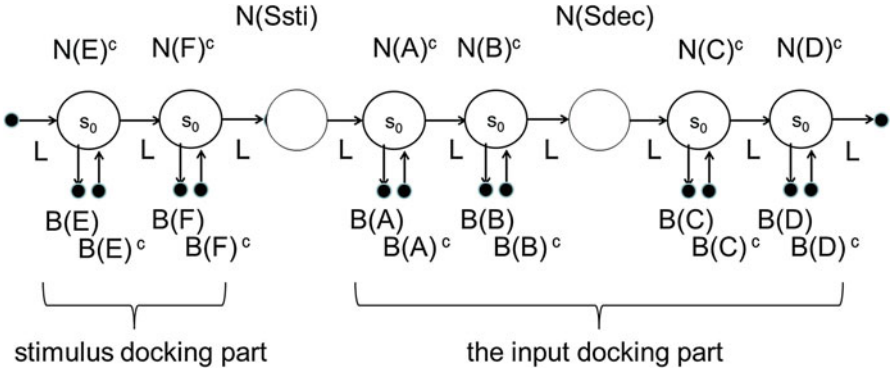


Fig. 15.26 The design of the first strand as the context of the decomposition reaction in Fig. 15.25

15.4 Concluding Remarks

This chapter first pointed out the necessity of a formal model and a middleware framework based on it for describing complex application scenarios using more than one dynamic proximity-based federation reactions of smart objects, and for rapidly developing these applications without any coding. Then we reviewed our three different levels of formal modeling. The first-level modeling formally defines a smart object, and describes the federation between one smart object and another within the scope of the former as the port matching process. The second level describes the dynamic change of federation structures as a graph rewriting system with nodes to represent smart objects and each edge to represent a channel connection between a pair of smart objects. The third level deals with complex application scenarios in which more than one smart object federation are involved, and describes each complex application scenario as a catalytic reaction network.

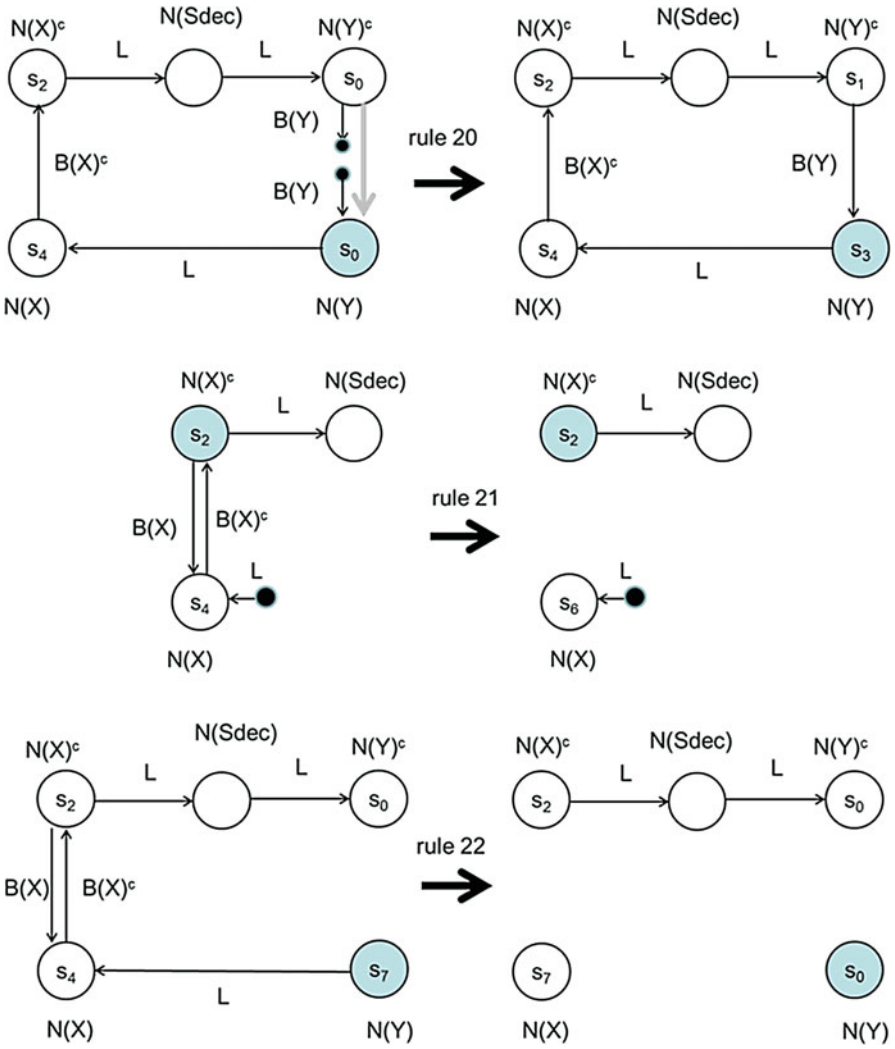


Fig. 15.27 Additional rewriting rules for decomposition reactions

Each composition reaction represents the federation of given input federations to compose an output federation under the help of a catalyst federation. Each decomposition reaction represents the defederation of an input federation into output federations under the help of a catalyst federation. A catalytic reaction network is a set of composition and/or decomposition reactions in which the output of some reaction may work as an input and/or a catalyst of another reaction.

Based on these three levels of formal modeling, this chapter proposed a novel middleware framework for rapidly developing complex applications of the proximity-based federation of smart objects. Our framework uses a special type

of smart objects working as a tag of identifying each different type of smart objects. These tag objects are called nucleotide smart objects. For each different tag, we can use the same prototype tag smart object, i.e., a small device, that allows us to manually set its type. Our framework also defined the whole set of rewriting rules for this prototype tag smart object to execute.

This chapter mainly not only focused on catalytic reactions using contexts, but also showed a simple example autocatalytic reaction network without using any context. Our future research will focus more on complex autocatalytic reaction networks, and the extension of the current version of our middleware framework to deal with reactions without using contexts.

References

1. Milner R (2004) Theories for the global ubiquitous computer. In: Foundations of software science and computation structures, LNCS, vol 2987. Springer, Berlin, pp 5–11
2. Henricksen K, Indulska J, Rakotonirainy A (2002) Modeling context information in pervasive computing systems. In: Mattern F, Naghshineh M (eds) Pervasive 2002, LNCS, vol 2414. Springer, Berlin, pp 167–180
3. Berry G, Boudol G. (1990) The chemical abstract machine. In: Proc. POPL'90, ACM, pp 81–94
4. Cardelli L, Gordon AD (1998) Mobile ambients. In: Nivat M (ed) Foundations of software science and computational structures, LNCS, vol 1378. Springer, Berlin, pp 140–155
5. Păun G (2000) Computing with membranes. J Comput Syst Sci 61(1):108–143
6. Milner R (2001) Bigraphical reactive systems. In: Proc. CONCUR 2001, LNCS, vol 2154. Springer, Berlin, pp 16–35
7. Castagna G, Vitek J, Zappa Nardelli F (2005) The seal calculus. Inf Comput 201(1):1–54
8. Schmitt A, Stefani J-B (2005) The Kell calculus: a family of higher-order distributed process calculi. In: Proceedings of the international workshop on global computing, LNCS, vol 3267. Springer, Berlin, pp 146–178
9. Ueda K, Kato N (2005) LMNtal: a language model with links and membranes. In: Proceedings of the fifth international workshop on membrane computing (WMC 2004), LNCS, vol 3365. Springer, Berlin, pp 110–125
10. Santi P (2005) Topology control in wireless and ad hoc sensor networks. ACM Comput Surv 37(2):164–194
11. Chinara S, Rath SK (2009) A survey on one-hop clustering algorithms in mobile ad hoc networks. J Netw Syst Manag 17(1–2):183–207
12. Dressler F, Akan OB (2010) A survey on bio-inspired networking. Comput Netw 54 (6):881–900
13. Heimbigner D, McLeod D (1985) A federated architecture for information management. ACM Trans Inf Syst 3(3):253–278
14. Edwards WK, Joy B, Murphy B (2000) Core JINI. Prentice Hall Professional Technical Reference, Englewood Cliffs
15. Tanaka Y, Fujima J, Ohigashi M (2004) Meme media for the knowledge federation over the web and pervasive computing environments. In: Proc. ASIAN 2004, LNCS, vol 3321. Springer, Berlin, pp 33–47
16. Tanaka Y, Ito K, Fujima J (2006) Meme media for clipping and combining web resources. World Wide Web 9(2):117–142
17. Tanaka Y (2006) Knowledge federation over the web based on meme media technologies. In: LNCS, vol 3847. Springer, Berlin, pp 159–182

18. Tanaka Y (2003) Meme media and meme market architectures: knowledge media for editing, distributing, and managing intellectual resources. Wiley-IEEE Press, Hoboken
19. Tanaka Y (2010) Proximity-based federation of smart objects: liberating ubiquitous computing from stereotyped application scenarios. In: Setchi R, Jordanov J, Howlett RJ, Jain LC (eds) In: Knowledge-based and intelligent information and engineering systems—14th international conference, KES 2010, Cardiff, UK, 8–10 Sept 2010, Part I. LNCS, vol 6276. Springer, Berlin, pp 14–30
20. Julia J, Tanaka Y (2016) Proximity-based federation of smart objects: its graph-rewriting framework and correctness. J Intell Inf Syst 46(1):147–178
21. Gelernter D (1985) Generative communication in linda. ACM Trans Program Lang Syst 7 (1):80–112
22. Picco GP, Murphy AL, Roman GC (1999) Lime: Linda meets mobility. In: ICSE'99: proceedings of the 21st international conference on software engineering, Los Alamitos, CA, USA, IEEE Computer Society Press, pp 368–377
23. Microsystems S (2001) Javaspaces service specification, version 1.2. https://river.apache.org/doc/specs/html/js-spec.html
24. Microsystems S (2001) Jini technology core platform specification, version 1.2. https://river.apache.org/doc/specs/html/lookup-spec.html
25. Collins J, Bagrodia R (2008) Programming in mobile ad hoc networks. In: Proceedings of the 4th conference on wireless internet, WICON'08, article no. 73
26. Kauffman S (2000) Investigations. Oxford University Press, Oxford

Chapter 16
Edge Computing for Cooperative Real-Time Controls Using Geospatial Big Data

Teruo Higashino

Abstract Recently, sensing technology and Internet of Things (IoT) have much attention for designing and developing affluent and smart social systems. Although huge sensing data are collected in cloud, generally cloud systems are facing poor scalability and difficulty of real-time feedback. In this paper, we focus on geospatial sensing data welled out continuously everywhere and consider how we can treat such huge sensing data. Here, first we introduce the notion of "Edge Computing," and explain its history, features, and research challenge. Then, we discuss about how we can apply this notion for designing scalable IoT-based social systems. As an example, we introduce our recent research work about the development of IoT-based cyber physical systems (CPS). Especially, we focus on safety management in urban districts by estimating up-to-date (real-time) population distribution and creating pedestrian mobility in urban districts from inaccurate sensing information. In urban areas, we might only be able to use heterogeneous sensors with different accuracy for crowd sensing. Thus, we propose a method for creating realistic human mobility using such heterogeneous sensors, and explain techniques to reproduce passages, add normal/emergency pedestrian flows, and check efficiency of evacuation plans on 3D map so that local governments can make efficient evacuation plans. We also introduce a method for the prediction of vehicle speeds in snowy urban roads. We believe those IoT-based social systems have enough scalability and dependability using the edge computing paradigm.

Keywords Edge computing • Cyber physical systems (CPS) • Geospatial big data • Internet of Things (IoT) • Urban planning

T. Higashino (✉)
Graduate School of Information Science and Technology, Osaka University,
Yamadaoka 1–5, Suita, Osaka 565-0871, Japan
e-mail: higashino@ist.osaka-u.ac.jp

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_16

16.1 Introduction

Recently, sensing technology and Internet of Things (IoT) have much attention for designing and developing affluent and smart social systems. Cloud computing is very popular and big data are collected at central cloud servers and their mining information are fed back to users. It makes large influence for business models of enterprises, and creates new individual life styles. However, with the progress of M2M and IoT, a huge amount of big data from giga ordered sensors might be generated from IoT-based social systems such as ITS (self-driving vehicles and collision avoidance), smart grid (power control), and crowd sensing systems from human beings with mobile devices (see Fig. 16.1). In such a situation, it is difficult to store all of those big data in central cloud servers with reasonable costs. In smart grid systems, tens of milliseconds order's controls are required in order for stable power control. Also, neighboring geospatial data might have strong correlation in power control systems while distant geospatial data might not have so strong correlation. Similar situations arise for social systems and applications using geospatial data such as crowd sensing data, floating car data, and personal data from mobile phones. Thus, it might be suitable for storing such geospatial data in local "edge servers" for quickly providing local dependent services in cooperation with both neighboring edge servers and central cloud servers. As a platform for IoT-based social systems, the notion of "Edge Computing" [1] has much attention where small-scale edge servers are located at users' neighborhood and they cooperate with the central cloud servers (see Fig. 16.2).

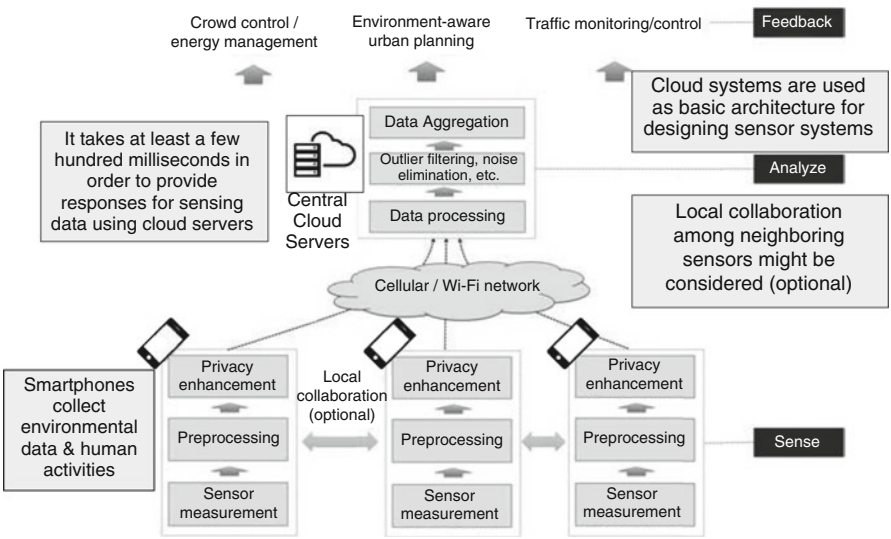


Fig. 16.1 Crowd sensing and human behavior sensing using central cloud servers

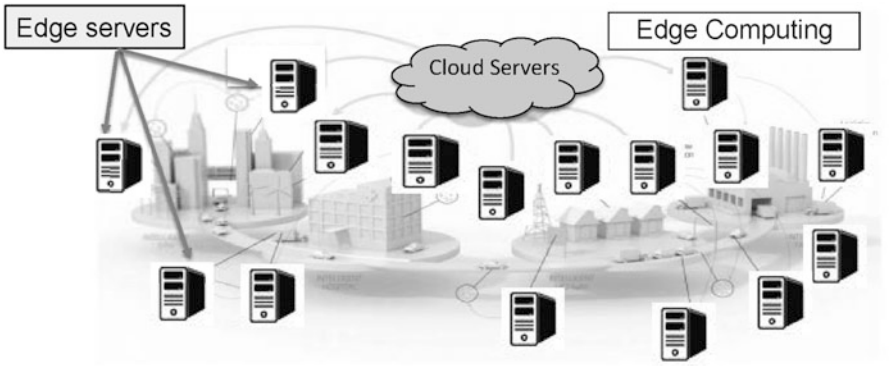


Fig. 16.2 Edge computing paradigm

In this paper, first we introduce the notion of "Edge Computing," and explain its history, features, and research challenge. Edge computing can reduce communication delay between edge servers and users. Thus, edge servers can provide quick responses to users. Also, edge servers can represent local computation at users' devices such as smartphones and small sensors. Soon, edge computing platforms can be used for several types of social systems such as ITS, smart grid, energy management, city management, smart city, health care, game, and AR.

Secondly, we propose a method combining curation mechanisms with simulation mechanisms for designing and developing scalable IoT-based social cyber physical systems (CPS). Especially, we focus on safety management in urban districts by estimating up-to-date (real-time) population distribution and pedestrian mobility in urban districts. In urban areas, we might only be able to use inaccurate sensing information from heterogeneous sensors with different accuracy for crowd sensing. Thus, we propose a method for creating realistic human mobility from such heterogeneous sensors, and explain techniques to reproduce passages, add normal/emergency pedestrian flows and check efficiency of evacuation plans on 3D map. For this purpose, we introduce our crowd sensing technique using laser range scanner (LRS) and smartphone-based crowd detection technique. After that, we introduce our technique for pedestrian flow estimation using those heterogeneous sensors where we propose a realistic pedestrian mobility called urban pedestrian flows (UPF) mobility in urban districts [2]. We also discuss about a sensor allocation problem for human mobility detection. Based on those techniques, we explain about how we can create efficient emergency plans based on such heterogeneous crowd sensing devices. In order to consider the guide for commuters unable to get home in wide urban areas, we need huge geospatial sensing data where neighboring geospatial data might have strong correlation while distant geospatial data might not have so strong correlation. Thus, we explain how the edge computing paradigm can be used for safety management in wide urban districts.

Thirdly, we introduce our recent IoT-based CPS research work for the prediction of vehicle speeds in snowy urban roads, which has been carried out as the Japanese

governmental CPS Integrated IT Platform (CPS-IIP) Project. We define "the deceleration amount of vehicle speed" as the difference between the average vehicle speeds on snowy seasons and non-snowy seasons for each target road segment. We propose a vehicle speed model to predict the deceleration amount of vehicle speed for each road segment using the regression analysis technique. In order to predict the deceleration amount of vehicle speed, weather information and vehicular traffic data are treated as the dominant factors in the regression model equation. Then, we represent the deceleration amount as the linear regression expression of those explanatory variables. For busy road segments with frequent traffic jam in the city centers and main road segments with enough traffic amount but infrequent traffic jam, we have proposed a method for estimating the deceleration amount of vehicle speed precisely. In snowy regions, the deceleration of vehicle speeds in bad weather often makes heavy traffic jams and hindrance. By cooperating with neighboring cities, the local government of each city can make an efficient snow removal plan and provide adequate traffic information to residents. We expect the edge computing paradigm can also be used for this research.

The paper is organized as follows: First, in Sect. 16.2, we explain the notion of "Edge Computing" and its history, features, and research challenge. Then, in Sect. 16.3, we introduce our recent IoT-based CPS research work for safety management in urban districts. After that, in Sect. 16.4, we also introduce our recent research work for the prediction of vehicle speeds in snowy urban roads. Finally, in Sect. 16.5, we conclude the paper.

16.2 Edge Computing

In the cloud computing paradigm, computing applications, data, and services are allocated to central cloud servers. Basically those cloud servers are located far from end users. Thus, it takes much time for end users to communicate with those cloud servers. On the other hand, as we have described in Sect. 16.1, IoT-based social systems often require quick responses for end users. Also, neighboring geospatial data might have strong correlation for such social systems. In "Edge Computing" paradigm, computing applications, data, and services are allocated to "edge servers" at peripheries of the network as shown in Fig. 16.2 where edge servers collect sensing data from their neighboring end users, store those data in their own edge servers, and provide services quickly to the end users. In the edge computing paradigm, edge servers autonomously collect neighboring geospatial data and decide local responses. Cloud servers cooperate with edge servers and regulate unbalance among edge servers for improving the entire quality of services. It is not a client–server model. It is a multi-layered hybrid model.

16.2.1 Real-Time Control of Geospatial Data

Recently, there are a lot of demands for requiring real-time responses for applications using geospatial data. In near future, much more accurate global navigation satellite system (GNSS) are expected to become popular. In such a situation, many vehicles, bicycles, and pedestrians might have sensors for identifying their own real-time positions. If such positioning data can be used, responses of several tens of milliseconds might be required for collision avoidance and autonomous running. More quick responses can lead more intellectual responses for preserving road safety. However, if we send those positioning data to central cloud servers, it takes at least several hundreds milliseconds to receive responses from those cloud servers. Thus some hybrid mechanisms combining central cloud servers and edge servers are needed where the edge servers have responsibility for tens of millisecond ordered responses for road safety and the central cloud servers have responsibility for total traffic controls in towns. In smart grid research, we also need similar real-time responses. Assume that several millions of smart meters are installed in a town. The supply of stable electricity requires 10 ms ordered power controls. Although all the data from those smart meters can be stored in cloud, 10 ms ordered responses are not possible. Similar situations also arise for research about smart city, future BEMS, environmental controls, and so on. Thus, the research for real-time controls combining cloud servers and autonomous edge servers is one of the new and important research themes in edge computing.

16.2.2 History and Research Challenge of Edge Computing

The term “Edge Computing” is not new. It appeared around 2002 associated with content deliver networks (CDN). Akamai Tech. Inc. provides cloud services. It has deployed more than 100,000 servers in more than 90 countries. It provides one of the world’s largest distributed computing platforms. Those servers keep several types of contents. In this early edge computing, edge servers correspond to CDN ones. “P2P Computing” started from around 2000. It is considered as the main precursor of edge computing. File sharing systems such as Napster and Kazaa are typical P2P systems, and distributed hash tables (DHTs) can be used for developing scalable P2P systems. Recently, there are a lot of efforts to combine P2P and cloud computing architectures. If peers contribute and combine their resources with cloud, the costs of cloud services can be reduced. Peer-assisted services construct hybrid architectures combining peer and cloud resources. Nano-datacenters, micro-clouds, community clouds, and edge clouds have been developed (for example, home alliance). “Fog Computing” substantially overlaps with edge computing. It is defined by CISCO. It extends cloud computing and services to edges of the network. Fog services may be hosted at the network and/or end devices such as set-top-boxes and access points. Fog computing is more related with networks

while edge computing is more related to P2P computing. Mobility is an important issue for IoT-based systems. Thus, "Mobile Edge Computing" is also studied.

There are several research challenges in edge computing. We need to consider edge architecture depending on target applications such as ITS, smart grid, energy management, smart city, and health care. We also need to consider efficient mechanisms for data aggregation. The paper [3] provides a survey about (1) routing protocols for data aggregation and (2) data aggregate functions for data mining. The paper [4] defines attenuation functions in accordance with the number of hops from the source. Closer nodes can obtain larger weights so that spatial neighboring relationship can be well considered. Scalability and real-time control are also important research issues. Since geospatial big data welled out continuously everywhere cannot be stored in cloud servers, we need to study the followings:

(a) what information processing mechanisms (hybrid mechanism/hierarchical mechanism) are needed,
(b) what geospatial analyzing mechanisms are useful,
(c) how and when we can discard obtained big data welled out continuously, and
(d) what mining data we should store for intellectual quick responses from edges, and so on.

The paper [5] summarizes techniques for crowd sensing and human behavior sensing. Finally, we need to consider security and privacy issues so that IoT-based social systems can be accepted from many residents. In the paper [6], privacy-preserving data aggregation protocols are summarized. Also, the paper [7] provides data aggregation mechanisms for several types of node failure. The paper [8] provides a technique for privacy-preserving data aggregation for mobile/opportunistic sensing.

16.3 Safety Management in Urban Districts

A small fire occurred at underground Metro Osaka Station in 2012 where a very small area in a warehouse under the platform was burned due to electric leak. Seventeen persons were conveyed to hospitals by ambulances. More than 3000 people tried to evacuate to the ground from underground. They could not find where the fire occurred and which directions they should run away. The number of pedestrians varies depending on weekday or weekend, and rush hour or daytime. Up-to-date crowd control at large stations, shopping malls, and underground malls are very important for disaster mitigation.

16.3.1 Urban Sensing from Inaccurate Sensing Information

In order for the safety management in urban districts, it is desirable that we can grasp pedestrian behaviors in target urban districts precisely. In order for grasping pedestrian mobility in urban districts precisely, we might need to place high-precision cameras with high density for the target urban districts. However, such placement might not be possible in many cases because of their costs and privacy problems. We might be able to use only heterogeneous sensors with inaccurate sensing ability. Here, we consider methods for creating pedestrian mobility from such inaccurate sensing information with reasonable precision and costs. In Fig. 16.3, we show a method for roughly grasping pedestrian behaviors in target urban districts. In the method, we use the following procedures:

(1) We enumerate movable routes from commercial city maps.
(2) Then, population distribution and mobility for each 100 m cell are estimated.
(3) In order to estimate population distribution of crowds in buildings, underground malls, and cities precisely, we use heterogeneous sensors such as LRS, cameras, and smartphones.
(4) We also use an urban simulator for creating pedestrian mobility in normal situations and emergency situations of urban districts (see Fig. 16.4).

The proposed method is an IoT-based CPS. Based on the notion of CPS, we create techniques to reproduce passages, add normal and emergency pedestrian

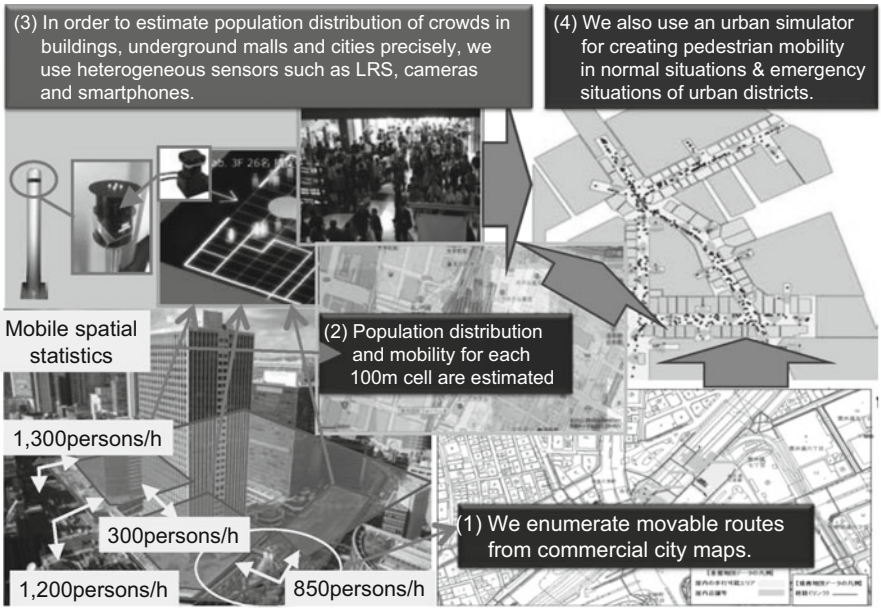


Fig. 16.3 Safety management in urban districts based on edge computing paradigm

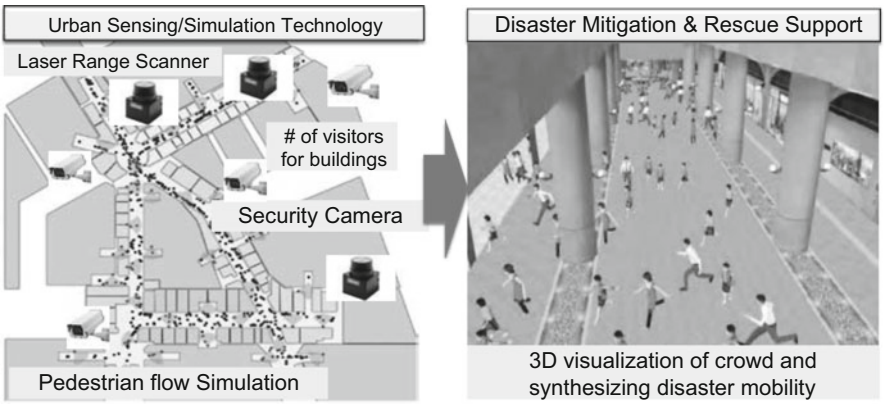


Fig. 16.4 Safety management in urban districts

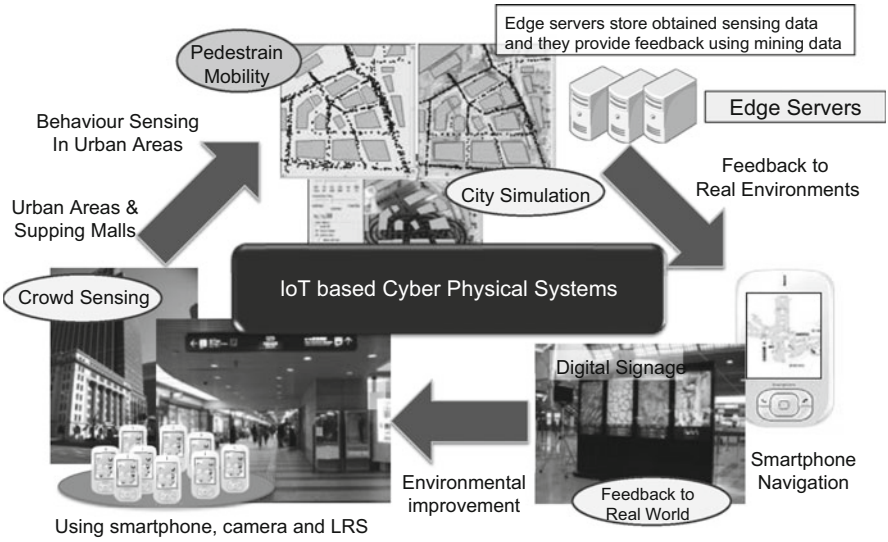


Fig. 16.5 IoT-based cyber physical systems (CPS)

flows, and check efficiency of evacuation plans on 3D map so that local governments can make efficient evacuation plans (e.g., evacuation planning for fire and flooding of target underground malls). Depending on date/time and real-time situations of disaster victims, we might create the corresponding pedestrian mobility and carry out several types of evacuation plans by simulation. By storing several types of simulation results in cloud (or edge) servers in advance, we can provide suitable evacuation routes and disaster information to their smartphones in real-time (see Fig. 16.5).

16.3.2 Crowd Sensing Using Heterogeneous Sensors

Here, we focus on up-to-date crowd control at large stations, shopping malls, and underground malls, and explain about our techniques for (1) LRS-based crowd sensing and (2) smartphone-based crowd detection. Note that the aim of this paper is not to improve the accuracy of sensing devices but to roughly grasping human behaviors in urban districts from heterogeneous sensors.

16.3.2.1 Laser Range Scanner-Based Crowd Sensing

In urban areas, several cameras might be installed for ensuring safety of the city. Several techniques for counting and/or tracking the number of pedestrians in urban areas have been proposed so far [9]. The paper [10] detects specific shape such as human face and counts the number of pedestrians. The paper [11] proposes a regression analysis-based method by considering the relationship between the image feature amount and the number of pedestrians. Those techniques are very useful. However, there exist privacy problems when we use cameras. Thus, we have proposed a LRS-based pedestrian tracking method [12].

LRS-based pedestrian tracking has been considered a reasonable solution for crowd behavior sensing in public space because it has privacy-preserving feature. LRS only captures distance to target pedestrians and thus tracking can be completed in an anonymous manner. Although it has powerful tracking capability, the number of pedestrians in crowded regions is often underestimated due to the occlusion problem when we measure human bodies. We might temporally lose back pedestrians when multiple people cross each other. Also, many people might stand in front of interesting booths where foreground people can be detected while inner/back people cannot be detected. To cope with this problem, we have built an empirical model that identifies the relationship between actual crowd density and the number of pedestrians captured by LRSs. In the paper [12], we propose a method combining a pedestrians' trace detection method and a crowded cell detection method, and estimate the total number of people in crowded regions.

The first step of human crowd detection is to find locations of individual pedestrians based on raw measurements from LRSs. At first, we identify stationary objects like walls and obstacles. For this purpose, we collect distance measurements from LRSs when there are no pedestrians in the environment, and store those values in the database. After that, we identify the reflection points that come from moving human bodies. In human body detection, the system first compares each LRS measurement d_i with the background distance $\widehat{d_i}$ in the corresponding direction. If the difference between d_i and $\widehat{d_i}$ is less than a pre-defined threshold, the system regards the reflection point to be formed by a stationary object and thus excludes it from the set of reflection points. Then, the system estimates the body of each pedestrian. The body of each pedestrian usually forms multiple reflection points in contiguous directions. Thus, the system seeks a cluster of reflection points

representing a single pedestrian. In a case that multiple LRSs are deployed in the environment, the system independently applies the above algorithm for the measurements. Thus several human locations can be generated for a single pedestrian when the coverage areas of multiple LRSs overlap. In such a case, the system merges human locations within 25 cm into one, and regards the centroid of those human locations as the position of a single pedestrian.

When the system connects sequential multiple traces, some sub-traces might be lost due to the occlusion problem. In such a case, the system fills vacant sub-traces so that two disjoint traces can be merged by considering the moving direction of two disjoint traces and their speeds. Since measurement intervals of LRSs are typically a few tens of milliseconds (e.g., 0.025 s for UTM-30LX [13]), the maximum distance that a pedestrian can move during contiguous time steps is no more than 10 cm. It is usually much less than the minimum distance to neighboring pedestrians, and thus we can accurately associate the human locations that belong to the same pedestrian by finding the nearest human location at the previous time step.

In sparse areas, we can fill vacant sub-traces of pedestrians in most cases even if the occlusion occurs. On the other hand, at interesting booths, many people might exist in narrow areas. Thus, they might not be able to be detected accurately due to occlusion by other pedestrians. In such a case, the system might frequently miss the presence of human crowds. Thus, the next step is to detect crowded cells and estimate their cell densities so that we can estimate the number of people in a target area. As a solution to this problem, we have built an empirical model that identifies the relationship between the estimated crowd density and the actual crowd density based on our preliminary simulation experiments. Then we use this empirical model to estimate the density of pedestrians in those cells.

We have conducted simulations to obtain the empirical model for this crowd density estimation. In the simulation, as shown in Fig. 16.6, a target area of 30 m × 30 m is divided into grid cells of 2 m × 2 m. We have deployed five LRSs at all the corners as well as at the center of the target area. We have synthesized several

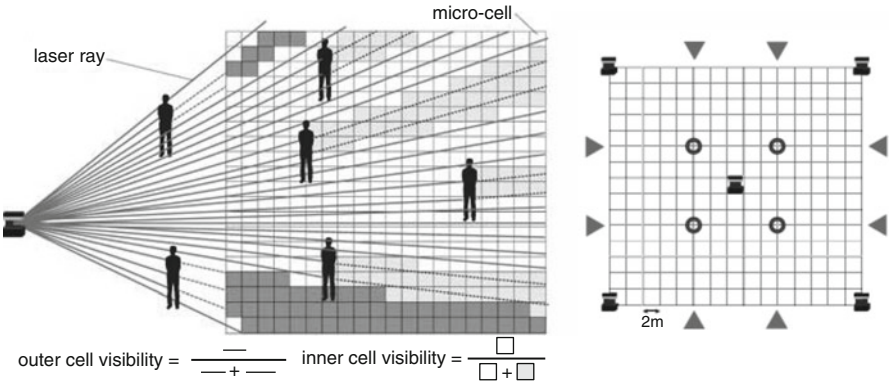


Fig. 16.6 Detection of crowded cells

human mobility models. We divide each grid cell into smaller square sub-regions of $0.2\,\mathrm{m} \times 0.2\,\mathrm{m}$ (i.e., micro-cells). We define that a micro-cell is covered if at least one laser ray from five LRSs passes over it. The inner cell visibility (ICV) is defined by the ratio of the micro-cells that are covered by laser rays over all of the micro-cells. If the value of the ICV is low, the pedestrian density of the corresponding cell is expected to be rather high. The ICV can be used for estimating the pedestrian density of each cell. On the other hand, the outer cell visibility (OCV) is defined as the visible ratio of target cells from outside using LRS. The OCV can be used for estimating the areas of crowded cells. We have carried out the simulation about 30,000 cell samples. Then, we empirically construct the visibility-density model, and estimate the number of people in the target crowded cells. From the experiments, about 70–80 % of human crowd detection becomes possible while the naive algorithm can only achieve 6–28 % of human crowd detection for crowded situations (for details, see [12]).

16.3.2.2 Smartphone-Based Crowd Detection

Next, we introduce our approach for estimating crowd density and smoothness of pedestrian flows in public space by participatory sensing with mobile phones [14].

By extracting the motion-based features and the audio-based features from the accelerometer readings and audio recordings, respectively, each phone classifies the behavior of its surrounding crowd into four categories: (i) low density (L), (ii) medium density (M), (iii-a) high density with smooth flows (H/Sm), and (iii-b) high density with intersections (H/Cr). The low density (L), medium density (M), and high density (H) denote (1) less than 1.0 persons/m^2, (2) 1.0–2.5 persons/m^2, and (3) more than 2.5 persons/m^2, respectively. The high density with smooth flows (H/Sm) denotes that the ratio of the moving directions with less than 45° between neighboring two persons is more than or equal to 70 %. It means that most of pedestrians move similar directions. The high density with intersections (H/Cr) denotes the ratio is less than 70 %. It means that many pedestrians move different directions in crowded situations.

From our observation, as the crowd density increases, low frequency components below 10 KHz exhibit larger power. To clarify this difference, we also show the average amplitude of each frequency component over all the collected audio recordings in Fig. 16.7. Due to the crowd noise, especially the frequency components below 2 KHz get significantly larger under the higher densities. Based on the observation, we extract frequency components below 2 KHz from the audio recordings of the recent 60 s, and calculate the sum of all the amplitude values of those frequency components. The estimation results from multiple phones are associated with their phone locations and collected to a server. By integrating data from those multiple phones, the system estimates the situations of pedestrian flows at each region. The congestion levels could be reliably distinguished by a machine learning algorithm. We employ the k-nearest neighbor algorithm to construct a classifier to estimate the congestion categories based on the audio-based features.

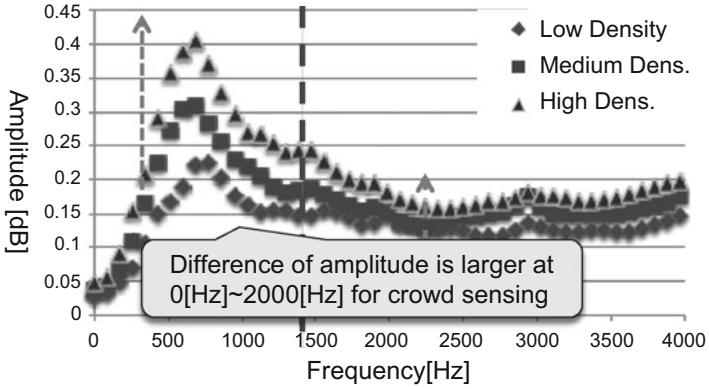


Fig. 16.7 Crowd sensing using noise detection of smartphones

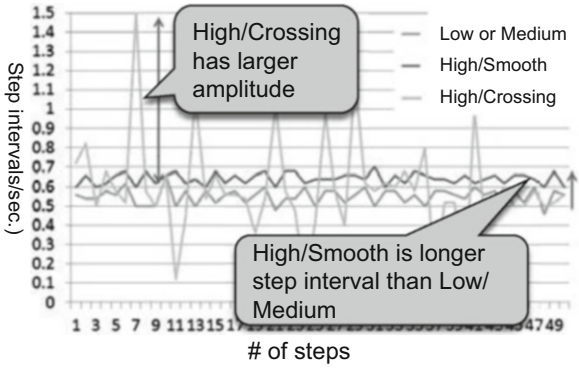
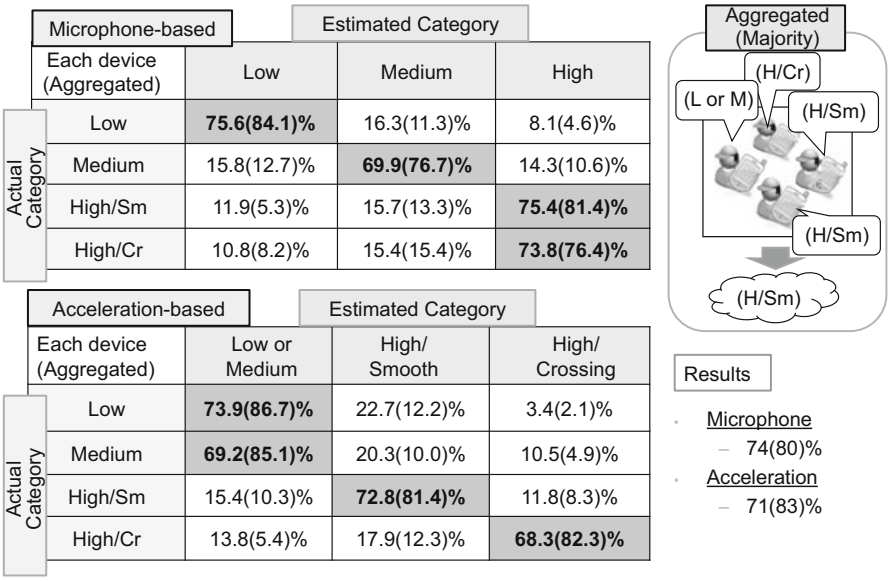


Fig. 16.8 Crowd sensing using accelerometers of smartphones

Analyzing the measurement data from our preliminary experiment, we have found that time intervals between the walking steps have strong correlation with the congestion levels. Figure 16.8 shows typical examples of step intervals with different congestion categories. As shown in the figure, the pedestrians walk with almost regular step intervals in the crowd categories with L, M, and H/Sm. On the other hand, under the category H/Cr, the step intervals significantly vary since they often slow down or even stop to avoid collision with his/her surrounding pedestrians who walk toward different directions. In Fig. 16.8, the H/Cr category shows larger amplitude for step intervals. In addition, the average step intervals with H/Sm and H/Cr tend to be longer than those with L and M. This is why pedestrians need to walk at a bit slow and similar speeds with their preceding pedestrians when they walk through in such crowded situations. In Fig. 16.8, the categories H/Sm and H/Cr show longer step intervals.

Figure 16.9 shows the results of crowd sensing using noise detection and accelerometers of smartphones. Here, the crowd sensing methods using noise detection and accelerometers of smartphones are denoted as "microphone-based



Microphone-based

Actual Category \ Each device (Aggregated)	Low	Medium	High
Low	**75.6(84.1)%**	16.3(11.3)%	8.1(4.6)%
Medium	15.8(12.7)%	**69.9(76.7)%**	14.3(10.6)%
High/Sm	11.9(5.3)%	15.7(13.3)%	**75.4(81.4)%**
High/Cr	10.8(8.2)%	15.4(15.4)%	**73.8(76.4)%**

Acceleration-based

Actual Category \ Each device (Aggregated)	Low or Medium	High/ Smooth	High/ Crossing
Low	**73.9(86.7)%**	22.7(12.2)%	3.4(2.1)%
Medium	**69.2(85.1)%**	20.3(10.0)%	10.5(4.9)%
High/Sm	15.4(10.3)%	**72.8(81.4)%**	11.8(8.3)%
High/Cr	13.8(5.4)%	17.9(12.3)%	**68.3(82.3)%**

Fig. 16.9 Results of crowd sensing

method" and "acceleration-based method", respectively. In the microphone-based method, each smartphone can classify the categories of L, M, and H with the accuracy of 70–76 % (their average is 74 %). In our method, the system collects multiple sensing data from neighboring smartphones, aggregates them, and finds their majority. By aggregating them, those smartphones can cooperatively classify the categories of L, M, and H with the accuracy of 76–84 % (their average is 80 %). On the other hand, in the acceleration-based method, each smartphone can classify the categories of (L or M), H/Sm and H/Cr with the accuracy of 68–74 % (their average is 71 %). By aggregating multiple sensing data from neighboring smartphones, those smartphones can cooperatively classify the categories of (L or M), H/Sm and H/Cr with the accuracy of 81–87 % (their average is 83 %).

In real usage, the positions of smartphones might be different. Some pedestrians might have their smartphones by hands, in their pockets or in their bags. The reliability of estimation is different from the positions of smartphones. We roughly estimate the positions of smartphones and apply more intellectual aggregation methods. Then, we can achieve 91 % as the average estimation ratio for classifying the four categories of L, M, H/Sm, and H/Cr; for details, see [14].

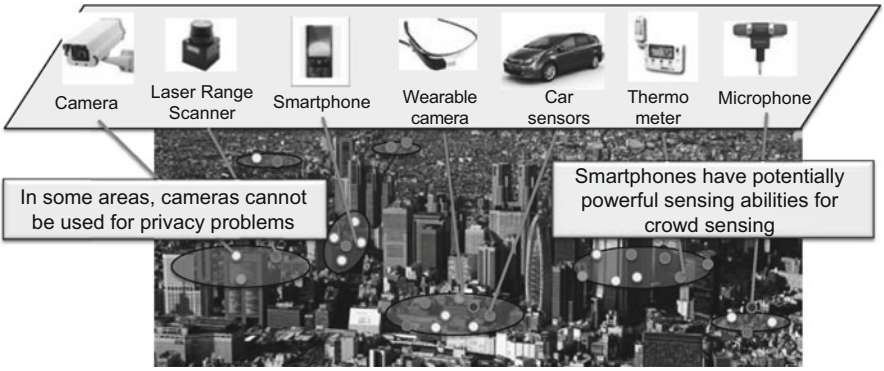


Fig. 16.10 Human mobility generation using heterogeneous sensors

16.3.3 Pedestrian Flow Estimation Using Heterogeneous Sensors

In real urban areas, the number (density) of pedestrians varies depending on time and locations. We might only be able to use heterogeneous sensors for crowd sensing (see Fig. 16.10). Here, we introduce our method to create realistic human mobility using sensing data from heterogeneous sensors, which reproduce the walking behavior of pedestrians in urban areas. In the method, we can only use inaccurate information about the densities of pedestrians observed at multiple observation points. We do not care about how such densities of pedestrians can be obtained. We can use several types of heterogeneous sensors as shown in Fig. 16.10.

16.3.3.1 Urban Pedestrian Flows Mobility

In [2], we have proposed a method to create the UPF mobility scenarios from given densities of pedestrians observed at several observation points. Our method derives a UPF mobility scenario that reproduces the walking behavior of pedestrians consistent with the observed densities, using linear programming (LP) techniques. The method targets reproduction of the walking behavior of pedestrians in city sections, stations, shopping malls, and so on. Given the average densities of pedestrians on certain streets, which can be easily obtained by fixed point observations, and a set of walking paths pedestrians are likely to follow, the method determines flows of pedestrians using linear programming (LP) techniques. Also, the maximum error between the observed density and the corresponding derived density is minimized so that we can reproduce realistic movement of pedestrians.

For creating the UPF mobility, we observe node densities at multiple observation points, enumerate pedestrians' moving routes for the target area (e.g., a route

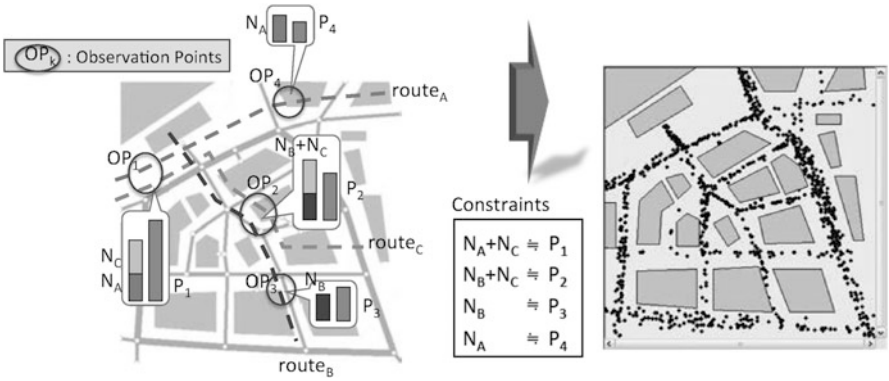


Fig. 16.11 Generation of urban pedestrian flows (UPF) mobility

from a station to a department store). Figure 16.11 shows blocks for a 500 m × 500 m area in front of Osaka Station, Japan. In Fig. 16.11, we specify multiple observation points ($OP_1, OP_2, OP_3, and OP_4$) and count the number (or density) of pedestrians ($P_1, P_2, P_3, and\ P_4$) for the corresponding observation points, respectively. At the same time, we enumerate pedestrians' moving routes ($route_A, route_B$, *and* $route_C$) for the target area. For the case of Osaka Station, there are several departure/arrival points for pedestrians' moving routes such as stations, department stores, major office buildings, and hotels. If there are n departure/arrival points, we might assign $n \times (n-1)$ shortest paths by enumerating all possible combinations as the pedestrians' moving routes. Let N_A, N_B, and N_C denote the numbers (or densities) of pedestrians walking along with $route_A$, $route_B$, and $route_C$, respectively. In Fig. 16.11, the number (or density) of pedestrians at the observation point OP_2 must be close to the sum $N_B + N_C$ of the numbers (or densities) of pedestrians walking along with $route_B$ and $route_C$. Such constraints for the four observation points are described in Fig. 16.11. We can describe the constraints for all the observation points as linear constraints for a linear programming (LP) problem. We can specify the objective function for the LP problem so that the numbers of pedestrians derived from the LP solver are close to the numbers of pedestrians observed at the observation points. If we can only use inaccurate sensors for counting the numbers of pedestrians at specific observation points, we might specify the constraints of the LP problem so that the differences (observable errors) between the derived numbers of pedestrians and the corresponding observed ones for such observation points might be a bit larger than the ones for accurate sensors. Using a LP solver, we can estimate the numbers (or densities) of pedestrians for all pedestrians' moving routes $route_A$, $route_B$, and $route_C$, which can minimize the observable errors.

In our experiments, we have measured the average densities of pedestrians on 33 streets in a 500 m × 500 m area in front of Osaka Station for about a half hour. The maximum error between the observed densities and derived densities was only 9.09 % [2].

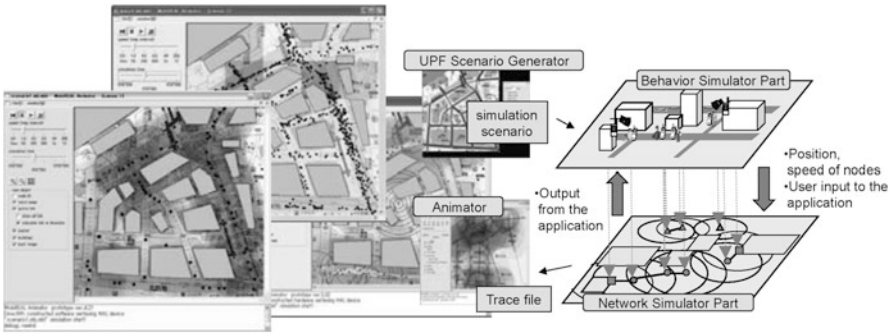


Fig. 16.12 MobiREAL simulator for designing MANET protocols and applications

Mobile ad hoc networks (MANETs) are expected to be very useful and important infrastructure for achieving future ubiquitous society. However, designing MANET protocols and applications is very complicated since it is hardly possible to build large-scale and realistic testbeds in the real world for performance evaluation. Thus there are demands for methodologies that allow us to design, analyze, and validate given applications in simple and inexpensive ways. Thus, we have designed and developed MobiREAL simulator [2] by extending the network simulator GTNetS [15], developed at the Georgia Institute of Technology. The main features of MobiREAL are twofold: First, MobiREAL introduces an original model called the condition probability event (CPE) model to describe the dynamic behavior of pedestrians, such as adjusting walking speeds and directions to avoid collision with neighbors and obstacles, and stopping at a traffic signal. By the CPE model, we can also describe the interaction between pedestrians and networks, e.g., we can describe a scenario that a mobile node makes a detour when it receives traffic jam information through MANETs or cellular networks. MobiREAL is developed by integrating the framework to enable the interaction between mobile nodes and network applications. By incorporating this CPE model into UPF mobility scenarios, the reality of simulations is considerably improved. Secondly, MobiREAL provides a suite of useful tools. With a visualization tool called an animator, the results of simulations can be analyzed easily and intuitively. The animator can visually animate the movement of nodes, network topology, packet propagation, and so on, and can also show statistical information like node density and the packet error rate observed in each sub-region (see Fig. 16.12).

16.3.3.2 Sensor Allocation for Human Mobility Detection

In order to create accurate pedestrian flows for a target area using UPF mobility, we need to install an adequate number of sensors at adequate places, which can count the numbers (or densities) of pedestrians. However, the number of sensors and their deployed locations greatly affects the performance of the human mobility detection.

In our experiments, we have found that the human mobility detection ability has significant performance difference for several scenarios with different sensor settings. Therefore, we have designed and developed a human mobility sensing system simulator called "HumanS" [16]. HumanS is a multi-agent simulator that works with geographic information system (GIS) and the UPF mobility is used for its mobility generation. It models realistic movement of pedestrians and behavior of sensors that capture their mobility.

There are many types of sensors such as LRSs, infra-red-based position sensitive detectors, and image/thermo analyzers to detect objects. We provide a generalized model of those sensors, and such sensors can be placed on any locations in a given map to detect the presence of people in the sensing area. The sensor model is specified by scan range, scan angles, and scan intervals, as well as scan blocking conditions that are directly related with detection errors. Motion detectors such as accelerometers and digital compasses can also be modeled and associated with human agents. It stores and manages the scanning data in a single GIS database by appropriately tagging their time and locations. Therefore, spatial and temporal queries, which are needed in sensor data analysis, can be processed in an efficient way. Figure 16.13 denotes a snapshot of HumanS. It visualizes sensor locations, sensing regions, and human locations and their mobility on GIS map, which helps intuitive awareness, recognition, and analysis of events in the simulation.

HumanS can produce passages and estimate how many percentage of pedestrian flows can be detected by the installed sensors. It can also evaluate where those

Fig. 16.13 Snapshot of HumanS

sensors should be installed. It provides a solution for the optimal sensor allocation problem.

In order to validate the usefulness of HumanS, we have modeled and simulated an underground city of Osaka that has a huge number of visitors (totally 600,000 people/per day). We have focused on a 300 m × 300 m square region in the district and set 22 origin/destination points in the region, which were selected from the entrances of office/commercial buildings and stations as well as stairs from/to the ground level. The generation rates of human agents have been decided based on the real population data of surrounding office buildings and stations to create a realistic flow of walking people. We have used the LRS model with 10 m maximum range, 60° of angle (with 3° angle resolution), and 2 s scan interval. Since the LRS model cannot detect objects behind others, there are always miscounting of people. We have given the following scenarios for observing the impact of sensor dependent properties on the performance of the target sensing system.

(a) Thirty-seven "perfect" sensors on all pathways are deployed where every pathway has a sensor, which can perfectly count the number of people that reside in the sensing region without errors
(b) Thirty-seven LRSs on all pathways are deployed where every pathway has a LRS
(c) Twenty-two LRSs on pathways connected to entrances are deployed where only pathways close to origin/destinations are monitored by LRSs
(d) Additional four LRSs are deployed in Scenario (c) where their locations are manually chosen based on the simulation results

Scenario (a) has achieved the best accuracy of the estimated density (the average error of density is 0.013 persons/m^2 (10 %) while that of Scenario (b) is 0.017 persons/m^2 (17 %). This 7% difference is due to the difference of sensor capability, which is significant in sensing systems. Also, only 11,031 people out of 14,013 can be detected by LRS in Scenario (b) (i.e., the detection ratio = 79 %). In Scenario (c), we can see large errors on pathways without LRSs, but the average error is 0.027 persons/m^2 (28 %). According to the simulation result of Scenario (c), we have tried to improve the accuracy by adding a limited number of LRSs in Scenario (d). As a result of adding only 4 LRSs, we could substantially improve the accuracy. The error has become 0.014 persons/m^2 (14 %), which is very close to the performance in Scenario (b).

16.3.4 Emergency Planning Based on Crowd Sensing

In emergency situations, people might take unusual behavior. There are some research work for creating pedestrian mobility in emergency situations. EXODUS is one of the most famous simulation software for analyzing emergency situations [17], which is developed by Fire Safety Engineering Group (FSEG), Faculty of Architecture, Computing and Humanities, University of Greenwich. It provides

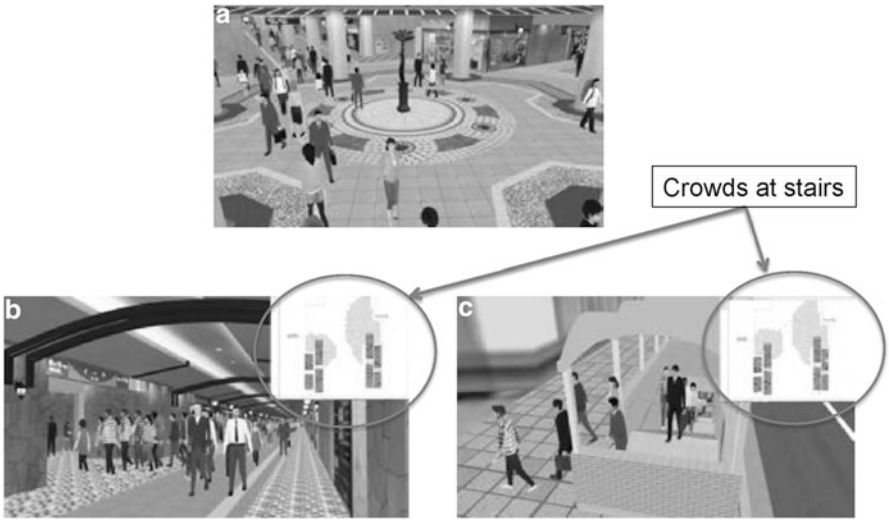


Fig. 16.14 Creation of pedestrian mobility for emergency situations. (**a**) Normal situation. (**b**) Pedestrians rush to stairs. (**c**) Pedestrians evacuate to the ground

several pedestrian mobility in emergency situations such as fire of buildings, disasters, accidents of vehicles, airplanes, and ships.

In Sect. 16.3.3, we have described a method for pedestrian flow estimation using heterogeneous sensors. The method can be used for pedestrian dynamics/circulation analysis in normal situations. We can evaluate time necessary to move from a station to a destination using MobiREAL simulation. Since MobiREAL has the CPE model, and the CPE model can make pedestrian behavior change using if-then rules with probabilities. Digital signage might be able to make pedestrian behavior change by showing alternative routes when a main route is very crowded. MobiREAL can evaluate such situations. It also has a facility to show pedestrians' moving behavior as 2D/3D animation. Here, we consider to use those facilities for emergency planning in an underground mall in front of Osaka Station. Figure 16.14a shows a normal pedestrian mobility of an underground mall in front of Osaka Station. Using the CPE model, we can describe a pedestrian mobility where most of pedestrians rush to neighboring stairs when a disaster occurs. Figure 16.14b shows such a situation. We can also specify the maximum traffic per minute for each stair. Our simulator can simulate a situation where those pedestrians evacuate to the ground from the underground mall based on the maximum traffic of each stair (see Fig. 16.14c). Thus, we can evaluate the total time necessary for those pedestrians to evacuate to the ground from the underground mall when a disaster occurs. Also, by providing several types of information to pedestrians using digital signage and smartphones, we can compare the performance of multiple emergency plans.

We create techniques to reproduce passages, add normal/emergency pedestrian flows, and check the efficiency of evacuation plans on 3D map so that local governments can make efficient evacuation plans (e.g., evacuation plans for fire and flooding at target underground malls). Depending on date/time and real-time situations of disaster victims, we can provide suitable evacuation routes and disaster information to their smartphones based on such simulation results.

16.3.5 Edge Computing Paradigm for Safety Management in Urban Districts

We assume that each building, mall, hotel, and station has its own crowd sensing information using the method shown in Fig. 16.3. By merging those crowd sensing information, we can create crowd sensing information for a given urban district.

For the guide for commuters unable to get home, we need to estimate commuters' mobility in urban areas. They might use public transportation, drive their own cars, or walk. How can we estimate commuters' mobility precisely? How can we obtain such mobility information in real-time? For estimating commuters' mobility, we need huge geospatial sensing data.

As shown in Fig. 16.15, in order to create vehicular mobility in urban areas, we might need (a) road maps and traffic jam information, (b) probe (floating) car data, and (c) mobile phone users' data. In order to create pedestrian mobility in urban areas, we might need (c) mobile phone users' data, (d) route map of trains/buses, and (e) passengers' getting on and off data. By combining those geospatial sensing data, we need to create vehicular and pedestrian mobility in urban areas. Based on those mobility, we can consider several guide plans for commuters unable to get home.

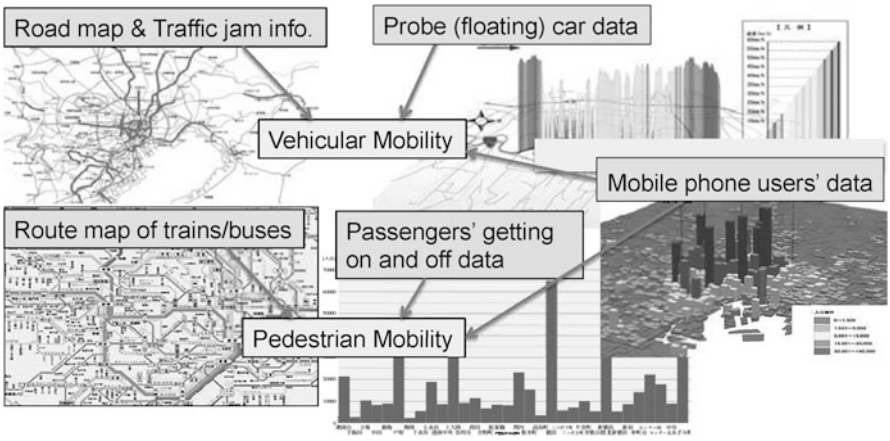


Fig. 16.15 Guide for commuters unable to get home

Thus, for the guide of commuters, we need huge geospatial sensing data. Those sensing data vary day by day, time by time. For the guide, neighboring geospatial data might have strong correlation while distant geospatial data might not have so strong correlation. Thus, it might be suitable for storing such geospatial data in local edge servers in order to provide location dependent services. Central cloud servers can cooperate with edge servers and provide total control mechanisms for metropolitan areas.

16.4 Prediction of Vehicle Speeds in Snowy Urban Roads

In this section, we introduce our recent CPS-based research work for the prediction of vehicle speeds in snowy urban roads [18]. Sapporo City, Japan has a population of 1.9 million, and the average snowfall in winter is 597 cm. Sapporo City is known as the most snowfall city in the cities with more than one million population in the world. Thus, efficient planning of clearing snow is a crucial problem in Sapporo City. As shown in Fig. 16.16, winter roads in snowfall cities have a lot of snow, and their road depths become rather narrow in winter. Also, their road surface conditions are drastically changed depending on snowfall, temperature, traffic amounts, and so on.

We have carried out the Japanese governmental CPS Integrated IT Platform (CPS-IIP) Project [19]. In CPS-IIP Project, we are studying efficient snow removal work in Sapporo City using floating car data and weather data. We also study to estimate travel time in winter (speed drop of vehicles in snowfall).

Here, we classify winter urban road segments into three categories: (1) busy road segments with frequent traffic jam (mainly, road segments in the city center), (2) main road segments with enough traffic amount but infrequent traffic jam (mainly, road segments from residential areas to the city center), and (3) other road segments such as community roads without traffic jam. From our preliminary investigation, we have found that the snowy roads of category (1) are frequently cleaned, and their vehicle speeds are mainly explained by the vehicle speeds of the previous day (or on the same day of a week) although they are also affected by the

Fig. 16.16 Typical snowy road surface conditions

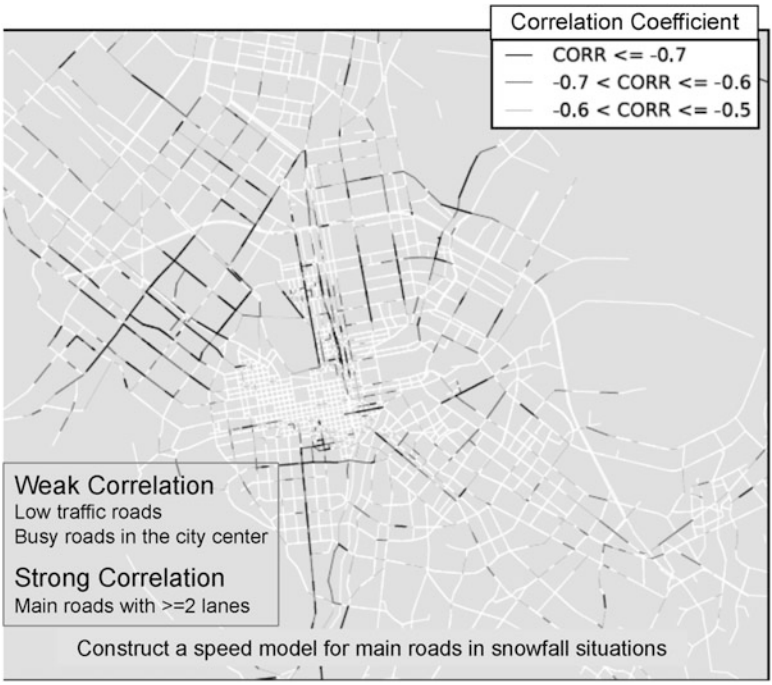


Fig. 16.17 Correlation between snowfall and speed drop

snow amount and temperature. On the other hand, the vehicle speeds on road segments in category (2) have a certain correlation with the temperature, snow depth, new snowfall, and vehicle speed on the previous day. For example, as shown in Fig. 16.17, the main roads in Sapporo City with at least two lanes have strong correlation between snowfall and speed drop while low traffic roads have weak correlation. Note that the road segments in category (3) have weak correlations with those factors, that is, vehicle speeds of community roads depend on locations, road widths, and frequency of snowplowing. Thus, we focus on roads in categories (1) and (2), and aim to predict their vehicle speeds so that the commuters can estimate how much time they need for commuting to work before they leave their home in the morning. We use weather information and vehicular traffic data to predict the deceleration amount of vehicle speed on each road segment.

Here, we explain the basic idea to predict the vehicle speed on snowy roads using weather information and vehicle traffic data. We define "the deceleration amount of vehicle speed" as the difference between the average vehicle speeds on snowy seasons and non-snowy seasons for target road segments. For example, we denote that the deceleration amount of vehicle speed is -15.0 Km/h if the average vehicle speeds are 20.0 and 35.0 Km/h on the road segment in snowy seasons and non-snow seasons, respectively.

We propose a vehicle speed model to predict the deceleration amount of vehicle speed for each road segment using the regression analysis technique. In order to predict the deceleration amount of vehicle speed, weather information and vehicular traffic data are treated as the dominant factors in the regression model equation. We represent the deceleration amount as the linear regression expression of those explanatory variables. Note that in Japan, the snow depth is measured at the weather stations and that the detailed weather information for most cities is obtained at Japan Meteorological Agency [20]. In the below, we explain how to find those explanatory variables and give their weights in order to estimate the deceleration amount of vehicle speeds with high accuracy.

We apply a multiple regression analysis method for finding major factors for estimating speed drop for each road segment, and we have found some major factors (explanatory variables) shown in Fig. 16.18. Then, we construct a linear regression expression of those explanatory variables as follows:

$$\begin{aligned}\text{Speed Drop (Km/h)} = {} & \text{Initial value } p_0 + p_1 \times (\text{Snow depth}) + p_2 \times (\text{New snowfall}) \\ & + p_3 \times (\text{Snow depth one day before}) + p_4 \times (\text{Highest temp. one day before}) \\ & + p_5 \times (\text{Speed in summer}) + p_6 \times (\text{Speed drop one day before})\end{aligned}$$

We use a floating car data for estimating the values of the coefficients $p_0, p_1, \ldots, p_6$. Depending on roads, the values (coefficients) of parameters $p_0, p_1, \ldots, p_6$ vary. Some explanatory variables do not affect the value of the above "Speed Drop." Thus, such explanatory variables are omitted using Akaike's information criterion [21, 22].

Figure 16.19 denotes the estimated speed drop and its real one for a typical road segment in category (2).

We have evaluated the difference between estimated speed drops and real ones for 7289 road segments in Sapporo City. We have collected floating car data and

Fig. 16.18 Multiple regression analysis

Parameters

	coefficient	t-value
Initial value	+8.071	+6.10
Snow depth(cm)	-0.0470	-4.63
New snowfall(cm)	-0.0453	-2.49
Snow depth one day before(cm)	-0.0539	-3.09
Highest temp. one day before(°C)	+0.207	+4.39
Speed in summer(km/h)	-0.202	-6.99
Speed drop one day before(km/h)	+0.632	+19.77

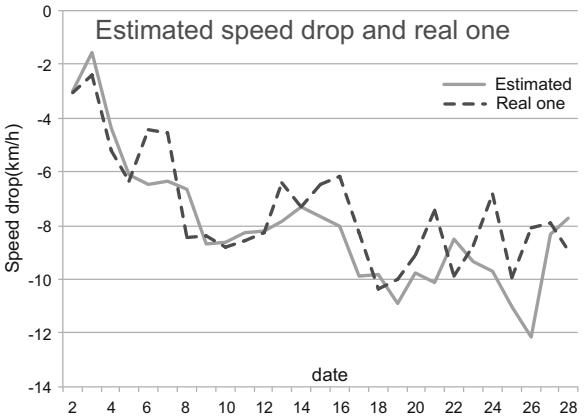


Fig. 16.19 Estimated speed drop and its real one for a typical road segment (Feb. 2013)

weather information for 20 days from Feb. 1, 2013 for those 7289 road segments. Then, we have estimated the values (coefficients) of parameters $p_1, p_2, \ldots, p_6$ for each of those 7289 road segments, and apply those values for estimating vehicle speeds of the corresponding road segments for 8 days from Feb. 21, 2013. The estimation errors for vehicle speeds of 3419 road segments are less than 4.0 Km/h. Our method could not precisely predict vehicle speeds for 2779 road segments because the accurate average speeds cannot be obtained for those road segments. Most of those 2779 road segments belong to the community roads in the category (3). The precision of the model depends on the number of floating cars. If we can obtain more floating car data for target road segments, more precise estimation of vehicle speeds becomes possible. From the graph in Fig. 16.20, if we can collect more than 30 floating car data per day for a target road segment, we can estimate its vehicle speed in winter with less than 4.0 Km/h estimation errors and more than 70 % probability; for details, see [18].

Here, we introduce a method for predicting the deceleration of vehicle speeds in snowy regions and apply the method for Sapporo City. There are several cities near Sapporo City. Those cities also need efficient snow removal plans and prediction of the deceleration of vehicle speeds in bad weather. Those neighboring cities also have similar trends for snowfall and deceleration of vehicle speeds in bad weather. Thus, by cooperating with neighboring cities, the local government of each city can make a more efficient snow removal plan and provide more adequate traffic information to residents. We expect the edge computing paradigm can be used for those purposes.

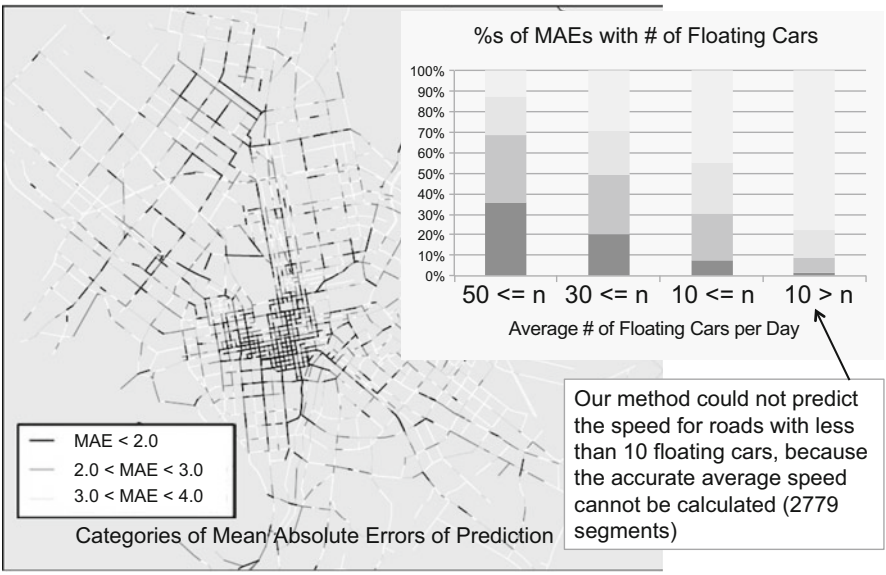


Fig. 16.20 Difference between estimated speed drops and real ones

16.5 Conclusion

In this paper we have introduced the notion of "Edge Computing," and explain its history, features, and research challenge. Then, we introduce our recent two CPS-based research work for safety management in urban districts and prediction of vehicle speeds in snowy urban roads. For constructing affluent and smart social systems, we need to consider scalability and real-time feedback mechanisms. We believe the notion of edge computing and CPS-based research can contribute for those purposes.

Our future work is to apply the methods introduced in Sects. 16.3 and 16.4 for wider urban districts and actually evaluate the usefulness of the edge computing paradigm quantitatively.

Acknowledgements This work is partly supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research(S) (KAKENHI) Grant Number 26220001. The research work introduced in Sects. 16.3 and 16.4 has been jointly carried out with Prof. Hirozumi Yamaguchi, Akihito Hiromori, Akira Uchiyama, Takamasa Higuchi, Takaaki Umedu, and graduate students in our laboratory.

References

1. López PG, Montresor A, Epema DHJ, Datta A, Higashino T, Iamnitchi A, Barcellos MP, Felber P, Rivière E (2015) Edge-centric computing: vision and challenges. ACM SIGCOMM Comput Commun Rev 45(5):37–42
2. Maeda K, Uchiyama A, Umedu T, Yamaguchi H, Yasumoto K, Higashino T (2009) Urban pedestrian mobility for mobile wireless network simulation. Ad Hoc Netw 7(1):153–170

3. Fasolo E, Rossi M, Widmer J, Zorzi M (2007) In-network aggregation techniques for wireless sensor networks: a survey. IEEE Wirel Commun 14(2):70–87
4. Cohen E, Kaplan H (2007) Spatially-decaying aggregation over a network. J Comput Syst Sci 73(3):265–288
5. Higuchi T, Yamaguchi H, Higashino T (2015) Mobile devices as an infrastructure: a survey of opportunistic sensing technology. J Inform Process 23(2):94–104
6. Bista R, Chang JW (2010) Privacy-preserving data aggregation protocols for wireless sensor networks: a survey. Sensors 10:4577–4601
7. Ozdemir S, Xiao Y (2009) Secure data aggregation in wireless sensor networks: a comprehensive overview. Comput Netw 53:2022–2037
8. Shi J, Zhang R, Liu Y, Zhang Y (2010) PriSense: privacy-preserving data aggregation in people-centric urban sensing systems. In: Proceedings of IEEE 29th conference on computer communications (INFOCOM'10), pp 758–766
9. Enzweiler W (2009) Monocular pedestrian detection: Survey and experiments. IEEE Trans Pattern Anal Mach Intell 31(12):2179–2195
10. Li M, Zhang Z, Huang K, Tan T (2008) Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. In: Proceedings of 19th international conference on pattern recognition (ICPR 2008), pp 1–4
11. Chan A, Liang Z, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. In: Proceedings of 2008 I.E. international conference on computer vision and pattern recognition (CVPR 2008), pp 1–7
12. Fujita K, Higuchi T, Hiromori A, Yamaguchi H, Higashino T, Shimojo S (2015) Human crowd detection for physical sensing assisted geo-social multimedia mining. In: Proceedings of 2015 I.E. conference on computer communications workshops (INFOCOM WS), pp 642–647
13. Hokuyo Automatic, Co., Ltd (2016) Scanning range finder, UTM-30LX. https://www.hokuyo-aut.jp/02sensor/07scanner/download/products/utm-30lx/
14. Nishimura T, Higuchi T, Yamaguchi H, Higashino T (2014) Detecting smoothness of pedestrian flows by participatory sensing with mobile phones. In: Proceedings of 2014 ACM international symposium on wearable computers (ISWC'14), pp 15–18
15. Riley GF (2003) The Georgia tech network simulator. In: Proceedings of ACM SIGCOMM workshop on models, methods and tools for reproducible network research (SOSP'03), pp 5–12
16. Kanaya T, Hiromori A, Yamaguchi H, Higashino T (2012) HumanS: a human mobility sensing simulator. In: Proceedings of 5th IFIP international conference on new technologies, mobility and security (NTMS 2012), pp 1–4
17. EXODUS (2016) Fire Safety Engineering Group (FSEG), Faculty of Architecture, Computing & Humanities, University of Greenwich. https://fseg.gre.ac.uk/exodus/
18. Tanimura R, Hiromori A, Yamaguchi H, Higashino T, Umedu T (2015) Prediction of deceleration amount of vehicle speed in snowy urban roads using weather information and traffic data. In: Proceedings of IEEE 18th international conference on intelligent transportation systems (ITSC2015), pp 2268–2273
19. NII (2016) CPS Integrated IT Platform Project https://www.cps.nii.ac.jp
20. Japan Meteorological Agency (2016) http://www.jma.go.jp/jma/indexe.html
21. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
22. Boisbunon A, Canu S, Fourdrinier D, Strawderman W, Wells MT (2014) Akaike's information criterion, CP and estimators of loss for elliptically symmetric distributions. Int Stat Rev 82 (3):422–439

Chapter 17
Challenges of Application of ICT in Cattle Management: Remote Management System for Cattle Grazing in Mountainous Areas of Japan Using a Smartphone

T. Gotoh, M. Maeda, O. Hirano, M. Nishiki, T. Fujita, T. Shibata, Y. Takayama, K. Yokoo, T. Nishidoi, H. Urabe, T. Ikenouchi, T. Ninomiya, M. Yoshida, J. Sugiyama, T. Sasaki, S. Sawane, and A. Muranishi

Abstract We created an information and communication technology system to monitor the feeding of grazing cattle, comprising the following elements: (1) actuators, including an audio player, feeder, and stanchion (an equipment for locking cattle neck during feeding); (2) wireless nodes with Wi-Fi connectivity to control the actuators; (3) Internet protocol (IP)-addressed cameras with Wi-Fi connectivity; (4) a wireless access point enabling the wireless nodes and IP cameras to connect to a wired network using Wi-Fi; and (5) a Web server that communicates with the wireless nodes and delivers a control graphical user interface for use by farmers. Farmers can call cattle towards an IP camera and feed them by using a smartphone without being physically present on the farm. Next, we constructed a received signal strength-based location system for grazing cattle. This included monitoring the location of cattle on pastures in mountainous areas, where issues to be considered included low energy consumption for battery operation and radio propagation degradation due to the undulating topography of the pastures. To evaluate the

T. Gotoh (✉)
Kuju Agricultural Research Center, Kyushu University, Kuju 878-0201, Japan
e-mail: gotoh@farm.kyushu-u.ac.jp

M. Maeda • O. Hirano
Project Support Department, Intellectual Property Management Center, Kyushu University, Fukuoka 812-8581, Japan

M. Nishiki • T. Fujita • T. Shibata • Y. Takayama
Research and Development Center, NTT West Corporation, 1-2-31, Osaka 530-0057, Japan

K. Yokoo • T. Nishidoi • T. Sasaki • S. Sawane • A. Muranishi
Innovation Promotion Office, Fujitsu Advanced Engineering Ltd., Tokyo, Japan

H. Urabe • T. Ikenouchi
Mobile Techno Corporation, Kawasaki, Japan

T. Ninomiya • M. Yoshida • J. Sugiyama
Advanced Wireless Technologies Lab., Fujitsu Laboratories Ltd., Kawasaki, Japan

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_17

proposed algorithm, an experimental study was conducted in a pasture with an area of 2 km^2. We also provided a means to view the locality on a smartphone by creating an application that connects to this system. In this study, we created two systems controllable via a smartphone, one system to call cattle and the other to monitor their locations. These systems could be very useful to farmers in controlling and monitoring cattle grazing.

Keywords Grazing cattle • ICT control • Smartphone • Remote feeding • Remote catching • Location monitoring of cattle

17.1 Background and Introduction

How can we apply information and communication technology (ICT) in agriculture in Japan? First of all, we have to know the actual situation of agriculture in Japan, especially the target area on which we want ICT to apply. We would like to introduce the current situation of agriculture, especially beef production situation of our target field and discuss how we apply ICT to improve the production system and apply for it.

17.1.1 The Current Situation of Agriculture in Japan

Japan exports industrial goods like cars and electrical products to overseas countries in exchange for foreign currency. In addition, domestic agriculture has traditionally been protected by refraining from importing food. However, at the beginning of 1990 as a result of the Uruguay Round, Japan agreed to import staple foods like rice, beef, and oranges [1]. Now Japan has low food self-sufficiency (40 %) [2]. The availability of cheap imported food has damaged Japanese agriculture. As much as 73 % of the total land in Japan is mountainous. Rice as a staple food is produced on flat areas.

17.1.2 Current Situation of Beef Production in Japan

Regarding beef production, Japan has a special breed, the "Wagyu (Japanese black)," which produces excellent marbled beef with more than 50 % intramuscular fat. Because the Japanese beef market prefers marbled beef with a large percentage of intramuscular fat, farmers have set up feed lots on flat areas without attached grass lands, and have managed intensive feeding using a vast amount of imported grain feed. To produce Wagyu beef, farmers require between 4000 and 5000 kg grain feed per head over 30 months, 90 % of which is imported. Moreover, the

Fig. 17.1 Cattle grazing on an abandoned agricultural land: (**a**) image when grazing started, and (**b**) image after 3 months

business of beef production is fraught owing to rising prices of imported feed and the distorted structure of the beef production business, which indicates the percentage of cost of current beef production to the price in Japan mostly become greater than 90 % (data not shown, authors calculated value in the university farm). Government and local administrations are protecting farmers by providing large subsidies. Nevertheless, heavy dependence on imported feed places Japan in a precarious situation. In addition, livestock-related epidemics, such as bovine spongiform encephalopathy (BSE) and foot and mouth disease, have caused serious damage to the management of beef production in Japan. The beef production system dependent on imported grain feed is destined to amplify anxiety over food safety in the future, as a result of possible outbreaks not only of BSE, but also of unknown infectious diseases, since contaminated food may be imported into Japan.

Japan, as one of the developed countries, should consider the food balance in the world and shift to safe high quality beef production using domestic grass resources, by shaking itself free from a system dependent on imported feed. To achieve this laudable idea, we need to focus on the mountainous and foothill areas, as well as agricultural lands in Japan that have been abandoned (Fig. 17.1).

Basically, cattle have an important ecological niche that capitalizes on the symbiotic relationship between fiber fermenting ruminal microbes and mammalian demand for usable nutrients as a ruminant. Beef cattle produce “meat” from grass resources as a source of protein after microorganisms in the rumen break down plant fibers that humans cannot digest. We would like Japanese cattle to supply beef for the Japanese population by utilizing domestic grass resources on mountain slopes, foothills, and abandoned agricultural lands. In fact, there are large tracts of land that remain unused, since 50 % of the Japanese population is concentrated on only 14 % of the total land. Typically, most of the Japanese population live in big cities such as Tokyo and Osaka, while in the mountainous and foothill areas the number of depopulated villages with more than 50 % of their inhabitants over 65 years of age is increasing rapidly. To conserve such areas, it is most important for Japan to retain the natural environment, as well as endemic plants and animals.

17.1.3 *Our Aim to the Future: A New Management System of Cattle by ICT*

This is also important in countries that aim for harmony between people and nature, material circulation, and higher food self-sufficiency. At present, the government is recommending using abandoned agricultural lands as grazing for cattle, but this system is still very primitive. To replace the current beef production system, we envisaged constructing a new beef production system for use in mountainous and foothill areas using ICT. This system would not only encourage food production, but also create a new ethical industry for future generations. Sixty-six percent of the national land is forest [3] in Japan. Industries related to forestry also seem to be under pressure, and there are not many successors. This situation is untenable. As previously mentioned, cattle are relatively large animals that can convert plant resources into protein, i.e., beef and milk. Therefore, relying on cattle grazing and ICT technologies, we set out to create a novel mountainous silvopasture system making use of depopulated remote areas (Fig. 17.2).

In the case of beef cattle in the mountain, farmers have to check their health, heat cycles, calving, safety, and so on. Mostly their pastures were located far away from their house because they live in the foot of mountains. Moreover, their checks to cattle have hard work and have to spend transportation fee daily. Normally when present in the pasture, farmers call their cattle to feed using behavior based on conditioned reflex. Previously we constructed calling and locking system by using

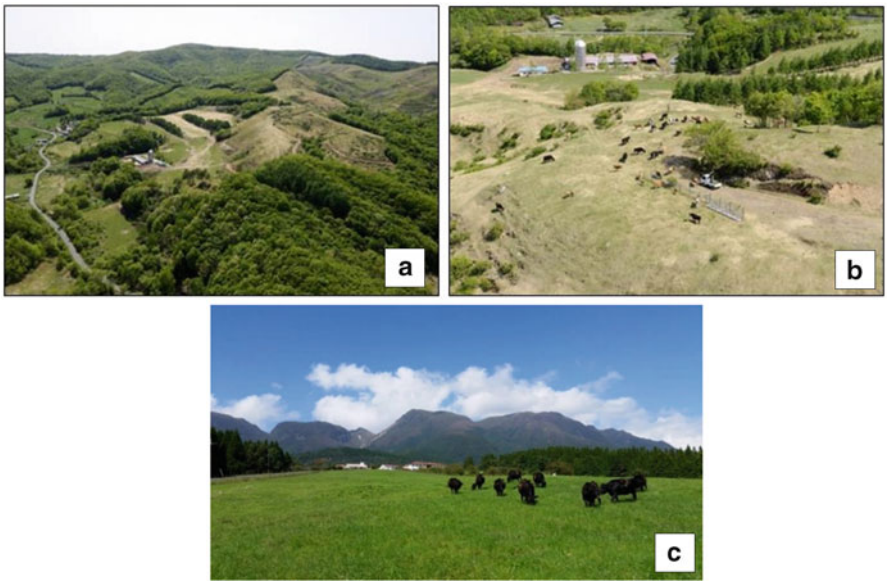


Fig. 17.2 Images of a farm in a mountainous area (Nakahora farm (**a & b**) pictures by Okada, and Kuju Agricultural Research Center, Kyushu University (**c**), by Gotoh)

PC [4]. At the time ICT technologies like smartphone have not been yet developed. In this study, we constructed an observation, calling, and herding system for cattle grazing in a pasture using a remote smartphone, obviating the need for the farmers to be present in the pasture. We set up a web-camera, speakers, an automatic feeder, and stanchions using a rocked system in the pasture. Using a smartphone, we attempted to call and herd the grazing cattle, while monitoring them and the pasture by means of the web-camera.

Additionally, we tried to monitor the location of grazing cattle via a smartphone. Given sufficient battery power, the global positioning system (GPS) is the most useful system for monitoring the location of cattle. Turner et al. [5] or Handcock et al. [6] reported to monitor behavior and change of pasture vegetation by using GPS and GIS. However, a GPS module for grazing cattle would require too much electricity for long continuous operation with small battery which can be mounted on cattle to transmit signals to the central server, and therefore, if farmers were to use it, they would need to exchange batteries daily for each member of the herd. This would require great effort and is not realistic or convenient for farmers. We implemented a system requiring low battery power to check the location of grazing cattle on a map of the pasture via a smartphone. There are few reports to control grazing cattle by ICT remote methods. This study resulted in the creation of two systems: one to call the cattle and the other to monitor their locality, both of which can be controlled by a smartphone.

17.2 Systems of ICT Remote Management of Cattle

17.2.1 Construction of System to Lead Cattle to Feed Using a Smartphone

We created an ICT system to monitor the feeding of grazing cattle, comprising the following elements (Fig. 17.3).

17.2.1.1 Actuator Included a Sound System, Automatic Feeder, and Motorized Stanchion

These actuators have a digital input interface, which is switched by a wireless node. When the sound system is switched on, it plays back a pre-recorded voice in the pasture and calls cattle to assemble in front of an IP camera (Fig. 17.4). When the auto feeder is switched on, it feeds cattle in front of the motorized stanchion. Then, when the motorized stanchion is switched on, it locks the cattle in place.

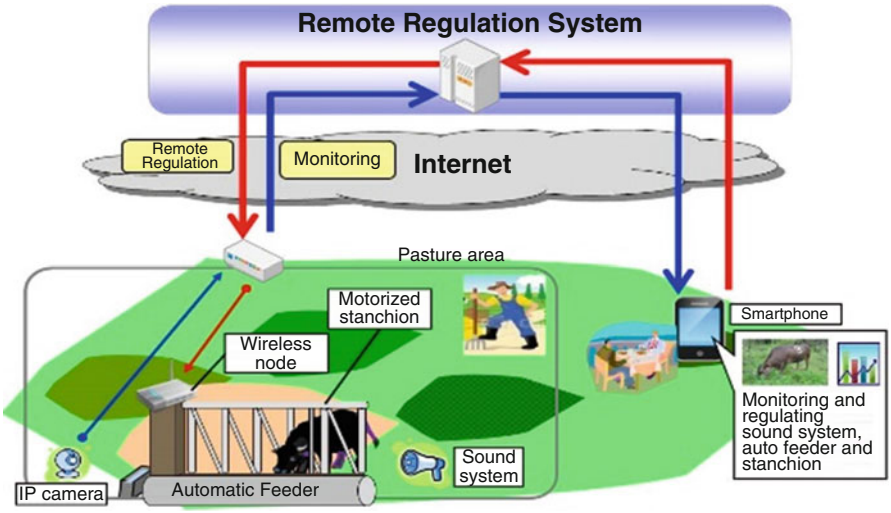


Fig. 17.3 Overview of ICT control system via a smartphone for cattle in remote areas

17.2.1.2 Wireless Node with Wi-Fi Connectivity

A wireless node is a small computer with Wi-Fi connectivity, temperature and humidity sensors, and a digital input/output interface. We installed it near the actuators, and connected it with each actuator by some cables. It uses an IPv6 address on the Linux operating system, and communicates with a Web server over HTTP protocol. It switches the digital output interface on and off, on request from the server, and then controls certain actuators.

17.2.1.3 IP Camera (Panasonic Co., Ltd. SW-174W)

The IP camera provides Wi-Fi connectivity, and sends and retrieves images via the Wi-Fi network. We installed it in front of the stanchion in order to monitor the feeding of grazing cattle.

17.2.1.4 Wireless Access Point (Gonet MBW3100)

A wireless access point is set up on top of a pillar. It allows the wireless nodes and IP cameras to connect to a wired network using Wi-Fi. We constructed the environment for a wireless network using a wireless access point with adaptive beamforming technology, which detects the packet direction-of-arrival and forms a single, narrow beam, directed according to the direction-of-arrival by combining the signals of different antennas. The technology realized a coverage of <1 km.

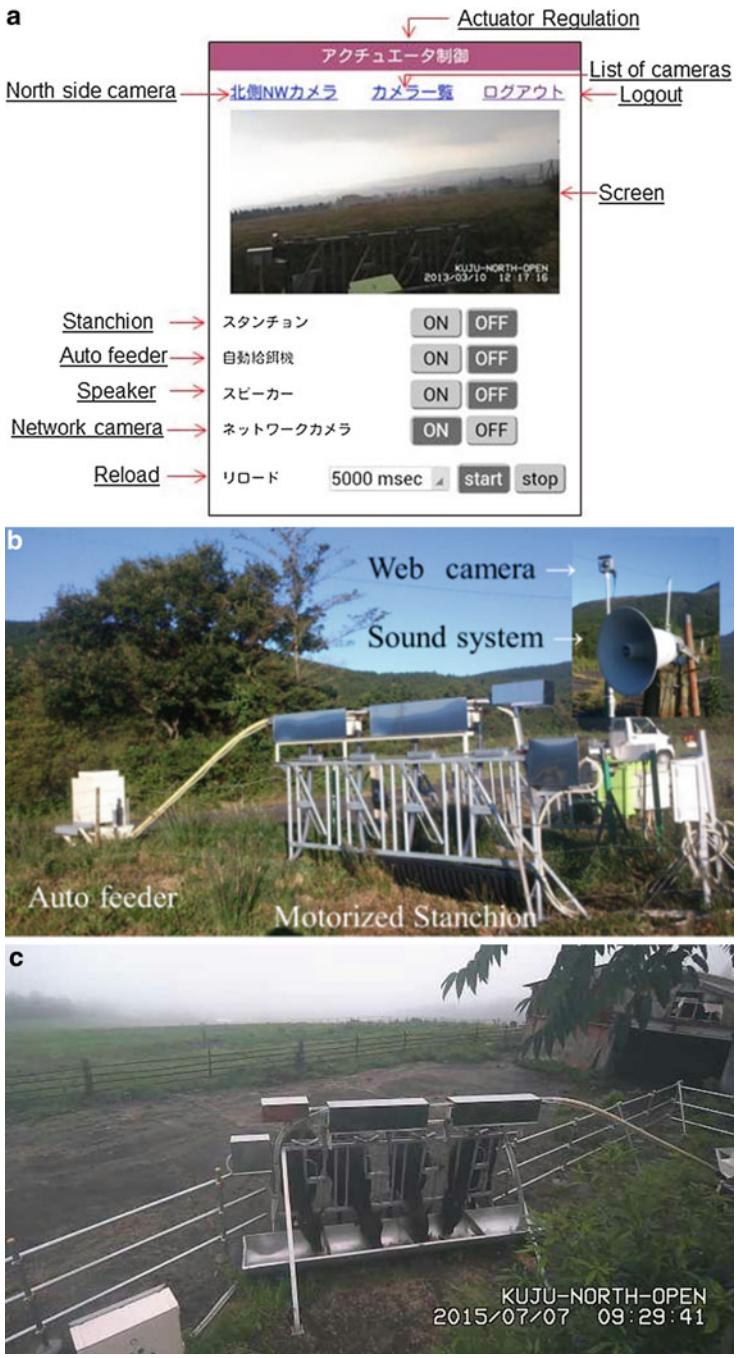


Fig. 17.4 Actuator control screen on a smartphone and the system. (**a**) Actuator control screen on a smartphone. (**b**). Auto stanchion system. (**c**). Picture of grazing cattle caught by auto stanchion

17.2.1.5 Web Server

A Web server is installed in the farmer's office to communicate with the wireless nodes and store data such as temperature and humidity, transmitted by the wireless nodes to the Web server. It also provides a control graphical user interface (GUI) for use by the farmer.

17.2.1.6 Smartphone

Farmers access the website via their smartphones to check the status in the field using the IP camera, and to control the actuators.

The actuator control screen for the smartphone is shown in Fig. 17.4. The image from the IP camera is displayed on the upper part of the screen, allowing farmers to monitor the field and their cattle via the IP camera. The ON/OFF buttons are on the lower part of the screen enabling the farmers to control the sound system, auto feeder, and motorized stanchion.

17.2.2 Practical Effects of System to Call Cattle to Feed

We tested the two created systems in this study on the field by using grazing cattle to practically evaluate the effects when farmers actually use them. We implemented this system at the Kuju Agricultural Research Center and evaluated the usefulness thereof. The coverage of Wi-Fi in the field is shown in Fig. 17.5. We measured the signal strength at various locations using a Wi-Fi Analyzer, an Android tool that tests signal strength for wireless devices.

Wireless nodes and an IP camera were located approximately 500 m from the wireless access point. At this location, the wireless signal strength was greater than −80 dBm because of good prospects. This confirmed that the Wi-Fi connection was stable. On the other side, however, at locations obstructed by trees, the Wi-Fi connection was not stable.

We tested this system by using grazing cattle ($n = 4$) and the actual pasture (5 ha). Grazing cattle were called by the sound controlled by a smartphone. Next, they inserted their head to stanchion by the feed from the auto feeder controlled by a smartphone. Finally they were locked their head by stanchion's stoppers controlled by a smartphone. These events could be observed by the screen through web-camera on the smartphone. If farmers find out cattle which be in the heat (estrus) or sick animals, beforehand they can go to the cattle with preparation of artificial insemination and medical treatment, respectively. Mating and medical treatment are very important things for farmers to keep economics and health of animals. This system would be able to save time and develop economic situation.

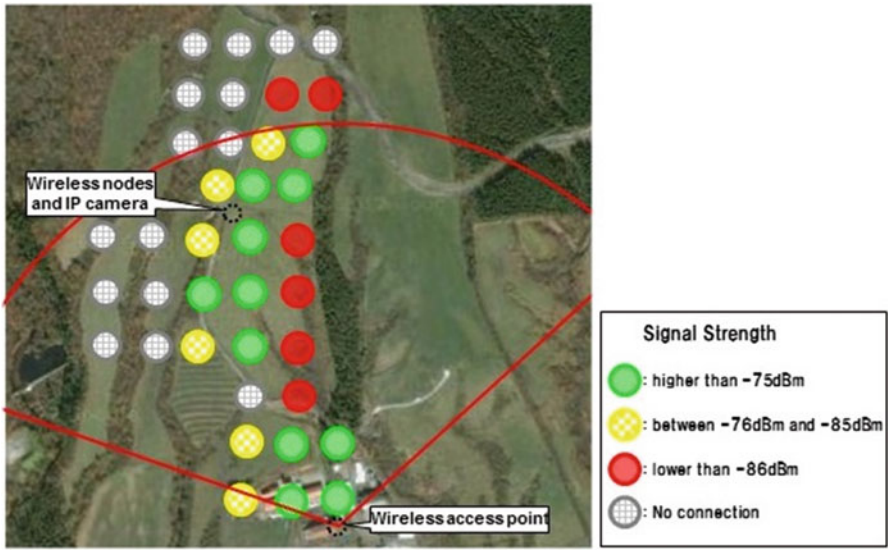


Fig. 17.5 Wireless signal strength at Kuju Agricultural Research Center

In this study, we used Wi-Fi as the field network. Although Wi-Fi is optimized for fast response, low latency, and high data rates, it has power consumption issues. Moreover, Wi-Fi in the 2.4 GHz frequency block cannot cover a large area such as a farm owing to 2.4 GHz band propagation characteristics. To solve these problems, we set about applying the wireless sensor network in a 920 MHz frequency block as the field network, including various standards such as IEEE802.15.4g, IEEE802.15.4e, 6LoWPAN, RPL, and CoAP. These standards are aimed at low power and lossy networks (LLN), which are composed of many embedded devices with limited power, memory, and processing resources interconnected by a variety of links.

17.3 System to Monitor the Location of Cattle in Pastures

17.3.1 Construction of System to Monitor the Location of Cattle in Pastures

17.3.1.1 Localization System for Cattle in Pastures

The prototype system to monitor the location of cattle in pastures was constructed [7]. Figure 17.6 shows the topology of a network for the system. A sensor node (SN) sends packets containing vital data such as body temperature obtained by sensors directly to the base station (BS), and listens for an acknowledgement if the

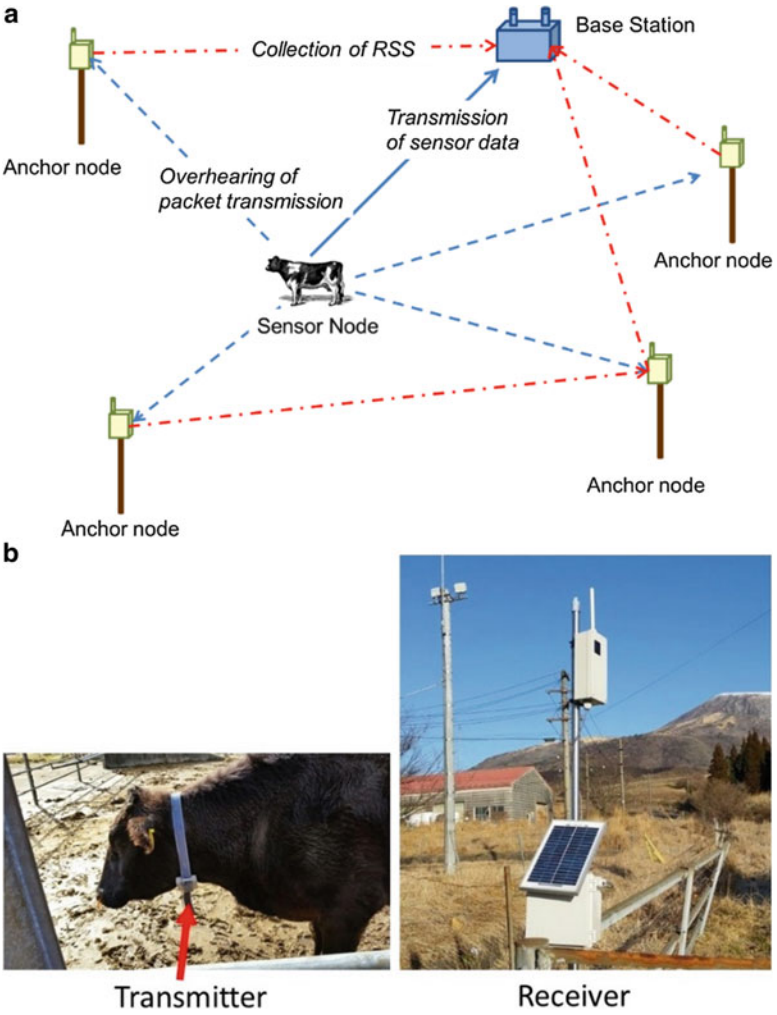


Fig. 17.6 Network topology and devices of cattle localization system [7]. (**a**) Network topology. (**b**) Transmitter and receiver of this system

transmission is successful. At the same time, anchor nodes (AN) are deployed in the pasture and upon receiving the packets, store both the received signal strength (RSS) and the data. If the BS has not successfully received the data, the AN may also transmit the stored packet, which means that it works as a relay. In addition, the ANs subsequently send stored RSS data periodically to the BS at different times or on different channels used by the SN. Thus, the BS receives both the vital information for the cattle and the RSS between the SN and AN. The RSS is used to calculate the location of the node, as described in detail in the following section.

17.3.1.2 Path-Loss Model for RSS-Based Trilateration

We applied RSS-based trilateration method as localization algorithm using PF to take topographical feature into account. We assumed that the relation between the distance and RSS P (dB) follows the path-loss model, where P_{tx}(dBm), G_{ant}(dB), d_0(m), λ(m), d(m), λ(m), α, and X_S(dB) are the transmit power, total antenna gain, reference distance, wave length of signal, distance between a transmitter and a receiver, path-loss exponent, and shadowing factor, respectively. We assume that the shadowing factor X_S follows a log-normal distribution with standard deviation σ_S(dB) and mean L_S(dB).

$$P = P_{tx} + G_{ant} - 20\,log\left(\frac{4\pi d_0}{\lambda}\right) - 10\alpha\,log\left(\frac{d}{d_0}\right) - X_S, \tag{17.1}$$

17.3.1.3 Considering Topographical Feature Using Particle Filter

A flowchart for the applied particle filter is given in Fig. 17.7, where t and m denote time and index of particles and time, respectively. For each step, the following assumptions are made:

(a) *Initialization step:* In the initialization step, each particle is randomly placed in the vicinity of the initial position x_0 if available, or distributed over the entire area otherwise. The position of **x** consists of the x-, y-, and z-coordinates, while elevation z is determined by x, y, and a fixed antenna height using a digital elevation map (DEM) [9].

(b) *State transition step:* The state transition probability $p\left(x_t \middle| x_{t-1}^{[m]}\right)$ is defined by (17.2), where the probability of displacement is equally distributed in radius $r_t = v_M \Delta t$, where Δt and v_M are the time difference from the previous update and the maximum speed of cattle, respectively. If the time difference exceeds a pre-determined value, all particles are randomly replaced over the area as the initialization step.

$$p\left(\boldsymbol{x}_t \middle| \boldsymbol{x}_{t-1}^{[m]}\right) = \begin{cases} \dfrac{1}{r_t} & if \ \left|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right| < r_t) \\ 0 & otherwise \end{cases}, \tag{17.2}$$

(c) *Measurement step:* The measurement probability $p\left(\mathbf{P}_t \middle| \mathbf{x}_t^{[m]}\right)$ for m-th particle is expressed as

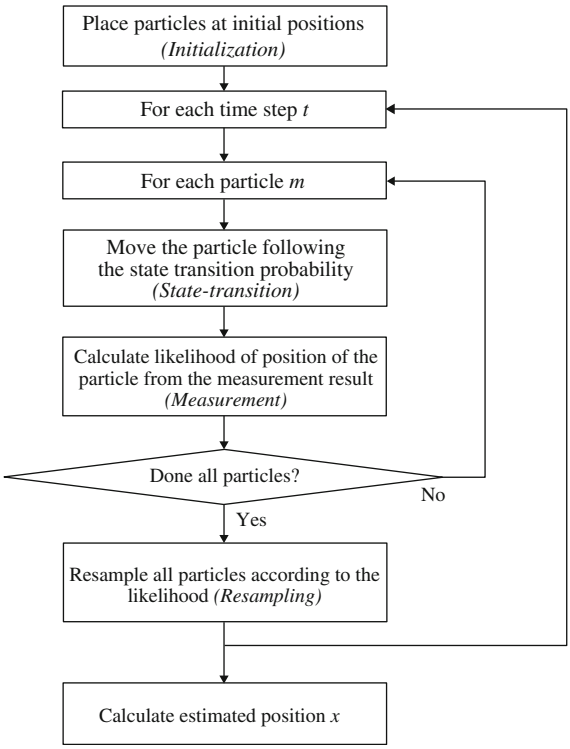


Fig. 17.7 Flowchart of particle filter algorithm for localization [8]

$$\mathrm{p}\left(\mathbf{p}_\mathrm{t}\middle|\mathrm{x}_\mathrm{t}^{[\mathrm{m}]}\right)=\mathrm{p}\left(\mathrm{P}_\mathrm{t}^{[1]},\mathrm{P}_\mathrm{t}^{[2]},\ \ldots,\mathrm{P}_\mathrm{t}^{[\mathrm{k}]},\ldots,\mathrm{P}_\mathrm{t}^{[\mathrm{K}]}\middle|\mathbf{x}_\mathrm{t}^{[\mathrm{m}]},\ \mathrm{v}\right)=\prod_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{p}\left(\mathrm{P}_\mathrm{t}^{[\mathrm{k}]}\middle|\mathbf{x}_\mathrm{t}^{[\mathrm{m}]},\mathrm{v}\right),\quad(17.3)$$

$$p\left(z_t^{[k]}\middle|x_t^{[m]}\right)=p\left(P_t^{[k]}\middle|x_t^{[m]},\nu\right)=\frac{1}{\sqrt{2\pi}\sigma_S^{[m][k],\nu}}exp\left\{-\left(\frac{P_t^{[k]}-\overline{P}_t^{[m][k],\nu}}{2\sigma_S^{[m][k],\nu}}\right)^2\right\},\quad(17.4)$$

where $\mathrm{P}_\mathrm{t}^{[\mathrm{k}]}$, K, and ν denote the measured RSS, number of anchor nodes, and the visibility defined below, respectively. If a straight line between a particle and AN does not intersect the elevation profile calculated from the DEM, visibility ν is defined as being line-of-sight (LOS), otherwise, it is assumed to be non-line-of-sight (NLOS). Additionally, $\overline{\mathrm{P}}_\mathrm{t}^{[\mathrm{m}][\mathrm{k}],\nu}$ and $\sigma_\mathrm{S}^{[\mathrm{m}][\mathrm{k}],\nu}$ are the mean and standard deviation of the estimated RSS at the k-th anchor node for the m-th particle as shown in (17.1), where the channel dependent parameters, i.e., σ_S, $\mathrm{L_S}$, and α are changed according to the visibility ν to consider the topographical features.

(d) *Resampling step:* All particles are resampled with probability proportional to the measurement probability obtained in the measurement step.

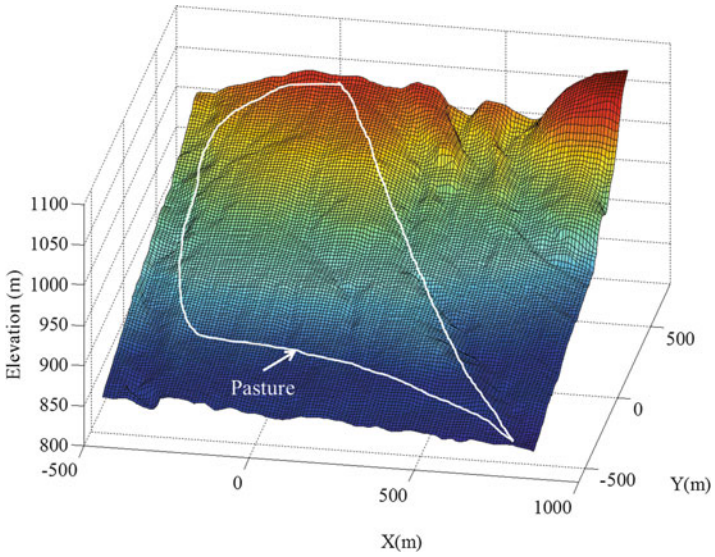


Fig. 17.8 Three-dimensional view of the pasture of Kuju Agricultural Research Center, Kyushu University

(e) *Localization:* For each update, a centroid of the position of particles is considered to be the localization result.

17.3.2 Practical Effects of System to Monitor Location of Cattle

We also evaluated the proposed localization algorithm at the Kuju Agricultural Research Center, which has a pasture with an area of 2 km^2 and a 180 m difference between the highest and lowest points. The area resembled an average pasture in Japan [10]. Figure 17.8 shows a three-dimensional view of the area.

17.3.2.1 Measurement Equipment

Table 17.1 gives the specifications of the transceiver module used in the evaluation. The 429 MHz frequency band is one of the unlicensed bands in Japan, and despite its small transmission power, can cover a large area up to 1 km.

Table 17.1 Specification of the transceiver for the experiment [7]

Transmit power	10 mW
Center frequency	429.5 MHz
Antenna gain	0 dBi
Modulation scheme	FSK
Bandwidth	12.5 kHz
Data rate	4.8 kbps
Adopted standard	ARIB STD T-67 [11]
Sensitivity	120 dBm

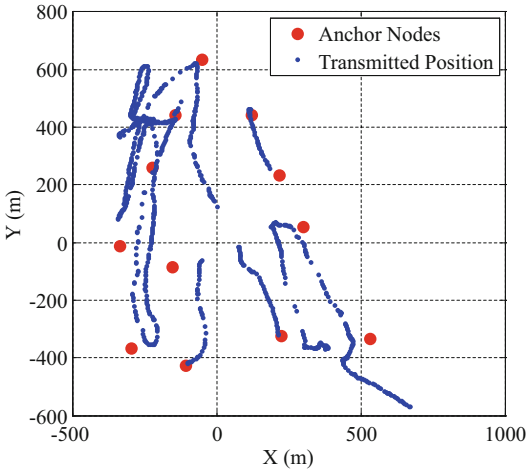


Fig. 17.9 Arrangement of anchor nodes for channel parameter estimation and experiment [7]

17.3.2.2 Channel Measurement

First, we performed channel measurements to derive σ_S, L_S, and α under both LOS and NLOS conditions. All ANs were arranged as shown in Fig. 17.9, with an antenna height of 2 m above the ground. A transmitter moved around the pasture at a height of approximately 1.5 m, as shown by the dots in the figure and periodically transmitted packets containing time and position information, obtained from the GPS. These were received by the ANs and stored together with the RSS. The channel parameters extracted from the data are summarized in Table 17.2, where "ALL" denotes the use of all data without distinction.

17.3.2.3 Localization Experiment

Then we performed a localization experiment with same arrangement of anchor nodes and transmitter emulating sensor nodes as shown in Fig. 17.9. All data were collected in the same way as for the channel measurement, and were processed as described in Sect. 2.2 with the obtained channel parameters in Table 17.2. Experimental conditions are summarized in Table 17.3.

Table 17.2 Extracted channel parameters for LOS/NLOS [7]

Parameter	LOS	NLOS	ALL
Standard deviation of shadowing σ_S	7.9 dB	6.9 dB	7.9 dB
Mean of shadowing L_S	−0.5 dB	13.4 dB	0.3 dB
Path-loss exponent α	2.7	2.4	2.8

Table 17.3 Parameters for the localization experiments [7]

Number of particles for PF	1000
Transmit power P_{tx}	10 dBm
Antenna gain G_{ant}	0 dB
Wavelength λ	0.70 m (429 MHz)
Reference distance d_0	1.0 m
Antenna height (anchor/sensor)	2.0 m/1.5 m
Maximum speed of animal v_{max}	36 km/h

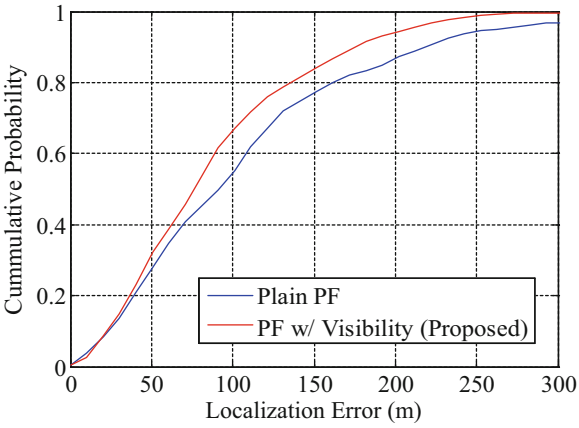


Fig. 17.10 Cumulative probability of localization error [7]

Figure 17.10 shows the cumulative probability function (CP) of the localization errors obtained for the proposed PF with the visibility algorithms and plain PF method for comparison. For the plain PF, we use same channel parameters denoted as "ALL" in Table 17.2 regardless of the visibility. The localization error of proposed PF with visibility and plain PF are 76.3 m and 91.5 m at 50 % of CP, respectively. Therefore, by taking topographical feature into account with the introduction of visibility, localization accuracy is improved by 16 %. Additionally, the proposed method performs 20 % better at 90 % of CP. These results are acceptable for cattle pasturing, which requires an accuracy of 100 m with very low energy requirements. Finally, we created a prototype incorporating the location information and map using a smartphone (Fig. 17.11) and tested it with actual grazing cattle in the pasture.

We successfully located them. The performance could be further improved by introducing more information, such as additional sensors or using more channel

Fig. 17.11 Prototype system for monitoring location of grazing cattle on a smartphone

information, which should be investigated in future. On the left side of smartphone, the map and trace of grazing cattle you chose are visible. On the right side of it, the cattle picture you choose and the body temperature of the cattle are visualized (this system is not explained in this book).

17.4 Conclusion

In this study, we created two systems controlled by a smartphone: one to call cattle and the other to monitor the location of grazing cattle. First, we created an ICT system to monitor the feeding of grazing cattle, comprising the following elements: (1) actuators included audio player, feeder, and stanchion; (2) wireless nodes with Wi-Fi connectivity to control the actuators; (3) IP-addressed cameras with Wi-Fi connectivity; (4) a wireless access point to allow wireless nodes and IP cameras to connect to a wired network using Wi-Fi; and (5) a Web server to communicate with the wireless nodes and provide a control GUI for use by farmers. Farmers can call cattle towards an IP camera and feed them using a smartphone without being physically present at the farm. Support for IP version 6, given that many devices would be connected to the network; and a control GUI that allows a user to control the actuators and monitor the field via IP cameras.

Secondly, we implemented the localization system for cattle in pastures. The system consists of sensor nodes, anchor nodes, a base station, and a server, and provides localization. To realize low energy localization in undulating pastures, we

proposed RSS-based trilateration using a PF algorithm with topographical features, referred to as "visibility." From experiments conducted in the pasture, this method was found to outperform the conventional plain PF approaches by 16 % at 50 % CP with a localization error of 76.3 m. Although this satisfies current system requirements, further investigation is needed for further improvement. Anyway, we created a prototype of the system to monitor the location of grazing cattle via a smartphone.

These systems could be very useful for farmers, assisting in the control of cattle grazing in inconveniently located fields. In this study, by using ICT technology, we constructed smart ways to control grazing cattle. To analyze visual data from web-camera by artificial intelligence and to realize what cattle do or predict their behavior would be more useful for farmers in the future. Moreover, at the next step, we are trying the monitoring of the health of cattle such as body temperature by implanted temperature sensor now. It would be our important next work.

Acknowledgment The authors thank Mr. T. Etoh, Mr. Y. Shiotsuka, Mr. R. Fujino, and Dr. H. Takahashi for their excellent advice and technical assistance. This work was supported by the NTT West Corporation, Fujitsu Limited, Strategic information and communications R&D Promotion Program (SCOPE) of Ministry of internal affairs and Communications (112310005), KAKENHI Grant Number 26310312 of Japan Society for the Promotion of Science (JSPS).

References

1. World Trade Organization (WTO) (2013) The Uruguay Round. http://www.wto.org/english/thewto_e/whatis_e/tif_e/fact5_e.htm
2. Ministry of Agriculture, Forestry and Fisheries of Japan (MAFF) (2010) Abstract of statistics on agriculture, forestry and fisheries census. Update No. 730, Available at: http://www.maff.go.jp/e/maffud/2010/730.html
3. Ministry of Agriculture, Forestry and Fisheries of Japan (MAFF) (2013) Abstract of statistics on agriculture, forestry and fisheries, forest agency. Update No. 730, Available at: http://www.rinya.maff.go.jp/j/kouhou/toukei/index.html
4. Gotoh T, Etoh T, Shiotsuka Y, Maeda M, Terauchi H, Saito N, Takeishi S, Utsunomiya S, Maeda Y, Honda T, Okubo K, Kinno Y, Mine S (2007) Internet control of cattle grazing in mountain and foothill areas of Japan. In: Proceeding of 8th international conference of construction, technology and environment in farm animal husbandry, 2007, Bonn, Germany, pp 407–411
5. Turner LW, Udal MC, Larson BT, Shearer SA (2000) Monitoring behavior and pasture use with GPS and GIS. Can J Anim Sci 80(3):405–413
6. Handcock RN, Swain DL, Bishop-Hurley GJ, Pation KP, WArk T, Valencia P, Corke P, O'Neil CJ (2009) Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collards and satellite remote sensing. Sensors 9:3586–3603
7. Yokoo K, Nishidoi T, Urabe H, Ikenouchi T, Ninomiya T, Yoshida M, Sugiyama J (2013) RSS-based localization considering topographical feature for pasturing. In: 2013 10th workshop on positioning navigation and communication (WPNC), 20–21 March 2013, pp 1, 5
8. Thrun S, Burgard W, Fox D (2006) Probabilistic robotics. MIT Press, Cambridge
9. GSI (September 2012) GSI home. Geospatial Information Authority of Japan, http://www.gsi.go.jp/

10. Tsujii H (2005) The present scenario of pasture in Japan (in Japanese). AFC Bulletin, Faculty of Agriculture, Shinshu University, vol 3:1–5
11. ARIB (2000) Telemeter, telecontrol and data transmission radio equipment for specified low-power radio station (ARIB STD-T67). Association of Radio Industries and Businesses (ARIB), Tokyo

Chapter 18
Health Sensor Data Analysis for a Hospital and Developing Countries

Yasunobu Nohara, Sozo Inoue, and Naoki Nakashima

Abstract We present two types of sensor data analysis for medical and healthcare. One sensor dataset is collected in a hospital for medical purposes. We gathered accelerometer data and RFID data of real nursing in the hospital. We provide the real nursing dataset for mobile activity recognition which could be used for supervised machine learning, and also the big data combined with the patients' medical records and sensors tried for 2 years. The other sensor dataset is collected in a developing country. We developed an eHealth system that comprises a set of sensor devices in an attache case. The first checkup was provided to 16,741 subjects. After 1 year, 2361 subjects participated in the second checkup, and the blood pressure of these subjects was significantly decreased ($P < 0.001$). Based on these results we proposed a cost-effective method using a predictor, to ensure sustainability of the program in developing countries

Keywords Mobile activity recognition • Accelerometers • Nursing activity • Big-data mining • Machine learning • Public health informatics • Preventive medicine • Sensor package • Developing countries • Body area network

18.1 Introduction

Evidence-based medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [1]. In order to decide best practice in medicine, there is a strong need to collect large amount of data and sensor is the one of the best method. There is a long history of using various sensors in medicine [2]. The modern medicine is not possible without sensor such as thermometers, blood pressure meter, and glucose meter. Moreover, the recent progress of sensor technologies enables measuring human activity using accelerometer, RFID, etc. The mobile phone networks, which have been spreading

Y. Nohara (✉) • N. Nakashima
Kyushu University Hospital, 3-1-1 Maidashi, Higashi-ku, Fukuoka 8128582, Japan
e-mail: y-nohara@info.med.kyushu-u.ac.jp

S. Inoue
Kyushu Institute of Technology, 1-1 Sensuicho, Tobata-ku, Kitakyushu 8048550, Japan

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7_18

all over the world, and wireless sensor network such as body area network enable collecting various sensor data easily from developing countries without computer expert. These sensors carry the promise of drastically improving and expanding the quality of care [3].

A health care process consists of three stages: primary care, secondary care, and tertiary care. Primary care is the first point of medical consultation for all patients and includes preventive medicine. The target of the primary care is large because it includes healthy people for preventing disease but each workload is light. As the people proceeds to secondary care and tertiary care, the number of targets become smaller but people needs more care and each load becomes heavier. At any stage of the health care, we need to manipulate large data and sensing and analysis of data are key components. In this paper, we present two types of sensor data analysis for medical and healthcare.

One sensor dataset is collected in a hospital for secondary and tertiary care. The dataset includes those of patients and nursing care in a cardiovascular center in the hospital. The patients provided vital sensor data such as a monitoring cardiogram and also medical information which were recorded in the electronic clinical pathways. At the same time, we also gathered accelerometer data and RFID data of real nursing in the hospital. We provide the real nursing dataset for mobile activity recognition which could be used for supervised machine learning, and also the big data combined with the patients' medical records and sensors tried for 2 years.

The other sensor dataset is collected in Bangladesh, one of the typical developing countries, for primary care, especially preventing medicine. We developed an eHealth system that comprises a set of sensor devices in an attache case. We provided eHealth checkups for people in Bangladesh. Individual health condition was automatically categorized into four grades based on international diagnostic standards. We provided teleconsultation for affected and emergent subjects and we provided teleprescription for these subjects. The first checkup was provided to 16,741 subjects. After 1 year, 2361 subjects participated in the second checkup, and the blood pressure of these subjects was significantly decreased ($P < 0.001$). Based on these results we proposed a cost-effective method using a predictor, to ensure sustainability of the program in developing countries.

18.2 Sensor Data Analysis in Hospital

Clinical pathways, also known as *Clinical* or *Critical pathways*, are ones of major tools to standardize care processes, in order to increase quality in healthcare [4, 5]. In meeting such an objective, the recognition and data mining of nursing activities can lead to a better understanding and improvements in medical care, and they can help prevent unnecessary activities and excessive work. These approaches are also beneficial to patients because the overall care process is optimized, thus resulting in shorter hospitalization times and lower costs.

Recently, researchers have explored the possibility of human activity recognition with mobile sensors; for example, accelerometers, gyroscopes, and low-frequency audio have been explored [6–8]. In addition, several researchers have applied such technology to domain-specific applications in nursing activities [9–11]. However, in the available methods, several unclear points still remain:

The nature of the real activities is not clear In the application of nursing activity recognition, the *activity classes*—the types of activities—are defined in a domain-specific manner. Here, the activities are not always easy to recognize because the table includes feature value varieties even for single classes, such as blood pressure measurements starting by attaching the corresponding equipment to a patient, followed by pushing air pumps periodically, and finishing with detaching the equipment.

The application is not clear In the application of nursing activity analysis, we can set up clear goals, such as improving nursing activities effectively for timing, duration, and patient satisfaction, or optimizing the costs of the nursing process.

No dataset with clear goals To overcome the aforementioned challenges, we require real data to evaluate or input into a machine learning algorithm. However, there is extreme shortage of such open datasets obtained from multiple subjects, and a set of entire days with densely annotated labels.

For this study, we collected (1) (*labeled data*) actual activities from nurses wearing accelerometers in a hospital for approximately 2 weeks and combined them with training labels, which resulted in 25 activity classes with 5,743 labels from 22 nurses, and (2) (*unlabeled data*) the open big data for 60 nurses for 442 [days $\times$ people] in the trial for almost 2-years with the duty days which could obtain agreements from the nurses and up to 100 patients, combined with patients' wearable, vital, and environmental sensor data and medical records. From the obtained labeled data, we observed that the activities have imbalances in the number of occurrences for each activity class, the starting times in a day, and the duration of each activity class.

Then, we propose a method for recognizing whole day activities using prior knowledge on the information of a sequence of activity segments which are obtained from whole day training dataset, such as the daily timestamps, duration, and imbalances among activity classes, based on our papers [12, 13]. We introduce an analysis of the unlabeled data utilizing the machine learning result of the labeled data, combined with nurses' profiles and medical records, and applying random forest algorithm to generate regression models with considering generalizing ability, and to investigate importances of each predictor variables as well as avoiding interactions between predictor variables, and visualize the effects between predictor and response variables.

18.2.1 Sensor Data Collection for Nursing Activities

We collected mobile-sensor data from the nurses of a hospital's cardiovascular center [14]. The experiment was exclusive to those nurses who agreed to usage of the sensor data, and to the duties related to patients who consented to participate in the experiment.

It includes labeled data for 2 weeks, and unlabeled data for the duty days which we could obtain agreements from up to 100 patients in 2 years. In this subsection, we describe the protocols for data collection and review both of the labeled and the unlabeled datasets.

18.2.1.1 Protocol

We requested the nurses to wear mobile devices (iPod touches) that record accelerations in their breast pockets in a generally fixed direction. They also attached a small accelerometer device on their right wrist, and another on the back of their waist. Figure 18.1 illustrates the attachments. Each sensor measured accelerations on three axes in the range of $\pm 2G$.

Labeled Data Collection

The daytime duties of 22 nurses over the period of 2 weeks on Feb. 2014 were labeled with mobile tablets by other nurses who acted as observers. Before the trial, we defined 41 activity classes from the clinical path, and asked the observers to record them.

Fig. 18.1 Nurses with three accelerometers: one on their right wrist, one attached to their breast pocket, and one on the back hip

Annotating labels for real activities requires careful design. Since the highest priority in real nursing activities is nursing for the patients, there occur a lot of missing labels or incorrect timestamps. Therefore, another nurse records the activities of the subject nurse as observer using iPad. On the software on the iPad, the observer selects the activity class which the subject nurse is about to start, and pushes the finish icon when the subject finishes.

In reality, if the observer waits for the subject nurse to start the activity, the start timestamp will be late. Therefore, they collaborated with each other to have correct start timestamps, such that the subject nurse declares the observer the activity before s/he will start it.

Unlabeled Data Collection

In the same department of the hospital as above, we collected unlabeled sensor data for 2 years from the nurses who wear three accelerometers in the same way as the labeled data collection.

Formatting the Dataset

The datasets for labeled and unlabeled data were formatted uniformly as well as possible. Both datasets are associated with nurse IDs.

Moreover, even the each sensor on each position of the body stores their sensor data separately on the device, it is useful for data analysis to be merged into one multi-column table. Therefore, we joined the data for three devices' data of a duty date to a single table in an offline manner. We firstly generated timestamps increasing by 20 Hz, which means 0.05 s, and adopted the closest sample within 0.025 s for each timestamps. If there are no samples within 0.025 s, we reused the last timestamp value.

Since each device has its own clock and they have no interaction for time synchronization with each other, there is a risk that the clock is not synchronized. To avoid this, we shook the devices together periodically—once in a day on average—as a reference timestamp, and used the relative time from the shaking time as well as possible.

18.2.1.2 Overview of the Dataset

As the result of the experiment, we collected 346.5 [hours × people] of sensor data from 22 nurses by the labeled data collection, and 1,655 [day × people] from 60 nurses by the unlabeled data collection. The activity classes actually observed were 25, listed in Table 18.1. The total number of labels was 5,743. The labels for each activity class are also listed in Table 18.1.

Table 18.1 Observed activity classes and numbers of labels

No.	Activity class	# of labels
1	Anamnese (patient sitting)	2
2	Measure height	45
3	Measure weight (dorsal)	8
4	Measure blood pressure (dorsal)	529
5	Sample blood (dorsal)	16
6	Start intravenous injection	61
7	Finish intravenous injection	40
8	Change drip/line	38
9	Assist doctor	19
10	Find artery	257
11	Examine edema (lie on back)	118
12	Check bedsore (sacrum/back heel)	10
13	Measure ECG	22
14	Attach ECG	54
15	Remove ECG	5
16	Attach bust bandage	29
17	Portable X-ray (prone)	5
18	Change bandage	30
19	Change posture	77
20	Clean body	27
21	Assist wheelchair	86
22	Assist walk	35
23	Move bed	19
24	Wash hands	117
25	Record work (PC)	912

Figure 18.2 shows the plot of the start times for each activity in a day range and the number of activities varies among activity classes. Moreover, we can see that not all the activities occur at any time uniformly. Some activities, such as No. 16, occur only during several hours in the morning or afternoon, and others occur continuously, such as No. 6. Compared with traditional experiment settings where the training data are collected in a balanced way or in a short time without considering the time of day, this may result in difficulties during activity recognition.

Figure 18.3 shows the duration of each activity class and activity duration varies considerably. For example, the maximum duration in the dataset was 9.35 min for "clean body," whereas the minimum was 0.03 min for "measure height." The variances within a class are large, such that the standard deviation of "measure weight" is 8.40 min, and that of "other" is 8.09.

In summary, the real activity dataset attempted for several entire days has imbalances in several aspects, such as class-wise, times of day, and activity duration. If such information is obtained in the training phase, we can expect it to be instructive for improving activity recognition.

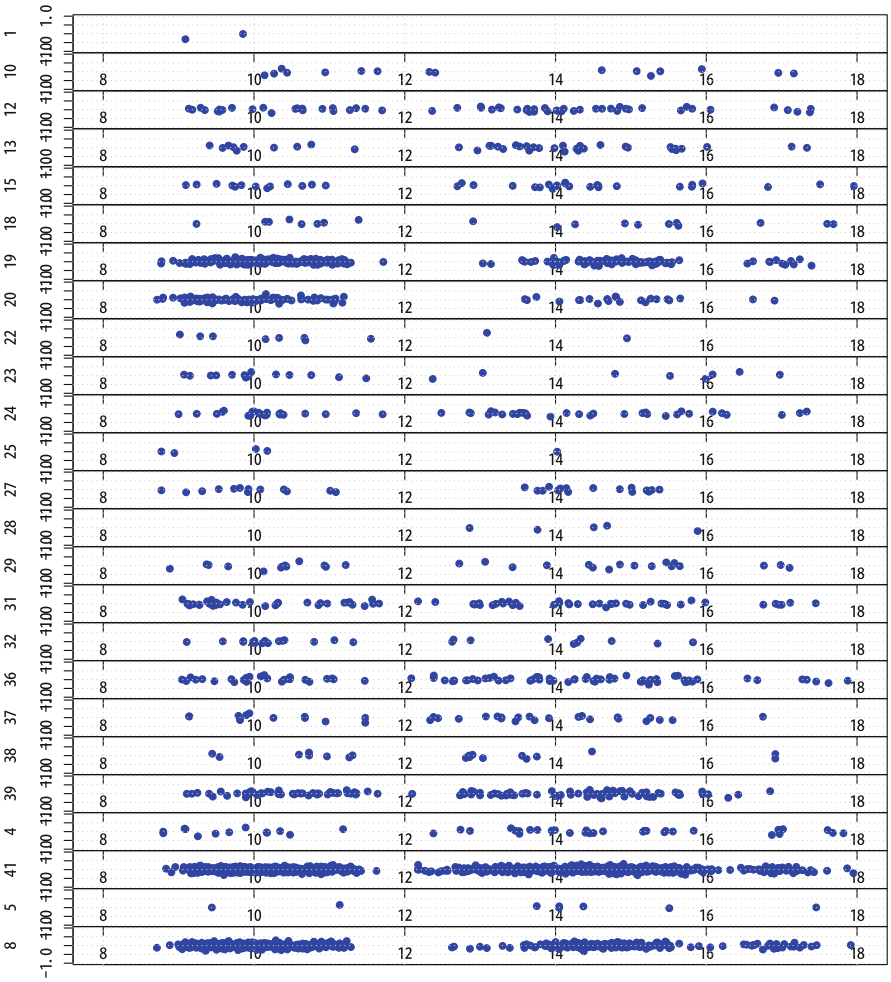


Fig. 18.2 Start time for each activity in a day range. Each row corresponds to an activity class (the number corresponds to the No. in Table 18.1), and the *x*-axis is the hour in a day. The *dots* are the recorded starting time of an activity. We can see imbalances between activity classes and times in a day

18.2.2 Activity Recognition Method

We explain a method for recognizing activities of a whole day [12, 13].

18.2.2.1 Approach

As shown in Fig. 18.2, nursing activities have different possibilities, depending on the time of day. If we have training data with labels and timestamps, we can convert the set of timestamps into the prior probability of the activity being performed.

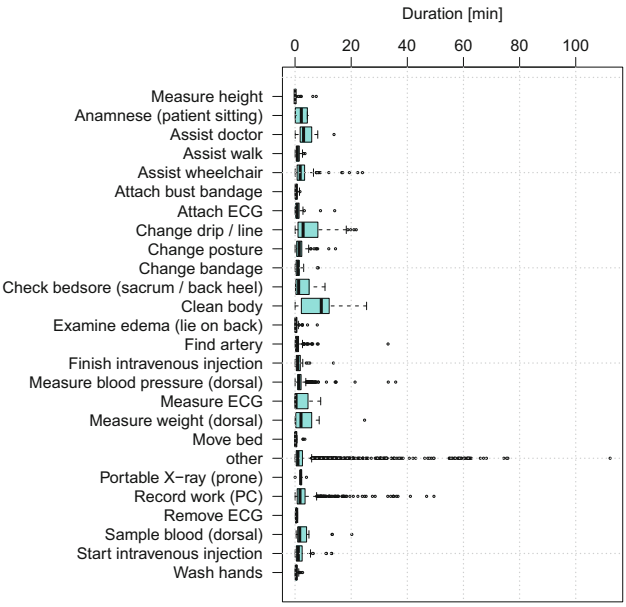


Fig. 18.3 Boxplot of durations of each activity label in the dataset

In addition, if we use both the starting and ending times of an activity, we can obtain information on the activity's duration. Such information of when and how long nursing activities are performed is important for analysis. In our approach, in addition to the traditional method for estimating activities from the sensor input of neighborhood time windows, we exploit the timestamp information in order to construct a prior probabilistic distribution on the activities of an entire day, implement them based on importance sampling, and utilize them for the Bayesian estimation of activities.

18.2.2.2 Method

Before explaining the method, we introduce the mathematical expressions. Figure 18.4 shows an overview of the expression for a single activity class c.

For simplicity, we assume that the time of day is expressed as an integer between one and T. We abbreviate the sequence $(1, 2, \cdots, T)$ as $1{:}T$. For each t, we assume that a feature vector is extracted, t. For each t, we assume that a *feature vector* is extracted that contains several statistic values from the time window of the sensor input around t. We specify the sequence of feature vectors $(x_1, x_2, \cdots, x_T)$ as $x_{1:T}$. Moreover, C refers to the set of activity classes to be recognized. We assume that at any time t multiple activities might be included, either because the nurse is performing several activities concurrently, or because the activity recognition

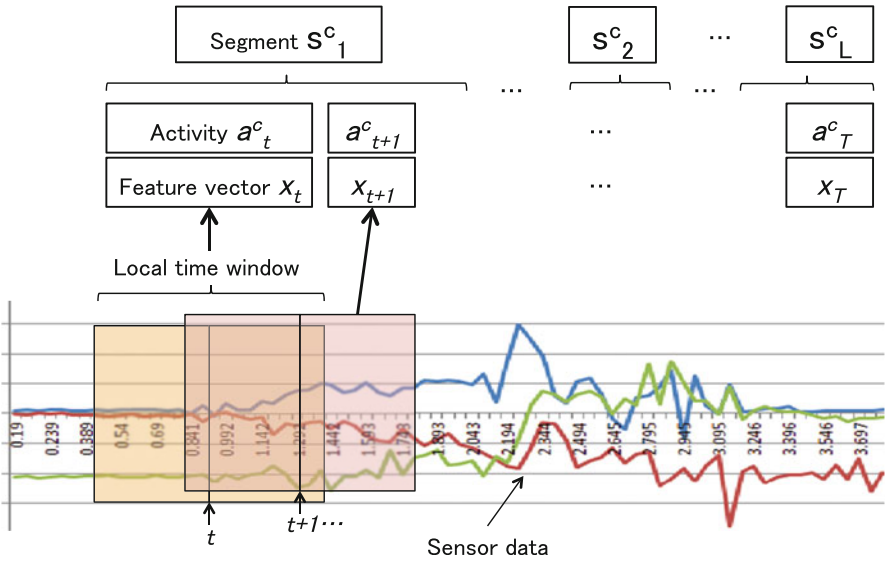


Fig. 18.4 Overview of one-day activities for single activity class $c, \in, C$

algorithm conducts fuzzy estimations. Therefore, we define whether the activity at time t is $c, \in, C$ or not as the binary value a_t^c.

We focus on the recognition of a single activity $c, \in, C$. In reality, we could apply the proposed method for each activity $c, \in, C$, and adopt either the most probable class $\arg_c \max \mathbf{P}(a_t^c)$.

We use the term *segment* as the continuous time range where the activity c is performed, and represent it as a pair of start and end times. When we assume that L^c segments are repeated for activity c in a day, the lth segment from time $b(l)$ to $e(l)$ is defined as:

$$s_l^c := (b(l), e(l)), \text{ where } 1 \leq b(l) \leq e(l) \leq T.$$

Our goal is represented as the problem of obtaining the probability of an entire day's activities $\mathbf{P}(a^c_{1:T}|x_{1:T})$. We assume the Bayesian network as shown in Fig. 18.5. This Bayesian network represents the conditional probabilities for one segment s_l^c. We assume that the probabilities between any segments s_l^c and $s_{l'}^c$ $(l \neq l')$ are independent.

In this model, the probability of an entire day's activities is given as follows:

$$\mathbf{P}(a^c_{1:T}|x_{1:T}) \propto \prod_{l \in 1:L^c} \left\{ \sum_{s_l^c} \mathbf{P}(s_l^c) \prod_{t \in b(l):e(l)} \mathbf{P}(x_t|a_t^c) \right\} \tag{18.1}$$

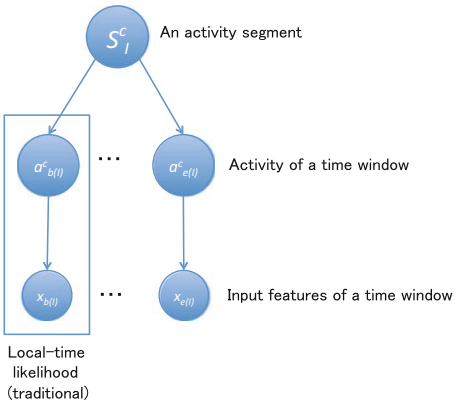


Fig. 18.5 Overview of the proposed method

18.2.2.3 Evaluation

We describe the dataset collected from actual nurses wearing accelerometers in a hospital for approximately 2 weeks, and we evaluate our proposed method by applying it to this collected data.

Preprocessing

We extracted feature vectors from the three axes using the accelerometer data. For the sensor data, time windows of 5 s were extracted, shifting every 2.5 s, as in Bao et al. [15]. For each time window, we calculated 47 feature values, following the literature of [16, 17].

We reduced the 47 feature variables to 27 by applying stepwise-feature selection [18] to 1,000 randomly sampled vectors over ten iterations. The feature variables that were selected are listed in Table 18.2.

Result

In order to evaluate our proposed method, we compared the *proposed* method with the prior knowledge about $\mathbf{P}(s_t^c)$, and the *naive* method without the prior knowledge. As underlying machine learning algorithms for $\mathbf{P}(x_t \mid a_t^c)$, which is the same as the naive method after applying the Bayes' theorem, we adopted naive Bayes (*NaiveBayes*) and evaluated it. In order to evaluate the accuracy of real usage where the training and usage data are different, we applied 1-*duty-day*-left-out cross validation, which means testing each nurse's working day with the model trained with the data that have either different days or different nurses.

Figure 18.6 shows the results for Naive Bayes as the underlying machine learning algorithm. When we adopt Naive Bayes as the underlying algorithm, BCR for the naive method is 55.15 % ($\sigma = 15.8$), and for the proposed method, it is 80.96 % ($\sigma = 14.5$).

Table 18.2 List of feature variables after feature selection

No.	Feature	Sensor	Axis (if any)
1	Mean intensity	Chest	
2	Mean intensity	Right wrist	
3	Mean	Chest	Y
4–6	Mean	Waist	X, Y, and Z
7–9	Mean	Right wrist	X, Y, and Z
10	Variance of intensity	Right wrist	
11–13	Variance	Right wrist	X, Y, and Z
14–15	Variance	Chest	Y and Z
16	Variance	Right wrist	Z
17–18	Mean FFT-domain energy	Chest	Y and Z
19–20	Mean FFT-domain energy	Right wrist	X and Z
21	Mean sum of the absolute values of each axis	Chest	
22	Mean sum of the absolute values of each axis	Waist	
23	Number of samples out of mean intensity $\pm\, 0.1G$	Right wrist	
24	Number of samples out of mean intensity $\pm\, 0.1G$	Waist	
25	Number of crosses of the zone of the mean intensity $\pm$ $0.1G$	Waist	
26	Number of crosses of the zone of the mean intensity $\pm$ $0.1G$	Right wrist	
27	Covariance between intensities	Chest and waist	

18.2.3 Activity Recognition for a Whole Day

For the unlabeled data, we extracted 265,002 time windows, which corresponded to 771 [duty-days $\times$ nurses], and applied our proposed method in order to estimate the real activities involved in nursing duties. The average time for the defined care time is 277.8 min with $\sigma = 55.7$.

Figure 18.7 is the estimated average care times for each activity class in one daytime. From the figure, we can see the types of activity on which the nurses spend more time, such as "Measure blood pressure" and "Find artery." We can also see that the nurses spend significant time recording their work on a PC, which were introduced after the electronic medical record system, and there is an opportunity for reducing this time.

18.2.4 Correlation with the Nurses' Profile

From joint data of the activity recognition results and additional data, we can mine further knowledge. To demonstrate this, we joined the activity results with nurses' profile, i.e., the number of experienced years, age, gender, title, and the ward

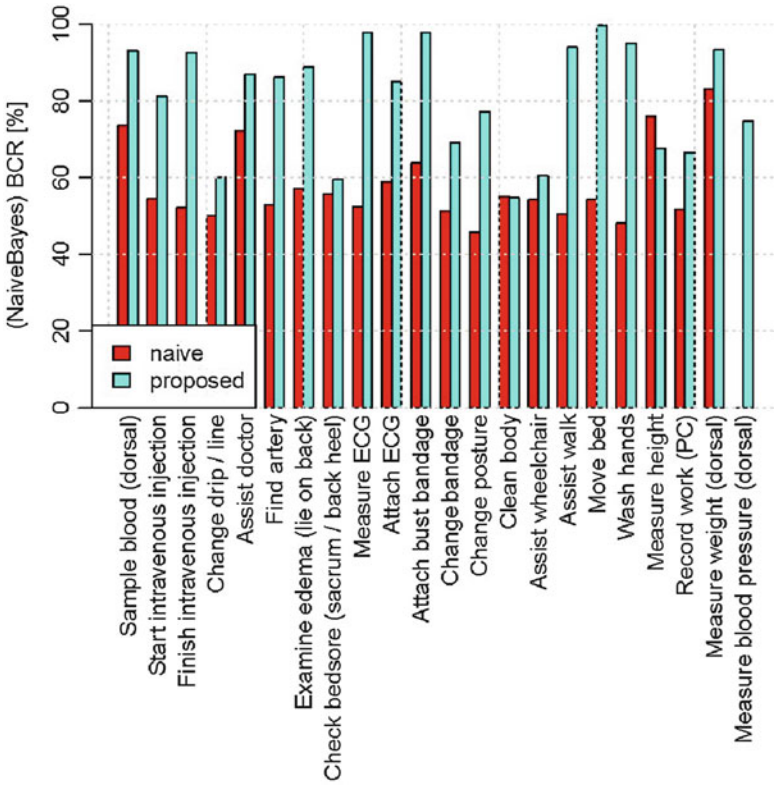


Fig. 18.6 BCR for naive/proposed methods for each activity with Naive Bayes (Average: 55.15) % for the naive method and 80.96 % for the proposed method

(west/east, where the west ward is for internal medicine, and the east is for surgery) for each nurse. The range of experienced years from 1-years to 25-years, and the mean is 7.36 years with $\sigma = 5.65$. The joined data consists of 64 samples each of which corresponds to each nurse.

For the joined data, we applied random forest algorithm for each activity class as a response variable, and the profiles as predictor variables. Random forest has several advantages compared with a traditional regression:

1. random forest automatically avoids over-fitting, and outputs general models
2. we can see the *importances* of variables after neutralizing interactions among variables unlike a traditional regression
3. we can also see the effect of each variable to the response variable by *a partial dependent plot* with neutralizing interactions
4. if we pick up a tree from the set of obtained trees, we can easily understand the partitioning conditions compared to other algorithms such as SVM

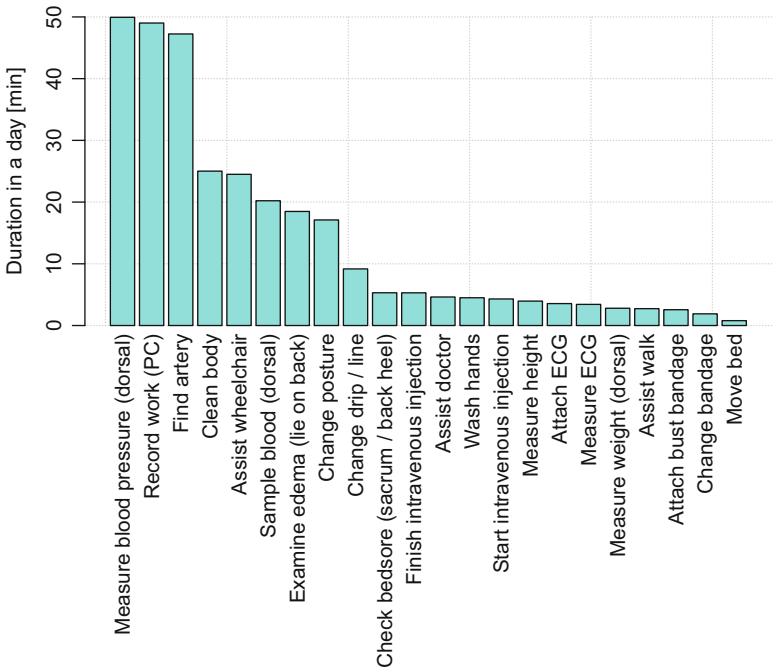


Fig. 18.7 The nursing times for each activity class in one daytime

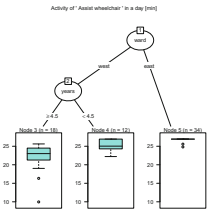
For the models for each activity class as response variable, we picked up the models which have the *pseudo* R^2, which is defined as

$$1 - \frac{(\text{mean squared error})}{(\text{variance of the response variable})},$$

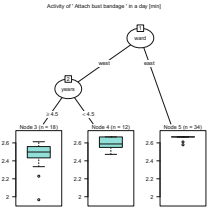
are more than 20 %, and showed the scores and (no random but naive) regression trees in Figure 18.8 to visualize example trees.

In any tree in the figure, the first partitioning is done by "ward." They have higher activity durations for the east ward, and are divided into experienced years < 4.5 [years] with middle durations and the rest with lower durations. It seems that there are differences in nursing activity durations in a day between the west ward, internal medicine department, and the east one, surgery department. For the internal medicine department, it seems that there are varieties of durations, and unexperienced nurses performed longer compared with experienced ones.

From such results, we or the nurses can estimate the differences of work load between wards or the years of experiences, and reallocate and equalize unbalanced work load, if necessary.

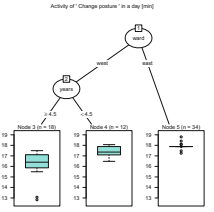


MSR: 6.98, R^2: 20.93%
(a) "Assist wheelchair"

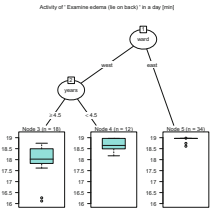


MSR: 0.012, R^2: 22.08%
(b) "Attach bust bandage"

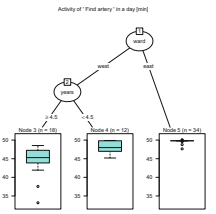
MSR: 0.018, R^2: 22.01%
(c) "Attach ECG"



MSR: 0.843, R^2: 21.61%
(d) "Change posture"

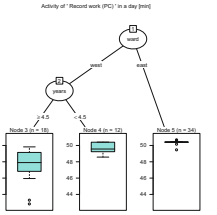


MSR: 0.253, R^2: 23.98%
(e) "Examine edema (lie on back)"



MSR: 7.32, R^2: 22.93%
(f) "Find artery"

MSR: 2.72, R^2: 22.85%
(g) "Measure blood preessure (dorsal)"



MSR: 1.98, R^2: 26.27%
(h) "Record work (PC)"

Fig. 18.8 Regression trees for each activity duration in a day by nurses' profiles with over 20 % of pseudo R squared with RandomForest. Each caption shows the mean of squared residuals (MSR) and the pseudo R squared with RandomForest. (a) Assist wheelchair. (b) Attach bust bandage. (c) Attach ECG. (d) Change posture. (e) Examine edema (lie on back). (f) Find artery. (g) Measure blood pressure (dorsal). (h) Record work (PC)

18.2.5 Correlation with Patients' Discharge Delays

We joined the estimated activity data with the patient record, and compared the amount of time spent by nurses for each activity with the duration of hospitalization, where within 4 inpatient days are normal, and over 5 days increase medical costs. In the experiment, we asked the nurses to attach RFID tags, which communicate with readers which were equipped at the entrances of the private patients' rooms. The log of the RFID readings provides relationships between nurses and patients who took care of/have taken care of at the day. We firstly joined the RFID records with the estimated activity data, and then joined with the patients' data about the hospitalization days. The number of patients after joining with the patient record is 28 with 24 nurses for 35 days, and the number of the samples is 54.

For the joined data, we applied random forest algorithm again with the numbers of hospitalization days as a response variable, and with the activity durations in a day as predictor variables.

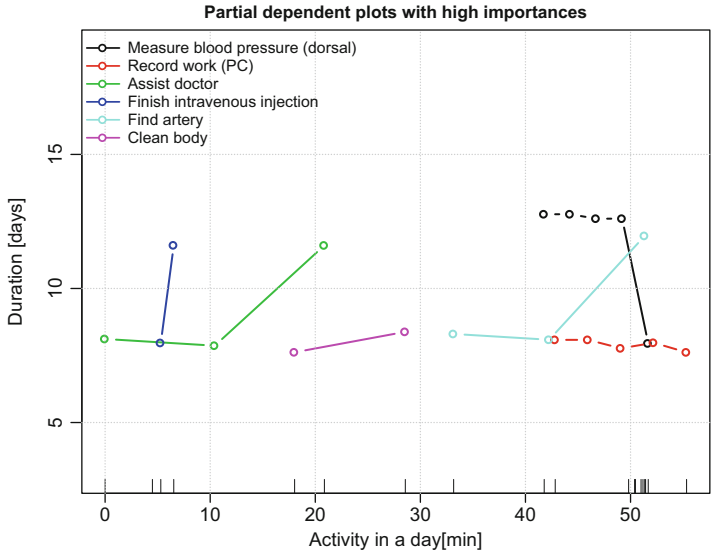


Fig. 18.9 Partial dependent plots to show the effect of each nursing activity time in a day (with high importance) to hospitalization durations of the cared patients

Then, we can obtain the importances for each predictor variables, which is measured as the mean increase of accuracy by the variable. For the importances, we picked up the variables with the importances of over 50, and showed the partial dependent plots in Fig. 18.9. Partial dependent plot is a plot between a predictor and the response variable, which plots the effect of the predictor to the response after marginalizing the other predictor variables. Unlike traditional regression, it can approximately eliminate the interaction effects between response variables.

From the figure, we can observe that "Measure blood pressure" takes shorter times for patients with longer hospitalizations, the time for "Assist doctor" increases if the patient have longer hospitalizations, "Finish intravenous injections" takes slightly longer if the hospitalizations take longer, and "Clean body" takes quite longer if the hospitalizations take longer. As such, we can estimate the effect of each nursing activity time to the hospitalization duration, and we can estimate the work load for each patient.

As shown in the above, by linking our proposed method with additional data which already exist in hospitals, we can produce a valuable knowledge for reflecting and improving medical processes.

18.2.6 Related Work

In the literature, many works attempted mobile activity recognition [6–8, 15]. A few papers also attempted to apply nursing activity recognition and application [9–11].

Several datasets for mobile activity recognition are available. Hattori et al. [19], with their large-scale activity collection, collected over 35,000 activities from more than 200 people over approximately 13 months. Kawaguchi et al. [20] was a unique trail to collect activity recognition datasets from the laboratories of multiple universities. In the 5 years, the total number of activities reached over 50,000 samples. Chavarriaga et al. [21] and Roggen et al. [22] provided an activity dataset with sensor-rich environment where the subjects wore multiple sensors on the body, with more than 27,000 activities from 12 subjects. Among them, Chavarriaga et al. [21] provided an entire day data/multi-day data as a part of them. However, the activity classes are common types, such as those that appear in activity in daily life (ADL) records, and not similar to our dataset, which is closely coupled with the application domain and domain data, such as medical records.

18.2.7 Conclusion

In this section, we collected a real nursing dataset for mobile activity recognition that can be used for supervised machine learning, and proposed a method for recognizing activities for an entire day utilizing prior knowledge about the activity segments in a day. The results showed accuracy improvement compared with the baseline method. We also demonstrated data mining by applying our method to bigger data combined with 2 years of patient medical records, and demonstrated the value of linking with additional day utilizing random forest regression.

18.3 Sensor Data Analysis in Developing Countries

The prevalence of non-communicable diseases (NCDs), such as heart disease, stroke, cancer, chronic kidney diseases, and diabetes mellitus, has been increasing rapidly worldwide. The World Health Organization reported that NCDs accounted for 63 % (36 million) of the 57 million global deaths in 2008 and approximately 80 % of all NCD-related deaths occurred in low- and middle-income countries [23]. In these developing countries, 29 % of NCD-related deaths occurred in the working-age group (in people aged < 60 years). This rate is higher than that for high-income countries (13 %) and contributes to declining labor productivity in developing countries. The total number of annual NCD-related deaths is estimated to reach 55 million by 2030 [24]. NCDs are no longer just a problem for high-income countries, but a problem that affects all countries.

Preventive medicine is the key to combat NCDs. Preventive medicine comprises three levels: primary prevention (maintaining a healthy condition), secondary prevention (avoiding the development of NCDs), and tertiary prevention (preventing the progression of NCDs into serious medical conditions). Over time, the focus of medical services in developed countries has changed from acute and

serious diseases to the management of chronic diseases. Developing countries have the opportunity to follow a different path, through the implementation of low-cost preventive and compassionate health care/medical services based on information and communication technology (ICT) [25].

In this study, we aimed to evaluate the impact of our preventive health care/medical program consisting of primary, secondary, and tertiary prevention services provided to > 10,000 subjects in Bangladesh. We selected Bangladesh as the research area because, while there are few medical institutions in rural areas, there are many pharmacies and the mobile Internet network has spread throughout the nation, as is the case in many developing countries.

We conducted the research over 2 years, applying eHealth solutions and telemedical interventions in an attempt to ensure an accurate stratification balance and assess the effects of intervention after 1 year in > 2000 subjects [26]. The target diseases in this program are chronic NCDs including diabetes mellitus and hypertension, which are rapidly increasing in developing countries. Sensor devices, including blood glucose meters and blood pressure meters, have been developed for the management of these diseases and are widely available. The World Bank Group's *Disease Control Priorities in Developing Countries* has also emphasized the paramount importance of risk management of chronic NCDs [27]. Using machine learning, we attempted to establish a method to decrease the cost of health checkups by predicting the results of expensive health check tests.

The main part of this section is first published in the Journal of Medical Internet Research [26].

18.3.1 Methods

18.3.1.1 Overview

We developed an eHealth system named the Portable Health Clinic (PHC). PHC comprises a set of sensor devices in an attache case, a data transmission system linked to a mobile network, and a data management application. The system can be used by operators with minimal information technology literacy to provide health checkup services, even in rural areas. We included a teleconsultation service using Skype over the mobile network to gather data on health. To assess the usability and sustainability of the system, we designed a study model including local pharmacies to provide a teleprescription service. We conducted a field study from July 2012 to March 2014 (first year: July 2012–February 2013; second year: June 2013–March 2014).

Fig. 18.10 The Portable Health Clinic System package

18.3.1.2 The Portable Health Clinic

We selected sensor devices based on international information standards and approved by Japanese pharmaceutical law. If a device did not have a standard transmission format, we attached a body area network (BAN) interface to the sensor. The BAN was published as IEEE802.15.6 in 2012 and uses frequency bands approved by national medical and/or regulatory authorities and the industrial, scientific, and medical (ISM) band [28]. In addition to the dedicated medical bands, it provides quality service, extremely low power, and a data speed of up to 10 Mbps, supports medical security, and emergency data handling.

For easy portability, we put the components into an attache case. The attache case was equipped with an Android tablet, consumable goods, including urine and blood sugar test strips, batteries, paper, and pens. The total weight of the attache case and contents was approximately 10 kg (Fig. 18.10).

An Android tablet served as a data input terminal, aggregating data via BAN and manual input and communicating with the sensor server. Results of individual health checkups, including the stratification level, were shown on a local site. The main server in a medical call center in Dhaka, the capital of Bangladesh, stored all sensor data. Data was available to doctors through the call center.

Local servers temporarily stored data from the Android tablet via wireless-LAN and synchronized data with the main server when an Internet connection was available. This use of local servers enabled PHC operators to upload their data even if an Internet connection was temporarily unavailable.

18.3.1.3 Stratification Algorithm

Before the study commenced, we established "Bangladesh-logic" for risk stratification using international diagnostic standards [29–33] to rank the risk grade into four groups—green (healthy), yellow (caution), orange (affected), and red (emergent) based on the results of each health checkup item (Table 18.3). The overall health condition of each subject was also determined by integrating the results of questionnaires into the four groups by the worst color of all health checkup items.

Table 18.3 Bangladesh-logic: criteria for risk stratification

		Green	Yellow	Orange	Red
Waist (cm)	Male	[0, 90)	[90, ∞)		
	Female	[0, 80)	[80, ∞)		
Waist/hip ratio	Male	[0, 0.90)	[0.90, ∞)		
	Female	[0, 0.85)	[0.85, ∞)		
Body mass index (kg/m^2)		[0, 25)	[25, 30)	[30, 35)	[35, ∞)
Blood pressure (mmHg)	Systolic	[0, 130)	[130, 140)	[140, 180)	[180, ∞)
	Diastolic	[0, 85)	[85, 90)	[90, 110)	[110, ∞)
Blood sugar (mg/dl)	Fasting	[0, 100)	[100, 126)	[126, 200)	[200, ∞)
	Postprandial	[0, 140)	[140, 200)	[200, 300)	[300, ∞)
Urine-protein		–	±	+, 2+	
Urine-sugar		–	±	+, 2+	
Urine-urobilinogen		±		+, 2+	
Pulse rate (bpm)		[60, 100)	[50, 60)	[0, 60)	
			[100, 120)	[120, ∞)	
Arrhythmia		None	+		
Smoking		None	+		
Body temperature (°C)		[0, 37)	[37, 37.5)	[37.5, ∞)	
Oxygen saturation (SpO2) (%)		[96, 100]	[93, 96)	[90, 93)	[0, 90)

Examples of the determination of overall health conditions based on the results of each health checkup item are as follows:

– Green, Orange, Green, ⋯, Yellow → Orange
– Green, Green, ⋯, Green (all Green) → Green

The presence or absence of arrhythmia was determined using a blood pressure meter. Data on smoking and time since the last meal were obtained from questionnaires.

18.3.1.4 Questionnaires on First and Second Visits

During health checkup visits, we surveyed subjects using questionnaires written in Bengali. Because the literacy rate is $< 60\,\%$ in Bangladesh, a staff member read the questionnaire to the subjects and entered response data into the system. On the first visit, we asked about literacy, occupation, time since the last meal, present symptoms, past diseases, medication, smoking, weight change, exercise, walking speed, eating behavior, sleeping, and the desire to have a healthy lifestyle. For orange- and red-grade subjects, we administered the questionnaire for teleconsultation with questions including information on drug allergy and surgical history. For subjects who participated in both the first and second years, we administered a different questionnaire at the second visit with questions regarding memory and effects of the

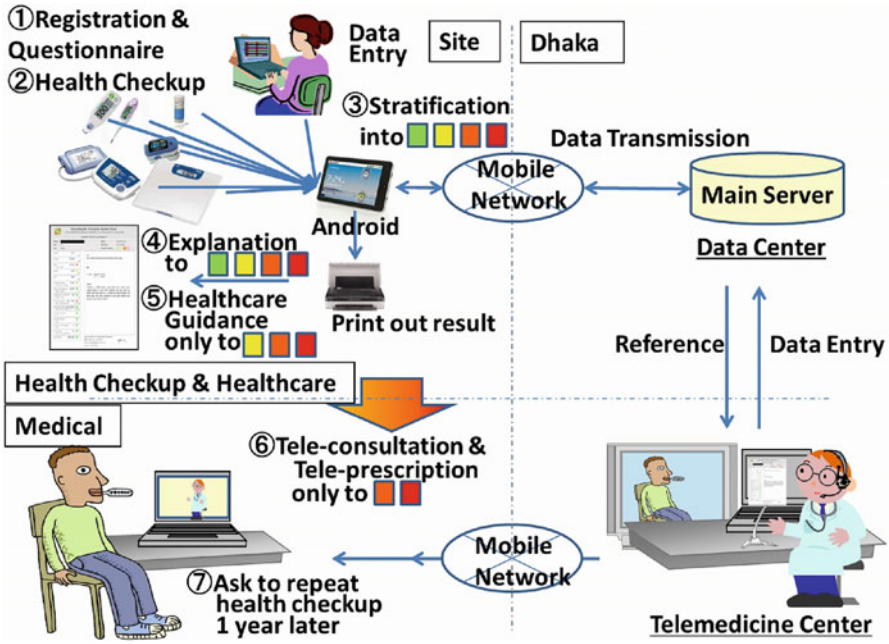


Fig. 18.11 Work flow (Steps 1–7) and data flow (*arrows*) of the service

first health checkup, psychological (Prochaska's) staging, present symptoms, and medication.

18.3.1.5 System Operation

We provided a health care service for the study, including a health checkup using sensor devices in the PHC, data storage in the call center, a health report, health care guidance according to the situation of individuals, and a teleconsultation with a doctor in the medical call center (Fig. 18.11). We conducted the study in five rural villages and five factories/offices in Bangladesh. In the first year, we conducted the study around Dhaka (Dhaka, Shariatpur, Chandpur, and Gazipur) because of a logistics problem. In the second year, we selected sites from all over the country to check the country's health status. The sites of the second year are Chittagong (south-eastern area), Rajshahi (western area), Thakurgaon (north-western area), and around Dhaka.

At the first visit, after registration, the subjects received an ID card with a barcode (Step 1 in Fig. 18.11). After completing the questionnaire, the subject underwent the health checkup with the sensor devices in the PHC (Step 2). Both the blood glucose and urine tests were performed by qualified health care professionals, whereas the other tests were performed by trained staff. We cross-checked the urine test results of the workers every 2 or 3 months because this test

requires visual judgment. Other devices, including blood glucose devices, display numerical results and we do not require calibration among workers. The data were stored in an Android tablet and in the main server in Dhaka. Categorized results for the four risk groups, graded from green to red according to the Bangladesh-logic, were printed out (Steps 3 and 4) and explained to the subject by the local staff (Step 4). A booklet was provided to all subjects graded yellow, orange, or red (Step 5). We provided telemedical intervention with a doctor in Dhaka for orange- and red-grade subjects (Step 6). Because we selected sites from around Dhaka in the first year, subjects in the village and factories/offices around Dhaka were asked to undergo the health checkup 1 year later to enable assessment of the effects of the program (Step 7).

18.3.1.6 Teleconsultation and Teleprescription

After the health checkup, we provided telemedical intervention for orange- and red-grade subjects via mobile network contact (Skype) with the medical call center in Dhaka. Because most areas in Bangladesh have Internet access (2G/3G), we brought laptop PCs or tablet PCs (iPad) with mobile routers to the checkup sites. The staff set up special rooms for teleconsultations at checkup sites and assisted subjects to communicate with remote doctors in Dhaka. Doctors had access to the results of health checkups via the Internet and they were able to provide advice on disease management and encourage subjects to visit a clinic. Where required, the doctors could send a teleprescription for anti-hypertensive medication via the network. In our program, subjects who received a teleprescription could visit their local pharmacy to purchase medication.

18.3.1.7 Booklet for Health Guidance

We provided an 11-page booklet to educate subjects graded yellow, orange, or red. The booklet contained information on the risks of NCDs, including obesity, hypertension, diabetes mellitus, smoking-related diseases, and chronic kidney disease. We prepared both English and Bengali versions of the booklet and provided the Bengali version to the subjects in the study. For subjects who could not read Bengali, a staff member explained the checkup results and provided health care guidance orally.

18.3.1.8 Ethical Considerations

The Kyushu University Institutional Review Board for Clinical Trials approved the protocol of this verification study in 2012. We applied for the IRB of Kyushu University, Japan, because participant groups in Bangladesh had no IRB. We prepared a consent form, etc., after discussing with the local doctors.

18.3.2 Results

18.3.2.1 Overview

There were 16,741 subjects assessed at the first health checkup, 9143 (54.61 %) males and 7598 (45.39 %) females (Table 18.4). There were 9309 (55.61 %) subjects from urban areas and 7432 (44.39 %) subjects from rural areas. Most of the subjects in urban areas were male (male/female=6299/3010], whereas female subjects were more numerous in the rural areas (male/female=2844/4588). Figure 18.12 shows images of the health checkup and teleconsultation process in a rural area.

Table 18.4 The number of subjects by sex and location (the number of subjects who participated in both the 2012 and 2013 checkups are indicated in parentheses)

Location	Male	Female	Total
Rural	2,844 (177)	4,588 (234)	7,432 (411)
Urban	6,299 (1,412)	3,010 (538)	9,309 (1,950)
Total	9,143 (1,589)	7,598 (772)	16,741 (2,361)

Fig. 18.12 Images of a health checkup and teleconsultation

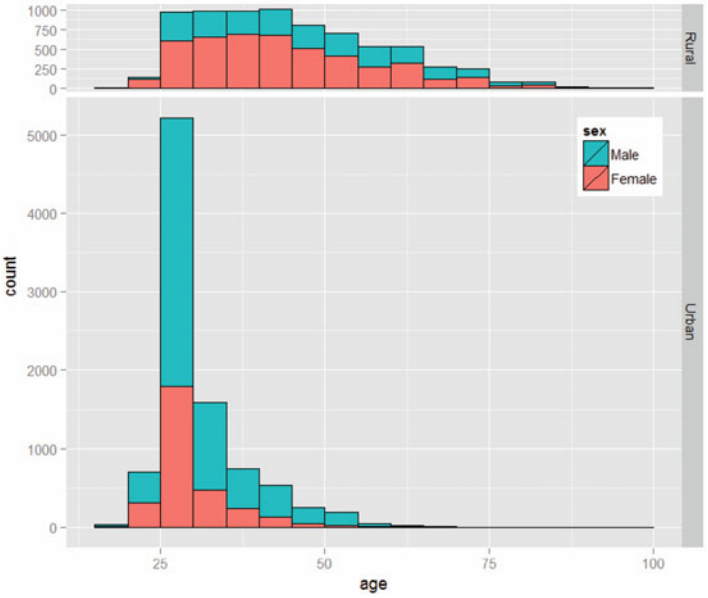


Fig. 18.13 Age distribution for rural and urban areas

Figure 18.13 shows the age distribution for rural and urban areas. There was a wide age distribution in rural areas, while there was a clear peak in the number of subjects aged in the late 20s in urban areas. The average age was 35.1 (SD 12.7) years for male subjects and 36.7 (SD 12.8) years for female subjects. The average age was 43.6 (SD 14.0) years for rural subjects and 29.6 (SD 6.9) years for urban subjects.

The results of the first health checkup are shown in Fig. 18.14. Based on the assessment of overall health condition, we identified 5419 out of 16,741 subjects (32.37 %) as affected (orange or red) and 9057 subjects (54.10 %) as caution required (yellow). There were 10,879 subjects (64.98 % of the total 16,741) graded yellow based on the waist/hip ratio, 5535 (33.06 %) graded yellow or higher based on a blood pressure test, and 1402 (8.37 %) graded yellow or higher based on a blood sugar test. Subjects were graded red (emergent) based on body mass index (BMI) (39 subjects), blood pressure (258), blood glucose (181), and oxygen saturation (SpO2) (6). We provided a teleconsultation service to affected subjects (orange or red, n=4899).

18.3.2.2 Risk Factors Associated with Overall Health Condition

Figure 18.15 shows the overall results with regard to age, sex, and area. To identify risk factors for NCDs related to overall health condition results, we used logistic regression analysis with the overall result (orange/red: true or green/yellow: false)

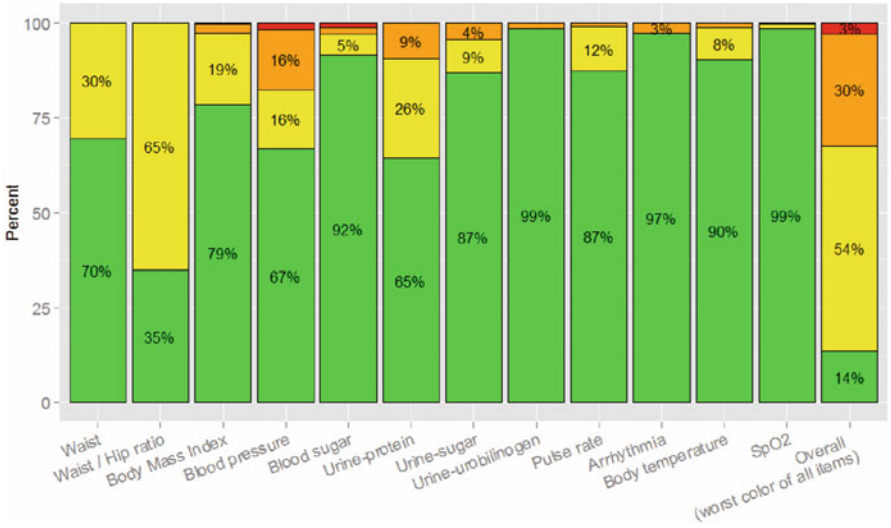


Fig. 18.14 Age distribution for rural and urban areas

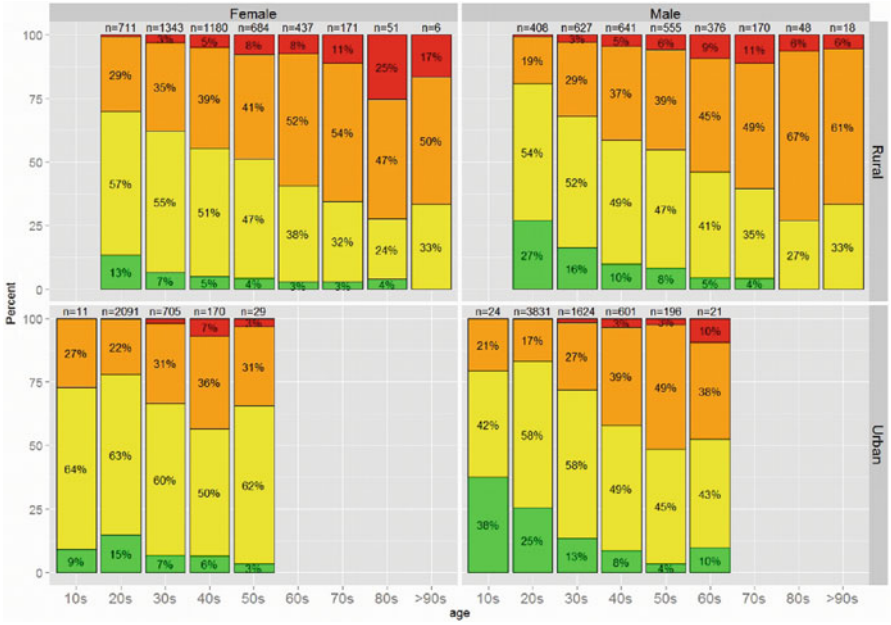


Fig. 18.15 Overall results by age, sex, and area (subject numbers are shown above the columns)

Table 18.5 Risk factors for NCDs associated with the overall health

Variables		Odds ratio (95 % CI)	*P* value
Age		1.04 (1.04–1.05)	< 0.001
Sex (male)		0.78 (0.71–0.86)	< 0.001
Area (urban)		0.66 (0.56–0.78)	< 0.001
	Daily labor	0.56 (0.45–0.69)	< 0.001
	Business	0.83 (0.68–1.01)	0.063
Occupation	Private/government service	1.00 (reference)	–
	Student	0.43 (0.27–0.66)	< 0.001
	Housewife	0.86 (0.71–1.03)	0.108
	Unemployed	0.80 (0.64–1.01)	0.055
Literacy		1.24 (1.14–1.36)	< 0.001

Bold variables were significantly associated with overall health condition($P < 0.05$)

as the outcome and age, sex, site type, occupation, and literacy as independent variables (n=16,315). The results are presented in Table 18.5. Variables that were significantly associated with overall health condition ($P < 0.05$) are noted.

The results of the analysis indicate that older age, female, and living in a rural area were risk factors for NCDs. Contrary to our expectations, literacy (and not illiteracy) was also a risk factor. When we changed the outcome of logistic regression analysis from the overall result to the result of each individual checkup item, we found that literacy was also a risk factor for high BMI, blood pressure, blood glucose, and urine glucose. Conversely, there was no significant difference between literacy and illiteracy for urine protein and urobilinogen levels, pulse rate, arrhythmia, body temperature, and SpO2.

Because significant variables were related to body mass, we generated the hypothesis that literate subjects: (1) earn more and tend to overeat, (2) do not get enough exercise because they use their own mode of transport or public transport, and (3) lack a basic awareness of health.

18.3.2.3 Comparison with Results of Health Checkups in Japan

Because NCD problems are spreading from advanced countries to developing countries, it is important for preventive medicine in developing countries to use past experiences in advanced countries. Moreover, experiences in developing countries could improve preventive medicine in advanced countries, the so-called reverse innovation. On the other hand, problems specific to each country exist and we need to cope with individual issues. In order to separate the problems, we compared the results of the present study with data from the 2012 National Health and Nutrition Survey in Japan (n > 15,000), which was conducted by the Japanese Ministry of Health, Labor, and Welfare [34]. The results of the present study were corrected to match the sex and age distribution of the Japanese dataset. Figure 18.16

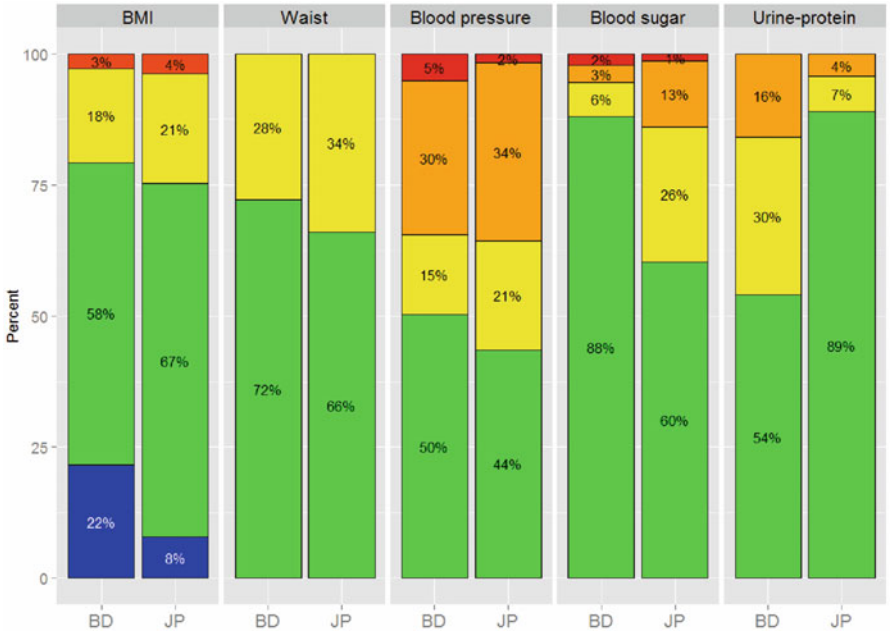


Fig. 18.16 Comparison of health checkup results for Bangladeshi (BD) and Japanese (JP) subjects

shows the results of the comparison. The blue color in the BMI column indicates BMI $< 18.5\,\mathrm{kg/m^2}$, meaning subjects were underweight.

The comparison shows that there were many underweight subjects (blue in BMI) in Bangladesh and that more Bangladeshi subjects were ranked green compared with Japanese subjects. However, the number of Bangladeshi and Japanese subjects ranked orange and red for BMI, waist/hip ratio, and blood pressure was similar, despite very different average income, living conditions, and eating habits in the two countries. Conversely, the results of blood sugar and urine protein tests were quite different between the two countries. This may be a result of regional differences because the results of the urine protein test differed among sites in Bangladesh (Fig. 18.17). We would like to conduct further research and learn to cope with the individual issues in Bangladesh.

18.3.2.4 The Second Health Checkup

There were 2361 subjects who participated in both the 2012 and 2013 health checkups. The details are indicated in parentheses in Table 18.4. Mean systolic blood pressure (SBP) in the first year (2012) was 121 (SD 17) mmHg and in 2013 it was 116 (SD 15) mmHg. Figure 18.18 shows the difference in SBP between the 2 years arranged by the ranked color of the first blood pressure test. There was a significant decrease in SBP for all color rankings ($P < 0.001$).

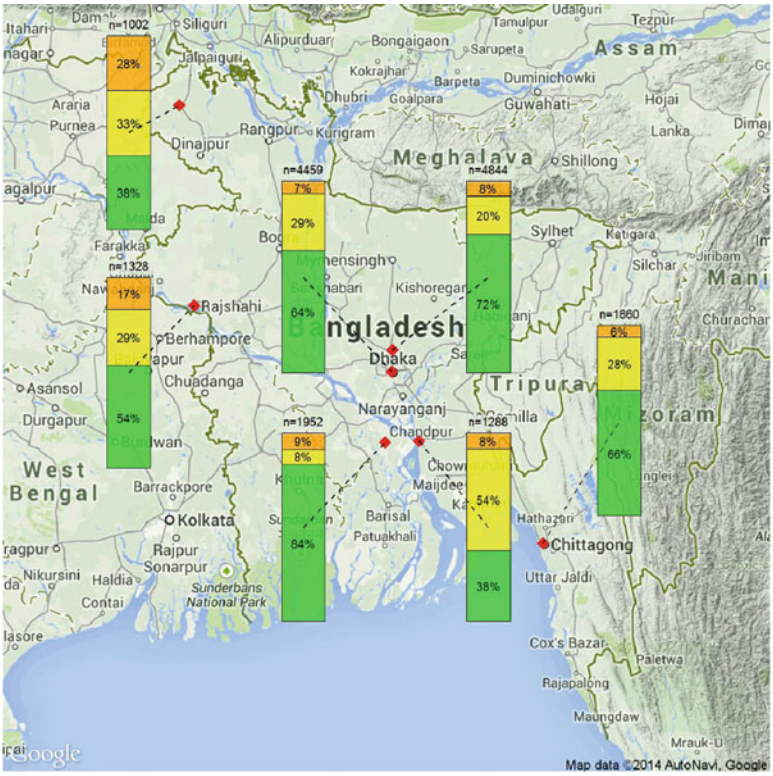


Fig. 18.17 Results of the urine protein test for each checkup site in Bangladesh

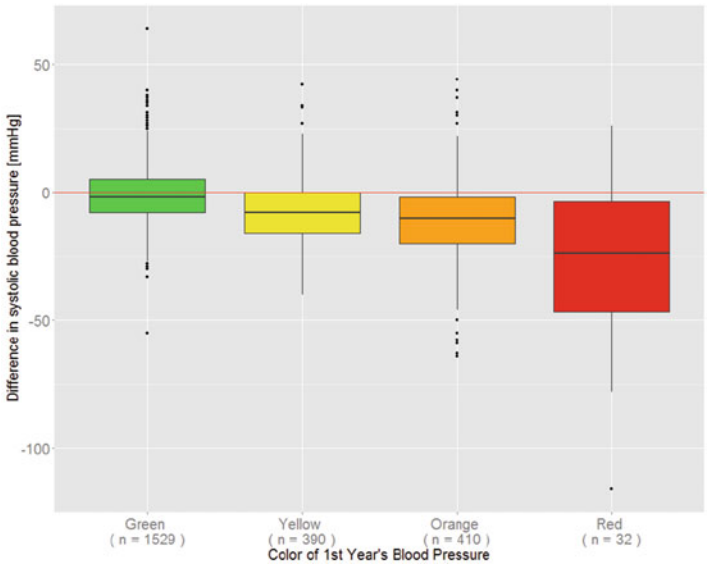


Fig. 18.18 Difference in systolic blood pressure based on color ranking at the first checkup

To determine which subjects showed improved health over the study period, we attempted to predict which subjects would have better health at the second checkup. There were 2110 subjects who had a medical interview at the first checkup and 640 of those subjects were graded as red or orange. Of those 640 subjects, 326 improved their health to green or yellow in the second year.

To further investigate the subjects with improved health, we applied a machine learning technique, the random forest method, using the medical interview, subject profiles, and checkup results as the explanatory variables. From the 640-subject dataset, we separated 60 % as training sets and 40 % as test sets without replacement and ran the estimation. The area under curve (AUC) for 20 trials was 0.7676 (SD 0.0267). The main factors that contributed to the estimation were age, BMI, waist/hip ratio, urine protein, and blood pressure. Based on these findings, we proposed a preferred intervention to help subjects to improve their health.

18.3.2.5 Predicting Blood Glucose Test Results

We applied the random forest method using the medical interview, subject profiles, and checkup results (excluding the blood glucose test result) as explanatory variables to estimate the Bangladesh-logic ranking of red and orange based on the blood glucose test. From the 15,705-subject dataset (true/false=462/15,246), we separated 60 % as training sets and 40 % as test sets without replacement, and ran the estimation. AUC for 20 trials was 0.9565 (SD 0.0072). Figure 18.19 shows a receiver operating characteristic (ROC) curve (AUC=0.9618).

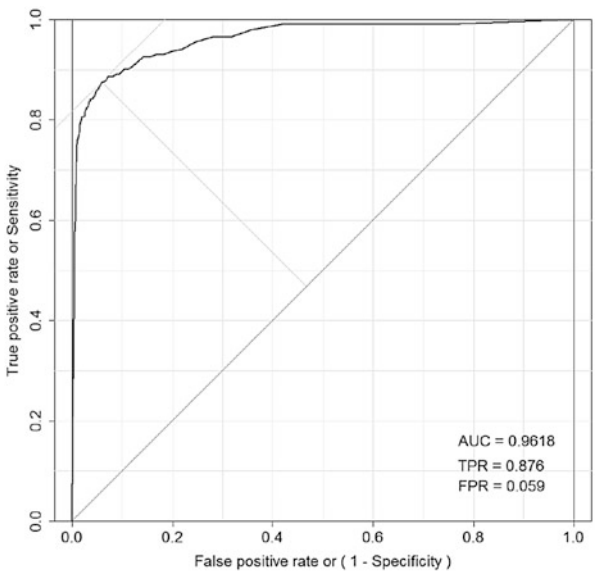


Fig. 18.19 ROC curve for predicting blood sugar

Based on Youden's index, which maximizes the distance from the 45° line on the ROC curve (the upper left point in Fig. 18.19), the true positive rate was 87.6 % and the false positive rate was 5.9%.

18.3.3 Discussion

18.3.3.1 Comparison with Prior Work

In recent years, there have been many projects aimed at improving health care in developing countries. Some projects have focused on more specialized and technical approaches, including immunological or enzymatic assays for bacterial toxins [35], and retinal photography, Doppler imaging, biothesiometry, and electrocardiography to detect diabetic complications [36]. Ramachandran et al. [37] showed that lifestyle modification could prevent type 2 diabetes in Asian Indian subjects. We targeted the general population and focused on primary prevention; therefore, our eHealth system comprised only basic biosensors to conduct the health checkup.

In this study, we detailed the design of our health care program and presented the results of the study conducted in five villages and five factories/offices in Bangladesh. A study of an Android-based mHealth system in South Africa showed that the system was more cost-effective than pen and paper alternatives [38]. This finding matches our experience in the present study. An intervention program in India found that mobile phone messaging (e.g., short messaging service) was an effective and acceptable method for the delivery of advice and support for lifestyle modification to prevent type 2 diabetes in men at high risk [39]. The results of our study suggest that literate subjects are at high risk of NCDs based on high BMI, blood pressure, blood glucose, and urine glucose results, and a mobile phone messaging system would be an effective approach to improving their health.

18.3.3.2 Results of the Health Checkup

We found a high rate of obesity based on a high waist/hip ratio (metabolic syndrome) and a high rate of hypertension at the first health checkup. A high carbohydrate and oil-rich diet may contribute to obesity in Bangladesh. In addition, the use of salt to preserve food where refrigeration is not widely available may increase the risk of hypertension. Conversely, despite the high prevalence of obesity, diabetes prevalence was not high, probably a result of high exercise levels. Many subjects were graded yellow based on the urine protein test. Chowdhury et al. [40] have reported widespread arsenic contamination of drinking water in Bangladesh. We are currently investigating whether pollution with arsenic and heavy metals, such as cadmium, affects urine tests directly or causes kidney dysfunction.

At the first checkup, 472 of 16,741 subjects were graded red. This is potentially an important outcome of the study because we were able to initiate intervention for these high-risk subjects with health care guidance, teleconsultation, and encouragement to visit a clinic.

At the second checkup, there was a significant decrease in SBP for all color rankings ($P < 0.001$), even if the subjects were graded as green or yellow at the first checkup. This result indicates that the health of the subjects improved even with knowledge of the initial result and basic health guidance without intervention by a doctor.

18.3.3.3 Cost Evaluation

In this study, we performed all the available tests in all subjects. In our estimation, to enable sustainable operation and widespread implementation of the program, we need to reduce the total cost to < US$3 per subject. However, the cost of the blood glucose test is high, at approximately US$0.60 per measurement. We identified effective ways to reduce this cost by estimating the risk for diabetes using predictors and measuring blood glucose only in high-risk subjects.

In designing a predictor system, there has to be a tradeoff between true positive rate (TPR) and false positive rate (FPR) results. For example, we selected a threshold for Youden's index that minimized the balanced error rate (=1-TPR +FPR), and generated a predictor with a TPR of 87.6 % and a FPR of 5.9 %. That result indicates that we can skip 14,344 (15,243×94.1 %) unnecessary tests if we accept 57 (462×12.4%) oversights. The predictor system would reduce the number of blood glucose tests from 15,705 to 1304; consequently, the measurement cost per subject would be reduced to one-tenth.

We need to design a predictor that maximizes TPR under existing budget constraints to manage the health of a large group with acceptable FPRs. To be more precise, if we arrange the risk value in the descending order first and count the number by the limits of inspection, then we can choose the risk value as the threshold.

The cost of teleconsultation is also high because of the high salary paid to doctors. Reducing the workload of doctors reduces the cost of the medical program. We are currently analyzing eHealth records using the association rule to support the clinical decisions of medical staff [41]. This analysis can help doctors to add prescription data into the system faster because the system predicts what the doctors want to do and can show candidate inputs and instructions. Machine learning techniques could substitute formulaic, insignificant, and cumbersome work of doctors, enabling them to concentrate on more specific and important issues of patients.

In this study, we provided a health guidance booklet for subjects and a staff member explained the checkup results and health care guidelines orally for illiterate subjects. Because cost control is a serious concern of this project, we plan to make educational videos for health guidance and screen them on devices such as tablet

PCs, at the checkup sites. The videos could be useful not only for illiterate subjects but also for literate ones to increase their health awareness.

18.3.3.4 Limitations

Because we selected sites around Dhaka in the first year due to a logistic problem, we assessed the 1-year after-effects of the program only around Dhaka. However, because the first checkup results around Dhaka are similar to those of the second year site, except for the urine protein test, we consider that the program has a similar effect even other areas.

18.3.4 Conclusions

The present study findings suggest that our eHealth system, combining a health checkup and teleconsultation via the mobile network, is an effective tool in the social health care system in developing countries. It also suggests that the stratification rule is working effectively.

18.4 Future Directions

This paper presented two types of medical sensor data analysis—Hospital study and Bangladesh Study. In the hospital study, we gathered and analyzed nursing activity data. The future work includes expanding the data mining in order to explore the knowledge about clinical paths, such as finding important activities that lead to earlier discharge from the hospital.

In the Bangladesh study, we developed an eHealth system that comprises a set of sensor devices in an attache case. We plan to continue large-scale research into the results of our program, evaluating long-term outcomes to better assess the quality of the service. We will investigate changes in mortality and the frequency of clinic and hospital visits as well as changes in the basic health level and the total costs involved.

References

1. Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. BMJ 312(7023):71–72
2. Oberg PA, Togawa T, Spelman FA (eds) (2006) Sensors applications, sensors in medicine and health care, vol 3. Wiley, New York

3. Ko J, Lu C, Srivastava MB, Stankovic JA, Terzis A, Welsh M (2010) Wireless sensor networks for healthcare. Proc IEEE 98(11):1947–1960
4. Panella M, Marchisio S, Di Stanislao F (2003) Reducing clinical variations with clinical pathways: do pathways work? Int J Qual Health Care 15:509–521
5. Rotter T, Kinsman L, James E, Machotta A, Gothe H, Willis J, Snow P, Kugler J (2010) Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database Syst Rev CD006632. http://www.ncbi.nlm.nih.gov/pubmed/20238347
6. Ward JA, Lukowicz P, Tröster G, Starner TE (2006) Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans Pattern Anal Mach Intell 28:1553–1566
7. Lane ND, Miluzzo E, Lu H, Peebles D, Choudhury T, Campbell AT (2010) A survey of mobile phone sensing. IEEE Commun Mag 48:140–150
8. Roggen D, Troster G, Lukowicz P, Ferscha A, Millan JDR, Chavarriaga R (2013) Opportunistic human activity and context recognition. Computer 46:36–45. http://www.computer.org/csdl/mags/co/2013/02/mco2013020036-abs.html
9. Naya F, Ohmura R, Takayanagi F, Noma H, Kogure K (2006) Workers' routine activity recognition using body movements and location information. In: 2006 10th IEEE international symposium on wearable computers, pp 105–108. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4067734
10. Tentori M, Favela J (2008) Monitoring behavioral patterns in hospitals through activity-aware computing. In: 2008 Second international conference on pervasive computing technologies for healthcare. IEEE, New York, pp 173–176. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4571062
11. Osmani V, Balasubramaniam S, Botvich D (2008) Human activity recognition in pervasive health-care: supporting efficient remote collaboration. J Netw Comput Appl 31(4):628–655. http://linkinghub.elsevier.com/retrieve/pii/S1084804507000719
12. Inoue S, Ueda N, Nohara Y, Nakashima N (2015) Mobile activity recognition for a whole day: recognizing real nursing activities with big dataset. In: ACM international conference on pervasive and ubiquitous computing (Ubicomp), Osaka
13. Inoue S, Ueda N, Nohara Y, Nakashima N (2015) Understanding nursing activities with long-term mobile activity recognition with big dataset. In: The 47th ISCIE international symposium on stochastic systems theory and its applications (SSS), Hawaii, p 10
14. Nohara Y, Sozo I, Nakashima N, Naonori U, Kitsuregawa M (2012) Large-scale sensor dataset in a hospital. In: International workshop on pattern recognition for healthcare analytics, Tsukuba, Japan, p 4. http://sozolab.jp/publications/176
15. Bao L, Intille SS (2004) Pervasive computing. Lecture notes in computer science, vol 3001. Springer, Berlin. http://www.springerlink.com/content/9aqflyk4f47khyjdhttp://link.springer.com/10.1007/b96922
16. Zhang M, Sawchuk A (2012) Motion primitive-based human activity recognition using a bag-of-features approach. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium (1), pp 631. http://dl.acm.org/citation.cfm?doid=2110363.2110433http://dl.acm.org/citation.cfm?id=2110433
17. Zhang M, Sawchuk AA (2011) A feature selection-based framework for human activity recognition using wearable multimodal sensors. In: International Conference on Body Area Networks, pp. 92–98. http://dl.acm.org/citation.cfm?id=2318776.2318798
18. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
19. Hattori Y, Inoue S, Hirakawa G (2011) A large scale gathering system for activity data with mobile sensors. In: 2011 15th annual international symposium on wearable computers, pp 97–100

20. Kawaguchi N, Ogawa N, Iwasaki Y (2011) HASC challenge: gathering large scale human activity corpus for the real-world activity understandings. In: Proceedings of the 2nd augmented human international conference, p 27. http://dl.acm.org/citation.cfm?id=1959853
21. Chavarriaga R, Sagha H, Calatroni A, Digumarti ST, Troster G, Millan JDR, Roggen D (2013) The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recogn Lett 34:2033–2042
22. Roggen D, Calatroni A, Rossi M, Holleczek T, Forster K, Troster G, Lukowicz P, Bannach D, Pirkl G, Ferscha A, Doppler J, Holzmann C, Kurz M, Holl G, Chavarriaga R, Sagha H, Bayati H, Creatura M, Millan JdR (2010) Collecting complex activity datasets in highly rich networked sensor environments. In: Seventh international conference on networked sensing systems (INSS)
23. Ala A (2011) Global status report on noncommunicable diseases 2010. World Health Organization, Geneva. ISBN: 978-92-4-156422-9
24. World Health Organization (2012) Non-communicable diseases - a major health challenge of the 21st century. World Health Statistics 2012. World Health Organization, Geneva, pp 34–37
25. World Health Organization (2011) mHealth - new horizons for health through mobile technologies: based on the findings of the second global survey on eHealth. Global observatory for eHealth series, vol 3. World Health Organization, Geneva. ISBN: 978-92-4-156425-0
26. Nohara Y, Kai E, Ghosh PP, 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 (2015) Health checkup and telemedical intervention program for preventive medicine in developing countries: verification study. J Med Internet Res 17(1): e2. http://www.jmir.org/2015/1/e2, PMID: 25630348
27. Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A, Musgrove P (eds) (2006) Disease control priorities in developing countries, 2nd ed. World Bank and Oxford University Press, Washington, DC. ISBN:10: 0-8213-6179-1
28. IEEE STANDARD ASSOCIATION (2012) IEEE standard for local and metropolitan area networks - part 15.6: wireless body area networks, IEEE Std 802.15.6-2012. ISBN: 978-0-7381-7206-4
29. Metabolic syndrome criteria of International Diabetic Federation (2006) http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf. Archived at http://www.webcitation.org/5a1oMi3ZC
30. Waist circumference and waist-hip ratio: report of a WHO expert consultation (2008) http://whqlibdoc.who.int/publications/2011/9789241501491_eng.pdf. Archived at http://www.webcitation.org/6EbZ4xEh2
31. National Institutes of Health, National Heart, Lung, and Blood Institute (1998) Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults; the evidence report. Obes Res 6(suppl 2):51S–209S.
32. Cifkova R, Erdine S, Fagard R, Farsang C, Heagerty AM, Kiowski W, Kjeldsen S, Lüscher T, Mallion JM, Mancia G, Poulter N, Rahn KH, Rodicio JL, Ruilope LM, van Zwieten P, Waeber B, Williams B, Zanchetti A, ESH/ESC Hypertension Guidelines Committee (2003) Practice guidelines for primary care physicians: 2003 ESH/ESC hypertension guidelines. J Hypertens 21(10):1779–1786 PMID:14508180
33. Global guideline for type 2 diabetes of the International Diabetes Federation (2005) http://www.idf.org/webdata/docs/IDF%20GGT2D.pdf. Archived at http://www.webcitation.org/6KfOmZlWs
34. Ministry of Health, Labor and Welfare of Japan (2012) National health and nutrition survey in Japan (written in Japanese). http://www.mhlw.go.jp/bunya/kenkou/eiyou/dl/h24-houkoku.pdf, Archived at http://www.webcitation.org/6R4ERJcJB
35. Balsam J, Ossandon M, Bruck HA, Lubensky I, Rasooly A (2013) Low-cost technologies for medical diagnostics in low-resource settings. Expert Opin Med Diagn 7(3):243–255. PMID:23480559

36. Mohan V, Deepa M, Pradeepa R, Prathiba V, Datta M, Sethuraman R, Rakesh H, Sucharita Y, Webster P, Allender S, Kapur A, Anjana RM (2012) Prevention of diabetes in rural India with a telemedicine intervention. J Diabetes Sci Technol 6(6):1355–1364. PMID:23294780
37. Ramachandran A, Snehalatha C, Mary S, Mukesh B, Bhaskar AD, Vijay V, Indian Diabetes Prevention Programme (IDPP) (2006) The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). Diabetologia 49(2):289–297. PMID:16391903
38. Rajput ZA, Mbugua S, Amadi D, Chepngeno V, Saleem JJ, Anokwa Y, Hartung C, Borriello G, Mamlin BW, Ndege SK, Were MC (2012). Evaluation of an Android-based mHealth system for population surveillance in developing countries. J Am Med Inform Assoc 19(4):655–659. PMID:22366295
39. Ramachandran A, Snehalatha C, Ram J, Selvam S, Simon M, Nanditha A, Shetty AS, Godsland IF, Chaturvedi N, Majeed A, Oliver N, Toumazou C, Alberti KG, Johnston DG (2013) Effectiveness of mobile phone messaging in prevention of type 2 diabetes by lifestyle modification in men in India: a prospective, parallel-group, randomized controlled trial. Lancet Diabetes Endocrinol 1(3):191–198. PMID:24622367
40. Chowdhury UK, Biswas BK, Chowdhury TR, Samanta G, Mandal BK, Basu GC, Chanda CR, Lodh D, Saha KC, Mukherjee SK, Roy S, Kabir S, Quamruzzaman Q, Chakraborti D (2000) Groundwater arsenic contamination in Bangladesh and West Bengal, India. Environ Health Perspect 108(5):393–397. PMID:10811564
41. Kai E, Rebeiro-Hargrave A, Inoue S, Nohara Y, Islam R, Nakashima N, Ahmed A (2014) Empowering the healthcare worker using the portable health clinic. In: Proceedings of 28th IEEE international conference on advanced information networking and applications (AINA2014), May 13–16, Victoria, Canada. IEEE, New York

Index

C.-M. Kyung et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-33201-7