Expert Systems with Applications 42 (2015) 4540–4555

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel multimodal communication framework using robot partner for aging population

Dalai Tang [a], Bakhtiar Yusuf [a], János Botzheim [a,b,*], Naoyuki Kubota [a], Chee Seng Chan [c]

[a] Graduate School of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
[b] Department of Automation, Széchenyi István University, 1 Egyetem tér, Győr 9026, Hungary
[c] Centre of Image and Signal Processing, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Available online 19 January 2015

Keywords:
Human–robot interaction
Robot partner
Intelligent robot
Aging
Gesture recognition

ABSTRACT

In developed country such as Japan, aging has become a serious issue, as there is a disproportionate increasing of elderly population who are no longer able to look after themselves. In order to tackle this issue, we introduce human-friendly robot partner to support the elderly people in their daily life. However, to realize this, it is essential for the robot partner to be able to have a natural communication with the human. This paper proposes a new communication framework between the human and robot partner based on relevance theory as the basis knowledge. The relevance theory is implemented to build mutual cognitive environment between the human and the robot partner, namely as the informationally structured space (ISS). Inside the ISS, robot partner employs both verbal as well as non-verbal communication to understand human. For the verbal communication, Rasmussen's behavior model is implemented as the basis for the conversational system. While for the non-verbal communication, environmental and human state data along with gesture recognition are utilized. These data are used as the perceptual input to compute the robot partner's emotion. Experimental results have shown the effectiveness of our proposed communication framework in establishing natural communication between the human and the robot partner.

1. Introduction

From the year of 2000 to 2050, it is expected that the proportion of elderly people (60 years old or older people) in the world's population will double from 11% to 22%, and the absolute number of elderly people is expected to increase from 605 million to 2 billion over the same period.[1] Meanwhile, in Japan, according to the Statistics Bureau at the Ministry of Internal Affairs and Communication,[2] the population of elderly people is expected to increase to 36 million, that is about 31% of the population in the year of 2030. In Tokyo itself, it is anticipated that the number of elderly people will reach 25.2% of the population in year 2015.

Along with the increasing number of elderly people, one must note that the number of those elderly people who are no longer able to look after themselves will also increase proportionally. Many of them will lose the ability to live independently because of limited mobility, frailty or other physical or mental health problems (Chernbumroong, Cang, Atkins, & Yu, 2013; Rueangsirarak, Atkins, Sharp, Chakpitak, & Meksamoot, 2012). In Japan, the increasing number of elderly people who live alone or independently has required a large number of nursing care to support them. However, since the number of caregivers is always limited, it is important to introduce alternative solution to tackle this problem. One of the solutions is the introduction of the human-friendly robot partner to support the elderly people in their daily life.

According to Broekens, Heerink, and Rosendal (2009), there are two types of robots that are able to support the elderly people. One is the rehabilitation robot and the other is the social robot. In the former, the robot focuses on physical assistance technology, whilst the latter is concerned as a system that has the capability in human–robot communication. It is also known as the robot partner. This paper focuses on the latter with a focus on realizing a natural communication between the human and the robot partner. The natural communication can be realized when robot can understand human intention or thought. We implemented the theory of relevance (Sperber & Wilson, 1995) to build mutual cognitive environment between human and robot partner into our system, which

* Corresponding author at: Graduate School of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan. Tel.: +81 42 585 8441.
E-mail addresses: tang@ed.tmu.ac.jp (D. Tang), yusuf-bakhtiar1@ed.tmu.ac.jp (B. Yusuf), botzheim@tmu.ac.jp, botzheim@sze.hu (J. Botzheim), kubota@tmu.ac.jp (N. Kubota), cs.chan@um.edu.my (C.S. Chan).

[1] See www.who.int/ageing/en/.
[2] See www.stat.go.jp/english/data/handbook/c0117.htm.

http://dx.doi.org/10.1016/j.eswa.2015.01.016
0957-4174/

called informationally structured space (ISS) to handle this problem. According to Sperber and Wilson (1995), relevance theory is very useful to discuss the multimodal communication, where each person has his or her own cognitive environment that make their communication restricted. Therefore, usually humans use their utterances or gestures to expand their cognitive environment by extracting person's attention into specific target object, event, or person. When human's cognitive environment became wider, they can share each other intention or thought as illustrated in Fig. 1. The implementation of this theory into our system can be observed in the structure of database in our system, which is explained more detailed in Section 6. Meanwhile in conducting communication between human and robot, we use the Rasmussen's behavior model to build the conversation system. In addition to verbal communication, we also implement non-verbal communication such as facial expression, emotional gestures and pointing gestures.

Our contribution of this paper is to treat all these elements (environment recognition, human recognition and emotional model) as an unified framework in the informationally structured space, so that a more natural communication between a robot partner and a person can be formed. In order to facilitate this, we built a new type of robot partner, named "iPhonoid". Experiments using three different case studies have shown the effectiveness of the proposed framework in establishing natural communication between the human and the robot.

The rest of the paper is organized as follows. Section 2 discusses the literature related to the proposed system. Here, we also explain the advantage and disadvantage of the proposed method compared to previous researches. Section 3 introduces the concept of informationally structured space. Section 4 deals with the environmental system, which includes sensor network, web system and robot system, while robot system is explained separately in the following section. Section 5 explains the robot partner including gesture recognition technique and emotional model. Section 6 discusses the database system and communication system. Section 7 details the conversation system. Section 8 presents experimental results of the proposed method. Section 9 summarizes and discusses the future direction of this research.

2. Literature review

In the proposed method, we implemented relevance theory to build mutual cognitive environment called informationally structured space; Rasmussen's behavior model for conversational system; emotional model and gesture recognition to realize natural communication between human and robot partner. In this section we discuss and compare our proposed method to previous researches. Since mutual cognitive environment has a close relationship with ambient intelligence, we will start to review previous researches on this field. Next, we will discuss emotional model, thereafter gesture recognition and finally conversation system.

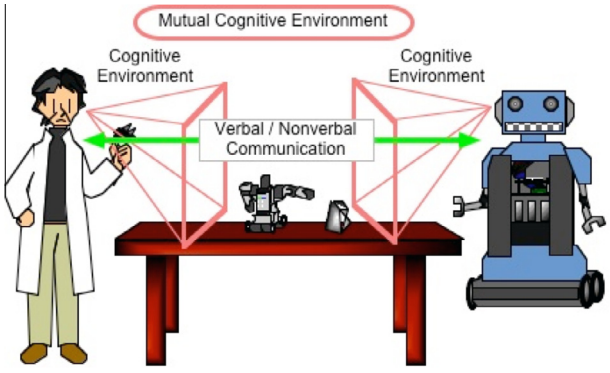


Fig. 1. Mutual cognitive environments via natural communication. In order to realize a natural communication with a person, the robot partner acquires information on the surrounding environment, as well the people's condition by processing the data obtained from various sensors. When the robot partner is able to build a mutual cognitive environment, the robot partner can understand the people's intention or thought.

In works related to ambient intelligence Montes, Ortega, Venzala, and Abril (2014) built smart environment based on software reference architecture. The smart environment is used to conduct the perception process in a standard office. However, the paper only used motion detection in order to measure data from sensor. Details such as gestures were not included. On the other hand, Lee, Lee, Kim, Wang, and Love (2014) proposed a method called mixed context-aware inference, which is a novel sensor-based context-aware system focusing on three inference processes: rule, inference and pattern driven. Forkan, Khalil, and Tari (2014) used various sensors to measure data, which enabled this method to get more accurate result. Moreover, the usage of cloud technology made the process time become shorter. However, this method is difficult to realize concerning the high cost. Furthermore, since user has to wear special clothes to get data from the sensors, it is very cumbersome. Forkan, Khalil, Tari, Foufou, and Bouras (2015) proposed fusion-based architecture, detection in activity and location patterns using Hidden Markov Model (HMM). Although it has a good accuracy in computation, HMM has a disadvantage in high computation cost.

For the emotional model concept, Tay, Jung, and Park (2014) proposed a method, which effects on occupational roles (security vs. health-care), gender (male vs. female), and personality (extrovert vs. introvert) on user acceptance of a social robot. However, they only use stereotype to conduct the evaluation, which makes the result arguable. Daosodsai and Maneewarn (2013) proposed a method to generate emotion of a robot using expert knowledge by fuzzy logic. The emotion of the robot is determined using 3 types of input data, such as the robot's personality, the ambient environment, and the interaction with human. However, since emotional expression is using LED only, the emotional expression done by the robot has a lot of limitations and it is difficult to be evaluated. Emotional model proposed by Kim, Yang, and Kwon (2013) used episodic memory system, as a result of long term human robot interaction and emotion generation reaction. Although the method of this paper is very interesting, the application is only possible in the virtual environment. Jitviriya and Hayashi (2014) used the integration of environment, robot self-states and feedback behavior for generating robot emotion (human data is not included). As a conclusion in the previous research the definition of the robot's emotion is not clear, while in our paper the change in environment, human state (gesture and distance) effects the robot's emotion, which linked into the conversation system content and robot's facial expression.

For the works related to gesture recognition, Iengo, Rossi, Staffa, and Finzi (2014) proposed a novel approach to real-time and continuous gesture recognition for flexible, natural, and robust human–robot interaction (HRI), and the generation of an ad hoc HMM. As mentioned before, one disadvantage of HMM is the high computational cost. Gesture recognition method proposed in Xiao, Yuan, and Thalmann (2013) is based on combination of the CyberGlove and Kinect sensor, which could recognize various gestures. The using of the CyberGlove for measurement device shows that this method cannot be directly realized in daily life now owing to its price. Liu, Hu, Luo, and Wu (2014) proposed a method in gesture recognition by using on-board monocular camera and specialized gesture detection algorithms. Here, also the dynamic movement primitives (DMP) model is employed. In this method, since the depth information is not acquired, the recognition has many limitations. The gesture recognition in our system used Kinect sensor has lower the cost, although our method is not as

accurate as the previous research. In addition, since our robot partner uses iPhone as a mainframe to speed up the computational cost, the growing neural gas algorithm is adopted to extract features of sensing data beforehand and then spiking neural network to recognize the gesture. A deep explanation of gesture recognition can be found in Section 5.2.

In the conversation system related research, Nesselrath (2013) proposed a dialogue system framework architecture that supports cognitive load prediction and situation-dependent decision making and manipulation of the HCI. This paper also proposed the multimodal fusion and fission which shows the system of how to interact with human and learning. On the other hand, Liu, Pasupat, Cyphers, and Glass (2013) proposed multilingual dialogue systems and seamless deployment to mobile platforms. English and Mandarin systems in various domains (e.g. movie, flight and restaurant) are implemented with the proposed framework. However, since the dialogue system works in online server (connected into Internet), it cannot work in offline state. Lopes, Eskenazi, and Trancoso (2015) presented the used of data-driven approach to improve Spoken Dialog System (SDS) performance by automatically finding the most appropriate terms to be used in system prompts. On the other hand, the conversation system of our method can be conducted in the server or in the robot partner. The conversation system in our method is built as the result of the connectivity between the environment and the robot individual intelligence. Basically, the conversation system contents is stored in the ISS server. However, minimum conversation system is also stored in the robot partner used as input–output interface. When the robot starts to communicate with the human, the robot will conduct the learning process based on the conversation contents time, human state, and environment state.

Similar systems to our proposed system were developed previously. For example, Takemura and Ishii (2011) explained the recognition of environment by the usage of color property; Shahdi and Bakar (2012) implemented face recognition technology and Botzheim, Obo, and Kubota (2012) employed the gesture recognition technology to perform human recognition. Moreover, Böhme et al. (2003) used person localization based on the face recognition and skin color detection (Tan, Chan, Yogarajah & Condell, 2012) to develop their robot. Since communication also involves the perception of intention and feeling, human emotion plays an important role in the act of communication, which leads to an action. Bien and Lee (2007) besides consider human gesture as the information for the robot, the emotional information such as facial expression and voice tones are also utilized to determine the robot action. Similar approaches can be found in Zhang, Jiang, Farid, and Hossain (2013). On another variant, according to Botzheim,Tang, Yusuf, Obo, & Kubota et al. (2013), a robot partner with emotional model can give meaning and value to the perceptual information, which leads to a decision based on internal and external state. That is when a person is in the state of sadness, his action will show the state of sadness. Therefore, it is also important to implement emotion into the robot partner. The comparison of these methods with our method can be found in Section 8.4 in the discussion of the experiments.

3. Informationally structured space

3.1. Human–computer interaction

Recently, ubiquitous computing (Cerpa et al., 2001; Chan, Liu, & Brown, 2007; Ingelrest et al., 2010; Mainwaring, Culler, Polastre, Szewczyk, & Anderson, 2002; Preuveneers et al., 2004) has become one of the main attentions in the development of information technology. Ubiquitous computing can be defined as the opposite of virtual reality. While virtual reality puts people inside a computer-generated world, ubiquitous computing forces the computer to integrate the world with people (Jeng, 2009; Johanson, Fox, & Winograd, 2002; Lim, Tang, & Chan, 2014; Römer, Schoch, Mattern, & Dübendorfer, 2004). This technology can also be described as pervasive computing and ambient intelligence. Ambient intelligence is an emerging discipline that brings intelligence to our everyday environments and makes those environments sensitive to us (Cook, Augusto, & Jakkula, 2009). The concept of sensor network and ubiquitous computing integrated into robotics can be called as network robotic and ubiquitous robotic (Kim, Kim, & Lee, 2004; Kubota & Nishida, 2006). The network robotic is basically divided into three parts: visible robots, unconscious robots, and virtual robots (Kubota, 2008). The visible robots use their body to act with human. The unconscious robots are used to acquire environmental data and the existence along with human is invisible. The virtual robot points out an agent or a software package in the cyber world. Based on these, we can make a conclusion that a robot can be used not only as a human-friendly life-support system, but it can also become an interface connecting the physical world with the cyber world (Costa, Castillo, Novais, Caballero, & Simoes, 2012; Kubota, 2005; Kubota, Nojima, Kojima, & Fukuda, 2006).

3.2. Informationally structured space

However, since the common problems in these researches are the distributed measurement and computing, the cooperative and distributed measurement used for realizing communication between a robot and a human has not been discussed frequently. In order to realize a natural communication between a robot partner and a human, the environment surrounds the human and the robot should have a structured platform to gather, store, transform, and provide information easily. This kind of environment is called the informationally structured space (Kubota & Yorita, 2009). Informationally structured space (ISS) basically has the following properties; (1) Generality and Shareability of Information, (2) Reversibility of Information, and (3) Human Understandability of Information (Kubota, Tang, Obo, & Wakisaka, 2010).

Information gathered in the environment is transformed by the robot partner and the sensor network device into a qualitative information to be uploaded to ISS using its own rule and reversibly can transform the information downloaded from ISS to measurement data by its own rule. This sharing information process within the environment, realizes natural communication between a human and a robot. Fig. 2 illustrates the concept of informationally structured space. In order to understand ISS deeply, we will discuss the ISS concerning life hub and robot partner in detail.

3.3. Informationally structured space for life hub

The word "life hub" is the extended concept of "digital hub" explained by the late Steve Jobs. He explained that Macintosh in a short time could serve as the Digital Hub that unites those disparate points in our digital life (January 9, 2001). In Life Hub, we unite people with physical and virtual information including (1) personal information, (2) environmental information, (3) Internet information, (4) people, (5) place, (6) goods, and (7) events, in addition to real world. These information will be structured and recorded in the database, which can be accessed by the robot partner.

Fig. 3 illustrates the ISS when gathering personal information to produce daily life log. This figure shows different levels of information supports, such as personal information, indoor life log, and outdoor activity log. Personal information can be gathered by a smart phone. Indoor life log can be created by sensor network. The next level is when the human activity log considers also out-

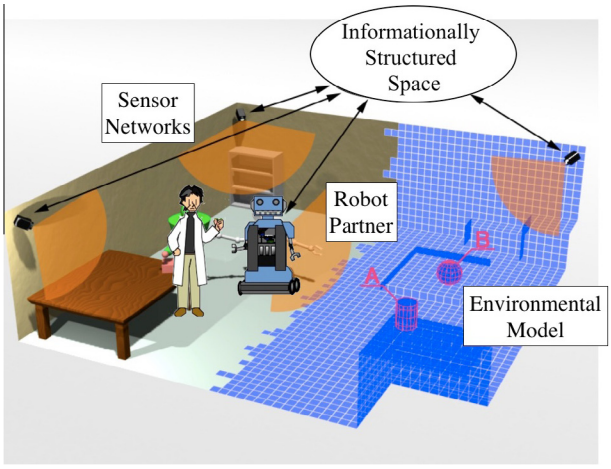


Fig. 2. Informationally structured space.

door information. As shown in Fig. 3 a rechargeable IC card is used to trace the shopping and traveling activities of people.

The gathered information can be used on different levels as well. The first level is the information support for family, or in the case of elderly people for the caregivers. The other level is social network in the sense of a local community on the Internet. The last level is the complete information support using Internet.

3.4. Informationally structured space for robot partners

Concerning a robot partner, we divide the robot partner modules into three parts, such as the perceptual modules, the decision making modules, and the action modules. In order to perform natural communication, the robot partner with perceptual modules is able to understand the people's behavior and the surrounding environmental condition. In this paper, we use a Sensor Network Module as the perceptual modules.

While the robot partner can get the people's state and the surrounding information by its perceptual module, the robot partner also needs to have decision-making modules to support people. The decision making module used in this paper is a basic communication module, which supports the daily conversation. Finally, in order to perform multi-modal communication, an action module is utilized. For the action module, we propose several sub-modules such as utterance module, gesture expression module, and emotional module. The utterance module performs actions such as planning conversation and speech synthesizing for conducting utterance. In the gesture expression modules, we defined many common gestures for different robot partners. Therefore, even for a different robot partner, the gesture expression is the same. For enhancing the robot partner, an emotional module is applied in the robot partner. Each of the discussed modules will be detailed in the next sections.

4. Environmental system

In this paper, we apply the concept of ISS to build the proposed system as shown in Fig. 4. This system is divided into an environmental system and a database. The environmental system is composed of a sensor network system, a web system, and a robot system. For measuring the environment and the human condition information, we apply the sensor network system. News or weather reports can be extracted using the web system, and the robot system is utilized for conducting directional interaction with the human while saving the human communication history into the database. The data processing work flow is as follows. First, the sensor network as an information collecting module gets all information required including environmental and human condition information periodically. These information is stored into the database server as a perceptual input for the emotional model. The emotional model processes all information to realize a robot partner which emotionally interacts with human. The output of the emotional model is sent to the robot partner as a signal to be converted into conversation contents, gestural and facial expression. The human reaction as the result of the robot partner's action is used again to update the robot partner's next action through the sensor network system, database and emotional model. In the next subsection, we will discuss the sensor network and web system. Whilst robot partner will be discussed in the next section.

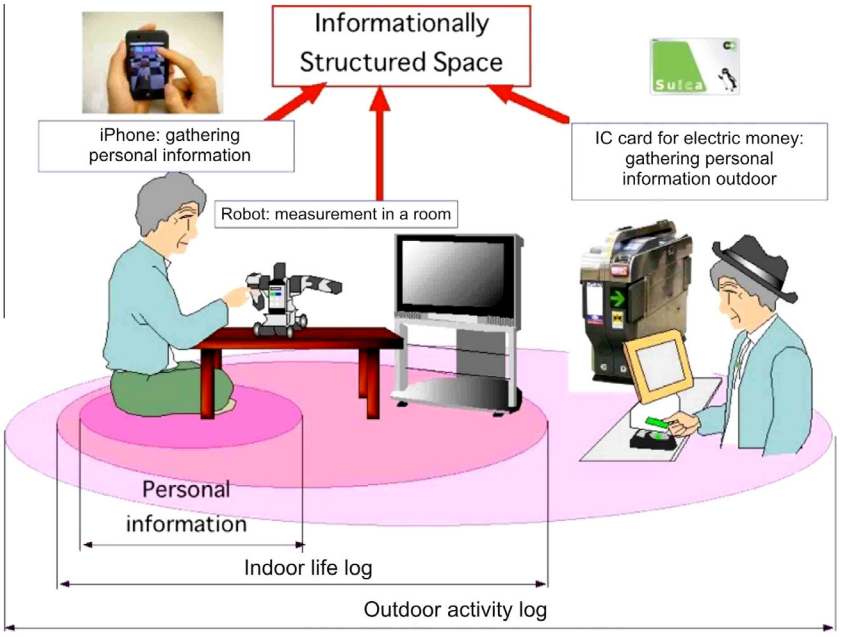


Fig. 3. Data gathering in informationally structured space.

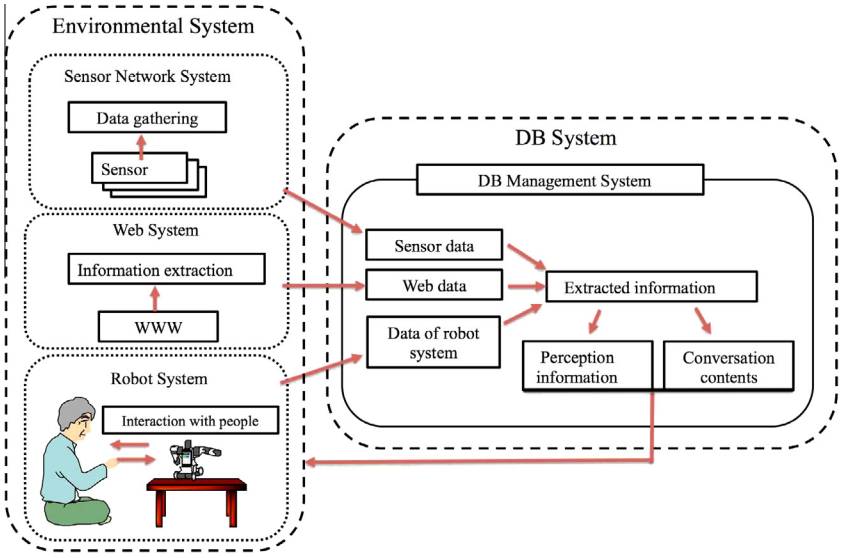
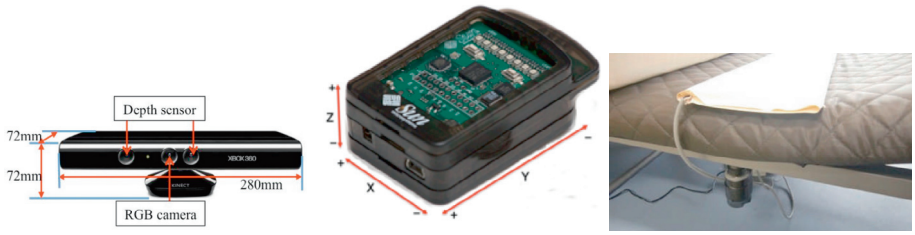


Fig. 4. The structure of our proposed system. It consists of an environmental system and a DB system.



(a) Microsoft Kinect (b) Sun SPOT (c) Wireless Optical Oscillo-sensor

Fig. 5. Sensor network system. Our sensor network system consists of a Kinect, Sun Spot and an optical oscillo-sensor.

4.1. Sensor network system

According to the reliability of recognition technologies for establishing natural communication, the range of communication is separated into three types: short range (∼ 700 [mm]), middle range (700 ∼ 2,000 [mm]), and long range (2,000∼ [mm]). For the short and middle range, the touch interface and Microsoft Kinect are used to gather the required data for the robot partner.

As an important part in gaining human perception, the visual system supports the cognitive associated process (Alexandre & Tavares, 2010; Wu & Hsu, 2011). In this paper, the visual system of the robot partner is used to measure not only environmental data, but also human state data such as gesture recognition measurement. In order to conduct these tasks, the robot partner's equipped sensors are not enough to collect the required information. Therefore, we use Microsoft Kinect[3] to perform these tasks as depicted in Fig. 5(a). With the technical specifications shown in Table 1, Microsoft Kinect is applied to extract information presented in Table 2. For measuring environment information, we use the RGB data of the environment extracted by Kinect to calculate brightness (br) and darkness (d) as perceptual inputs for the emotional model using Eqs. (1) and (2), respectively.

$$br = \frac{\sqrt{r^2 \times 0.241 + g^2 \times 0.691 + b^2 \times 0.068}}{350} \quad (1)$$

$$d = \frac{255 - \sqrt{r^2 \times 0.241 + g^2 \times 0.691 + b^2 \times 0.068}}{350} \quad (2)$$

[3] See www.xbox.com.

Table 1
Specification of Microsoft Kinect.

Size	$282 \times 72 \times 72$ [mm]
Horizontal field of view	57 [deg]
Vertical field of view	43 [deg]
Physical tilt range	±27 [deg]
Measuring range	1.2–3.5 [m]
Resolution	$320 \times 240, 640 \times 480$ [pixel]
Frame rate	30 [fps]

Table 2
Extracted data from Microsoft Kinect.

Sensory data	Input data for emotion
RGB data	Brightness
	Darkness
Skeleton data	Distance data
	Human activity
	Human gesture
	Human detection
	Number of people

For the distance data, the z axis of Microsoft Kinect is used for calculation. The human activity is calculated by counting the difference of human position in x and y axis at time t and $t-1$. Human gesture computation will be explained in the following section. Human detection will determine the human existence in the room, while the number of people shows the number of people in the room. These data are used as the input data for perception.

Table 3
Specification of SunSPOT.

Size	41 × 23 × 70 [mm]
Weight	54 [g]
3-Axis accelerometer range	2G/3G
Light sensor range	0–750 [raw reading from 1x]
Battery	720 [mAh] lithium-ion battery
OS	Squawk VM
Wireless radio	2.4 GHz, IEEE 802.15.4

Meanwhile, for the long range, SUN Spot and Wireless optical oscillo-sensor are employed. The Sun SPOT (Sun Small Programmable Object Technology) is a small and battery-powered wireless sensor network (WSN) developed by Oracle Corporation (Sun Microsystems).[4] As shown in Fig. 5(b) and Table 3, the Sun SPOT is built by the IEEE 802.15.4 standard, which can be used for a wide range of applications including robotics, environmental monitoring, asset tracking, proactive health care and many others. The Sun SPOT is also powered by a specially designed small-footprint Java virtual machine called Squawk, which can host multiple applications concurrently requiring no underlying operating system. The wireless optical oscillo-sensor is a sensor used for estimating human states on a bed developed by NEW SENSOR Incorporated.[5] The sensor composed of a pneumatic sensor and an ultrasonic sensor as displayed in Fig. 5(c).

4.2. Web system

As the widespread of tablet-PC and smartphone enforces the development of communication technology, people easily use cloud technology anywhere and anytime. Moreover, using twitter and facebook as social media, people around the world can easily share their opinion, feeling and information by connecting to the Internet. The latest news around the world even some gossips about movie stars can be acquired using the RSS (Rich Site Summary) service. Additionally, we can even know which and where the best restaurant around us is, using Google Web API or Yahoo Web API.

However, for the elderly people or people with cognitive decline, information provided by computer or smart devices cannot be acquired, because for them these new technologies are not accustomed yet. Because of this reason, we develop an information support system for these people, which can extract information from Internet. This system is called web system and shown in Fig. 6. Web system is composed of web information extraction and database, where the extraction information called meta data can be seen in Fig. 7. The web information extraction uses RSS (Rich Site Summary) and yahoo web API for acquiring weather and news information in XML file. This extracted data is stored as contents in database to be reused. Table 4 shows the example of weather news, while Table 5 shows the news example.

5. Robot partner

In the previous researches, we have developed and applied various types of robot partners such as MOBiMac, Hubot, Apri Poco, and Miuro to be used as a support system for elderly people (Kubota, 2005; Kubota & Yorita, 2009). However, since the widespread of smartphone and tablet PC makes their price decreasing along the time, we have been developing a robot partner, which combines smartphone and embedded system into a small, mobile, and economical device. The word "Economical" means that as smartphones are equipped with various sensors like accelerometer, gyro, camera, and microphone, we can decrease the price of the robot partner. In this paper the robot partner will act not only to measure the human condition using touch sensor and voice recognition, but iPhonoid will also process the collected data through sensor network using emotional model to perform a particular action. The architecture of the iPhonoid can be seen in Fig. 8.

[4] See www.sunspotworld.com/index.html.
[5] www.new-sensor.com.

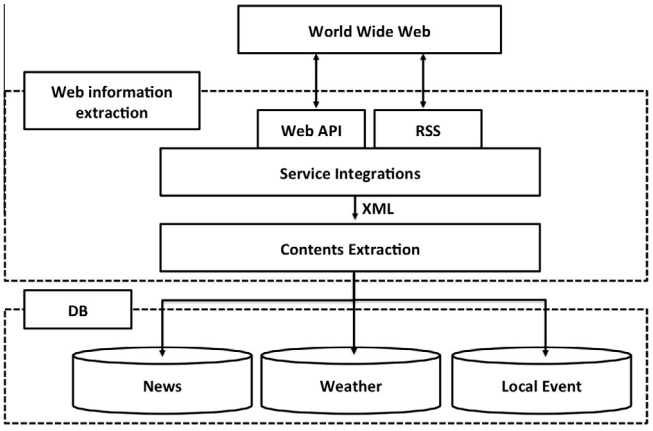


Fig. 6. Web information extraction system structure.

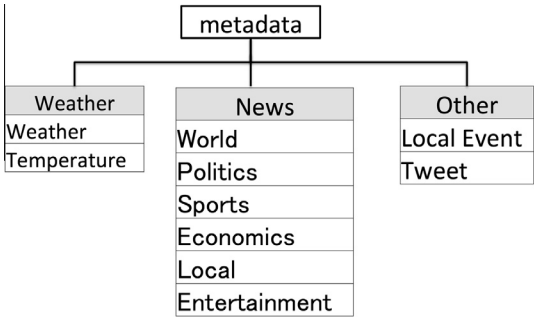


Fig. 7. Metadata of web information.

Table 4
Example of weather news database.

Day	Date	Temp (max)	Temp (min)	Weather
Wednesday	140507	22	16	Sunny
Thursday	140508	24	17	Sunny
Friday	140509	25	19	Sunny and cloudy

Table 5
Example of the latest news database.

No	Topic	Time	Contents
10001	World	140517202514	Laos government's plane crash cause death of vice prime minister
10002	World	140517131908	China Republic requests payment from Vietnam
10003	World	140517153048	CNN fired editor of 50 plagiarism

5.1. Emotional model

We develop emotional models composed by emotion, feeling, and mood measured based on a time scale. This development is done by assuming that emotion change temporally based on the perceptual information on the internal state and the external envi-

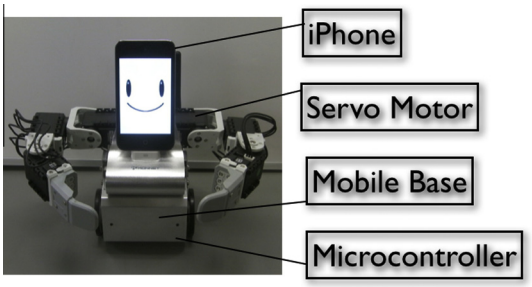


Fig. 8. Our new robot partner, namely the iPhonoid.

ronment (Kubota & Wakisaka, 2010; ; Yorita, Botzheim, & Kubota, 2013). Emotion is used for intermediating the perceptual system and emotional model, while considering it as an intense short-term mental state based on perceptual information. This leads to the assumption of the emotion changing, which depends on specific perceptual information. As a part of the robot partner, smartphone converts independently the gestures and the environmental information as perceptual information to emotional input based on predefined data. Each feeling is updated as the summation of emotions.

The ith emotional input $u_i^E(t)$ is generated based on the $u_{j,k}^I(t)$ perceptual information as follows:

$$u_i^E(t) = w_{i,j,k}^E \cdot u_{j,k}^I(t), \tag{3}$$

where $w_{i,j,k}^E$ is the degree of contribution from the jth gesture and kth environmental data to the ith emotion ($-1 \leqslant w_{i,j,k}^E \leqslant 1$).

Yorita et al. (2013) defined five different feeling models. In this paper, we apply the model where the state of the ith feeling $u_i^F(t)$ is updated by the emotional input from the viewpoint of bottom-up construction and the top-down constraints from mood values are also considered as displayed in Fig. 9:

$$u_i^F(t) = \tanh\left(\kappa u_i^F(t-1) + (1-\kappa)[E + F_i]\right), \tag{4}$$

where

$$E = \sum_{j=1}^{N^E} u_j^E(t-1)$$

$$F_i = \sum_{j=1, j\neq i}^{N^F} w_{i,j}^F \cdot u_j^F(t-1)$$

$$\kappa = \frac{\gamma^F}{1 + u_1^M(t-1) - u_2^M(t-1)}, \tag{5}$$

where γ^F is the temporal discount rate of feelings ($0 < \gamma^F < 1$), N^E is the number of emotional inputs, N^F is the number of feelings, $w_{i,j}^F$ is

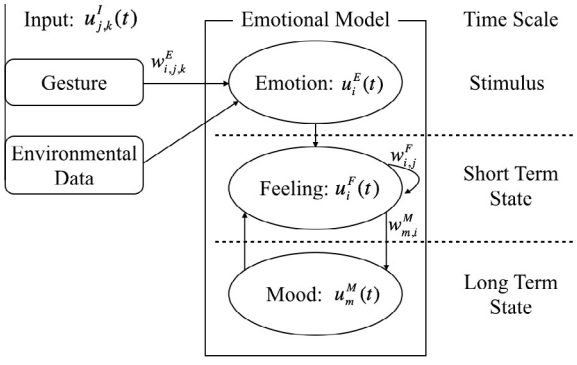


Fig. 9. The structure of the proposed emotional method.

the stimulation or suppression coefficient from the jth feeling to the ith feeling ($-1 \leqslant w_{i,j}^F \leqslant 1$), and $u_m^M(t)$ is the value of the mth mood. We use positive mood ($m = 1$) and negative mood ($m = 2$). The hyperbolic tangent is used to regulate the values of feelings.

Mood is defined as the long-term state updated by a change in feelings, and governs changes in feelings. While feeling is defined as a short-term state updated by a change in emotion. The state of the mth mood is updated by the sum of feelings:

$$u_m^M(t) = \tanh\left[\gamma^M u_m^M(t-1) + (1-\gamma^M)\sum_{i=1}^{N^F} w_{m,i}^M u_i^F(t)\right], \tag{6}$$

where γ^M is the discount rate and $w_{m,i}^M$ is the stimulation or suppression coefficient from the ith feeling to the mth mood ($-1 \leqslant w_{m,i}^M \leq 1$). The structure of the model is shown in Fig. 9. In this figure, we can see how the feeling and mood influence each other and the emotion can be considered as an input impulse to the feeling.

5.2. Gesture recognition

In this paper, the structured learning (SL) is adopted as similarly in Botzheim et al. (2013) and Botzheim and Kubota (2012). SL contains two stages, a topology learning phase and a spatio-temporal learning phase. The growing neural gas (GNG) is applied in the first phase for information extraction, and the spiking neural network (SNN) is applied in the second stage to recognize the gesture.

5.2.1. Growing neural gas for information extraction

Unsupervised learning is performed by using data without any teaching signals (Fritzke, 1992, 1995; Kohonen, 2001; Martinetz & Schulten, 1991). Self-Organizing Map (SOM) (Kohonen, 2001), Neural Gas (NG) (Martinetz & Schulten, 1991), Growing Cell Structures (GCS) (Fritzke, 1992), and Growing Neural Gas (Fritzke, 1995) are some well known unsupervised learning methods that use the competitive learning approach. In SOM, the number of nodes and the topological structure of the network are designed beforehand Kohonen (2001). In NG, the number of nodes is also constant, however its topological structure is updated according to the distribution of sample data (Martinetz & Schulten, 1991). On the other hand, GCS and GNG can dynamically change the topological structure based on the adjacent relation (edge) referring to the ignition frequency of the adjacent node according to the error index. GCS does not delete nodes and edges and it must consist of k-dimensional simplices whereby k is a positive integer chosen in advance. On the other hand, GNG can delete nodes and edges based on the concept of ages (Fritzke, 1995). The initial configuration of each network is a k-dimensional simplex, if $k = 1$ then it is a line, if $k = 2$ then it is a triangle, and if $k = 3$ then it is a tetrahedron (Fritzke, 1992). GCS has been applied to construct 3D surface models by triangulation based on 2-dimensional simplex. However, because the GCS does not delete nodes and edges, the number of nodes and edges is over increasing. Another disadvantage of GCS is that it cannot divide the sample data into several segments. GNG can overcome these drawbacks. When applying GNG, the distance criterion is used for extracting human motions. The GNG algorithm is described in Algorithm 1.

The following notations are used in the learning algorithm of GNG (Fritzke, 1995, 1996): $\mathbf{r}_i$ is the 3-dimensional vector of a node (reference vector, $\mathbf{r}_i \in \mathbb{R}^3$); $\mathbf{v}$ is the 3-dimensional input data, calculated from the Kinect data, describes the relative position from shoulder where shoulder position is set at $(0,0,0)$, A is a set of node indices, N_i is a set of node indices connected to the ith node, and $a_{i,j}$ is the age of the edge between the ith and the jth node.

Algorithm 1. GNG ALGORITHM

Step 1: Generate two units at random position, $\mathbf{r}_1, \mathbf{r}_2$ in $\mathbb{R}^3$. Initialize the connection set.

Step 2: Generate an input data $\mathbf{v}$ randomly according to $p(\mathbf{v})$ which is the probability density function of data $\mathbf{v}$.

Step 3: Select the nearest unit (winner), s_1 by Eq. (7) and the second-nearest unit, s_2 by Eq. (8).

$$s_1 = \arg\min_{i \in A} ||\mathbf{v} - \mathbf{r}_i|| \tag{7}$$

$$s_2 = \arg\min_{i \in A \setminus \{s_1\}} ||\mathbf{v} - \mathbf{r}_i|| \tag{8}$$

Step 4: If a connection between s_1 and s_2 does not exist already, create the connection. Set the age of the connection between s_1 and s_2 to zero, $a_{s_1,s_2} = 0$.

Step 5: Add the squared distance between the input data and the winner to a local error variable (which is initialized as 0): $E_{s_1} \leftarrow E_{s_1} + ||\mathbf{v} - \mathbf{r}_{s_1}||^2$.

Step 6: By using the total distance to the input data, update the reference vectors of the winner node, s_1 (see Eq. (7)) using Eq. (9) and its direct topological neighbors using Eq. (10) by the learning rate η_1 and η_2, respectively.

$$\mathbf{r}_{s_1} \leftarrow \mathbf{r}_{s_1} + \eta_1 \cdot (\mathbf{v} - \mathbf{r}_{s_1}) \tag{9}$$

$$\mathbf{r}_j \leftarrow \mathbf{r}_j + \eta_2 \cdot (\mathbf{v} - \mathbf{r}_j) \quad \text{if } j \in N_{s_1} \tag{10}$$

Step 7: Increment the age of all edges emanating from $s_1 : a_{s_1,j} \leftarrow a_{s_1,j} + 1 \quad \text{if } j \in N_{s_1}$

Step 8: Remove edges with an age larger than a_{max}. If this results in units having no more emanating edges, remove those units as well.

Step 9: If the number of input signals generated so far is an integer multiple of a parameter λ, insert a new unit using the following steps:

a. Select the unit q with the maximum accumulated error according to Step 5.

b. Add a new unit r to the network and interpolate its reference vector from q and f using Eq. (11), where f is that neighbor of q which the largest error has according to Step 5.

$$\mathbf{r}_r = 0.5 \cdot (\mathbf{r}_q + \mathbf{r}_f) \tag{11}$$

c. Insert edges connecting the new unit r with units q and f, and remove the original edge between q and f.

d. Decrease the error variables of q and f by a fraction α:

$$E_q \leftarrow E_q - \alpha \cdot E_q \tag{12}$$

$$E_f \leftarrow E_f - \alpha \cdot E_f \tag{13}$$

e. Interpolate the error variable of r from q and f:

$$E_r = 0.1 \cdot (E_q + E_f) \tag{14}$$

Step 10: Decrease the error variables of all units:

$$E_i \leftarrow E_i - \beta \cdot E_i \quad (\forall i \in A) \tag{15}$$

Step 11: Continue with Step 2 if a stopping criterion is not yet fulfilled. The net size or some performance measure can be used as a stopping criterion.

5.2.2. Spiking neural network for gesture recognition

In the second stage of the proposed method, the spiking neural network is applied to recognize the human gestures. However, in order to reduce the computational cost, a modified spike response model is applied in this paper as to Botzheim and Kubota (2012), Tang, Botzheim, Kubota, and Yamaguchi (2013) and Kubota, Toda, Botzheim, and Tudjarov (2013). We use two-layered SNNs which are composed of an input layer and an output layer. Each gesture is recognized by one SNN. The number of spiking neurons in each SNN is N, which is the same as the number of reference vectors. The proposed model is depicted in Fig. 10.

The internal state $h_{k,i}(t)$ of a spiking neuron i in the input layer for the kth gesture is calculated as follows:

$$h_{k,i}(t) = \gamma^{syn} \cdot h_{k,i}(t-1) + h^{syn}_{k,i}(t) + h^{ref}_{k,i}(t) + h^{ext}_{i}(t), \tag{16}$$

where γ^{syn} is a temporal discount rate, $h^{syn}_{k,i}(t)$ includes the pulse outputs from the other neurons, $h^{ref}_{k,i}(t)$ is used for representing the refractoriness of the neuron, $h^{ext}_{i}(t)$ is the input to the ith neuron from the environment.

The input to the ith neuron from the external environment is calculated by the difference between the reference vector and the input vector:

$$h^{ext}_{i}(t) = \exp\left(-\gamma^{env} \cdot (\mathbf{v} - \mathbf{r}_i)^2\right), \tag{17}$$

where γ^{env} is a coefficient, $\mathbf{v}$ is the input vector.

The pulse outputs from the other neurons, $h^{syn}_{k,i}(t)$ is calculated by:

$$h^{syn}_{k,i}(t) = \begin{cases} \tanh\left(\sum_{j=1, j\neq i}^{N} w^{ges}_{k,j,i} \cdot h^{PSP}_{k,j}(t-1)\right) & \text{if } h^{ext}_{i}(t) \geqslant \theta^{syn}, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

where $w^{ges}_{k,j,i}$ is the weight from the jth neuron to the ith neuron in the kth SNN (kth gesture), θ^{syn} is a threshold, $h^{PSP}_{k,j}(t)$ is the PostSynaptic Potential (PSP) approximately transmitted from the jth neuron in the kth SNN at the discrete time t. The hyperbolic tangent is used to avoid the repeated firing by several neurons without an efficient input (without reaching the θ^{syn} threshold).

When the internal state of the ith neuron reaches a predefined threshold level, a pulse is outputted as follows:

$$p_{k,i}(t) = \begin{cases} 1 & \text{if } h_{k,i}(t) \geqslant \theta^{pul}, \\ 0 & \text{otherwise}, \end{cases} \tag{19}$$

where θ^{pul} is a threshold for firing. In case of firing, R is subtracted from the $h^{ref}_{k,i}(t)$ value of neuron i as follows:

$$h^{ref}_{k,i}(t) = \begin{cases} \gamma^{ref} \cdot h^{ref}_{k,i}(t-1) - R & \text{if } p_{k,i}(t-1) = 1, \\ \gamma^{ref} \cdot h^{ref}_{k,i}(t-1) & \text{otherwise}, \end{cases} \tag{20}$$

where γ^{ref} is a discount rate of $h^{ref}_{k,i}$ and $R > 0$.

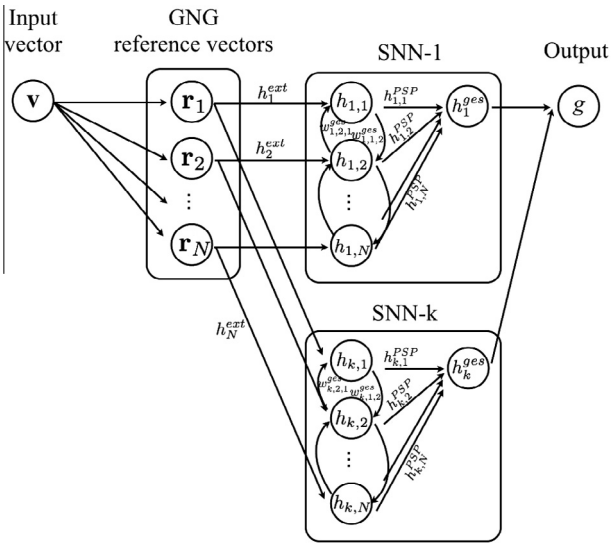


Fig. 10. Our proposed spiking neural network for gesture recognition.

The presynaptic spike output is transmitted to the connected neuron through the weight connection. The PSP is calculated as follows:

$$h_{k,i}^{PSP}(t) = \begin{cases} 1 & \text{if } p_{k,i}(t) = 1, \\ \gamma^{PSP} \cdot h_{k,i}^{PSP}(t-1) & \text{otherwise}, \end{cases} \tag{21}$$

where γ^{PSP} is a discount rate of $h_{k,i}^{PSP}$ and $0 < \gamma^{PSP} < 1$. The PSP is excitatory if the weight parameter, $w_{k,j,i}^{ges}$ is positive. If the condition $h_{k,j}^{PSP}(t) < h_{k,i}^{PSP}(t)$ is satisfied, the weight parameter is trained based on the temporal Hebbian learning rule (Hebb, 1949):

$$w_{k,j,i}^{ges} \leftarrow \tanh\left(\gamma^{wgt} \cdot w_{k,j,i}^{ges} + \xi^{wgt} \cdot h_{k,j}^{PSP}(t) \cdot h_{k,i}^{PSP}(t)\right), \tag{22}$$

where γ^{wgt} is a discount rate of the weights and ξ^{wgt} is a learning rate.

The evaluation value for the kth gesture in the output layer is calculated by:

$$h_k^{ges}(t) = \gamma^{ges} \cdot h_k^{ges}(t-1) + \sum_{i=1}^{N} h_{k,i}^{PSP}(t), \tag{23}$$

where γ^{ges} is a discount rate. The gesture recognition at the discrete time t is done by:

$$g(t) = \arg\max_k h_k^{ges}(t-1). \tag{24}$$

Finally, the overall recognition result is calculated as the most frequently selected gesture over time. The information flow is illustrated in Fig. 10.

6. Database system

We propose the database structure as illustrated in Fig. 11 to realize the informationally structured space. The database structure is divided into eight parts; (A) Human condition, (B) Personal model, (C) Life log, (D) Human behavior, (E) Conversation log, (F) Conversation contents, (G) Web information, and (H) Sensor raw data.

The sensor raw data (H) is acquired from the measurement by sensor network inside the room and by the smartphone outside the room. Web information (G) is the database that stores the information extracted from the Web to generate sentences in the conversation. Conversation contents (F) provides robot partner with conversation contents to support the communication between the robot partner and the human. The conversation contents is composed of (F1) Scenario conversation, (F2) Daily conversation, and (F3) Information support. Conversation between the robot and the human recorded as Conversation log (E).

Human behavior (D) records human behavior estimation results, which are estimated by active sensing from the robot partner (E) and passive sensing from the sensor network (H). Life log (C) is the database of life log. It is constructed by storing human behavior in time. Human condition (A) and Personal model (B) are statistical analysis result of Life Log. Personal models are the database of personal models, which lead to individual lifestyle and preference extraction. Human states are the database for estimating regular and irregular human state.

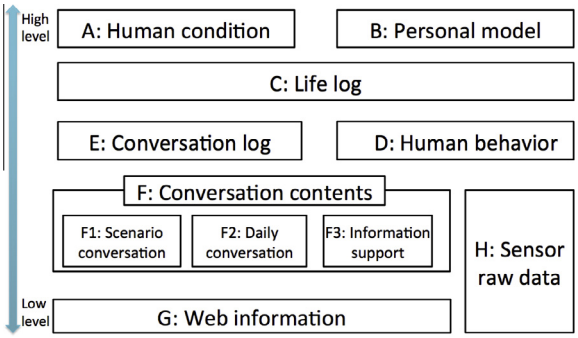


Fig. 11. Database structure.

7. Conversation system

The conversation system has been developed for many years with various architectures and standardizations. In the interaction between a human and a robot partner, when the robot partner leads the conversation, the robot partner has a good performance if the human's expectation can be fulfilled. However, if the human's expectation cannot be fulfilled, the interaction between them will be broken. On the other hand, if the human leads the conversation, interaction building between the human and the robot partner is difficult owing to the current technology. Therefore, in order to realize natural communication, we suppose that first the robot partner leads the conversation, ideally in the middle of interaction the human also takes place to lead the conversation interactively. This process can be performed, as long as the robot partner conducts sequential transitional behavior, while in arbitrary timing it performs some action reflectively to human interruption behavior.

According to Rasmussen, Pejtersen, and Goodstein (1994), human behavior based on information processing is composed of skill level, rule level, and knowledge level. The skill level is a daily common and repetitive behavior, which does not need memory and knowledge referring process while performing it. In other word, the skill level can be defined as an unconscious, reflective, and short time behavior. Meanwhile, the rule level is defined as a behavior based on customs and rules. This kind of behavior needs human memory and knowledge referencing process in order to perform the behavior correctly. Comparing to skill level, the rule level needs more time to conduct. The knowledge level is behavior performance for unknown or unfamiliar situations. In order to perform this kind of behavior, sufficient knowledge is needed. Otherwise, during the performance, the human can have new knowledge while doing some optimization to get the best result.

In the wider application, Rasmussen's behavior model has been applied to human and voice interface. We also believe that Rasmussen's behavior model can be applied to conversation system as well. Here, we propose a human conversation system based not only on Rasmussen's behavior model, but also on existed conversation architecture. As shown in Fig. 12, the conversation system is divided into three parts, such as skill, rule, and knowledge based conversation system. In the skill based conversation system, daily conversation with reflective, repeated, and short conversation is performed. The rule based conversation system takes place as information support conversation. Here, using the word understanding model, the latest news or weather condition can be requested. In the other situation, information support conversation can be conducted based on time and on the human's condition. The knowledge based conversation is performed based on the understanding of conversation keyword. Through this keyword, conversation is conducted by picking suitable dialogue topic scenario.

From conversation architecture, the conversation system can be structured into detailed parts as shown in Fig. 13. The conversation structure is divided into four layers, including mode, node1, node2 and contents. The first layer "mode", pointed the conversation mode. While the second and third layer show subcategory of the conversation mode. The last layer mentions the conversation contents. Using the conversation structure, the robot partner can perform the learning process to select suitable conversation contents based on the people's state and time. Figs. 14 and 15 show the daily

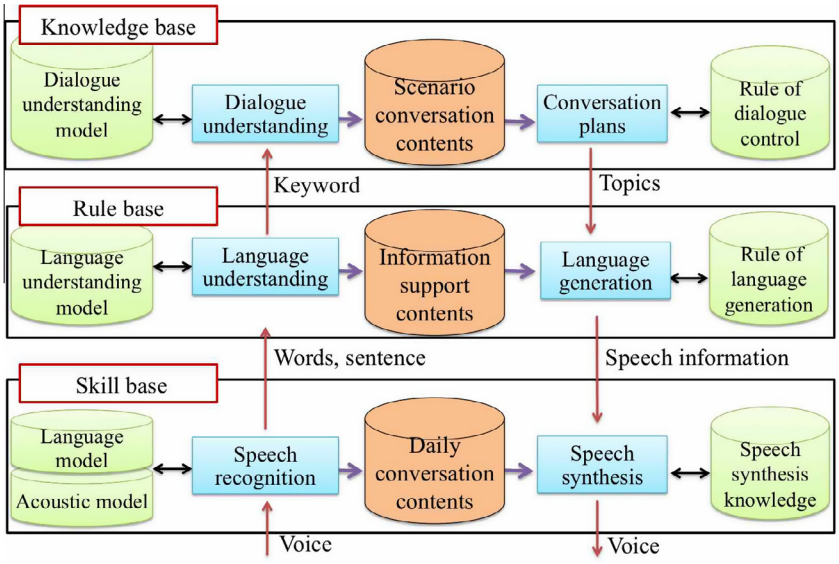


Fig. 12. Conversation system architecture.

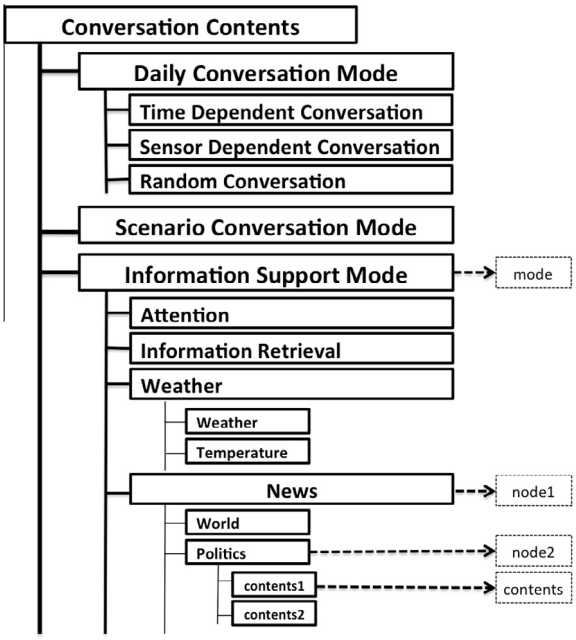


Fig. 13. Conversation contents structure.

conversation contents and information conversation contents used in the experiment process. The parameters are defined in the left side of conversation contents column as classifiers in selecting conversation contents.

In the experiment part, as shown in Fig. 16, we define the details of some parameters such as human state, human behavior, human gesture, robot emotion and robot action. For the utterance selection module, we propose a method that composed of two stages: (1) utterance group selection and (2) word or sentence selection. Basically, an utterance group is composed of different words or sentences having the same meaning, e.g., "hello"=hi, hello, ya. First, the utterance group is selected according to the flow of the context in the conversation control module and perceptual information. Next, one word or sentence is stochastically selected from the group according to the state of feelings. The state of feelings corresponds to the utterance group is calculated using spiking neural network.

When the spiking neuron corresponding to the ith utterance group fires, the selection strength ($s^G_{i,k}$) of the kth words in the ith utterance group related to the jth feeling is calculated by

$$s^G_{i,k} = u^F_j(t) \cdot \exp\left(-\left(u^F_j - F^G_{j,i,k}\right)^2\right). \tag{25}$$

After that, the selection probability $\left(p^G_{i,j}\right)$ is calculated using Boltzmann selection scheme as follows;

$$p^G_{i,j} = \frac{\exp\left(s^G_{i,k}/T^G\right)}{\sum_{g=1}^{N^G_i} \exp\left(s^G_{i,g}/T^G\right)} \tag{26}$$

where T^G is a positive value called the temperature, N^G_i is the number of candidate words in the ith utterance group. According to the equation, when the temperature is high, the robot partner will randomly select utterance words from the ith utterance group. As the temperature decreases, the robot partner deterministically selects the utterance words with high selection strength. At the same time, the robot partner selects hand gesture corresponding to the selected utterance (Fig. 17).

8. Experiment

In this paper, we divided the experiment into three case studies. In the first case study, we want to investigate and validate the proposed emotional model through the computation of environmental conditions and human behaviors (movement and distance) effects to robot partner. For the second case study, we conducted the experiment much further about the robot partner's emotions in conjunction with human gesture recognition. Finally, in the third case study, we investigate the feasibility of our integrated system, starting from data collection through sensors, data processing, and conversation system.

8.1. Case study 1

In this experiment, we considered the effects of environmental conditions during the communication with the robot partner. Table 6 shows the contribution parameters from perception input to feeling. These parameters are acquired as the optimum result of the trial and error parameter settings. Fig. 18 displays the

contents_id	mode	node1	node2	time	h_state	h_behavior	h_gesture	r_action	speech	contents
11001001000	1	100	100	420	1	null	0	1	Hi,Hello,Morning	Good morning
11001011000	1	100	101	465	1	null	0	0	null	It's time to take breakfast
11011011002	1	101	101	450	1	13	0	0	null	Haven't you feel hungry already
11011011003	1	101	101	450	1	13	0	0	null	It's good to take breakfast before 9 o'clock

Fig. 14. Daily conversation contents.

contents_id	mode	node1	node2	time	h_state	h_behavior	h_gesture	r_action	speech	contents
21001001000	2	100	100	420	1	1	0	5	Weather	Today's weather is sunny
21001011000	2	100	101	420	1	1	0	5	Wather,Temperature	Today's maximum temperature is 28 degree
21011001000	2	101	100	420	1	1	0	5	News,World	Laos government's plane crash cause death of vice prime minister
21011011000	2	101	101	420	1	1	0	5	News,Politics	Don't forget to bring towel
21021001001	2	102	100	435	1	14	0	5		The floor is slippery, please watch your step
21021011000	2	102	101	495	1	13	0	5		Don't forget to have a glass of water

Fig. 15. Information support contents.

No	Human state
0	nobody in the sensing range
1	human detection
2	more people detection

(a) Human state

No	Robot emotion
0	happy
1	sad
2	fearful
4	angry

(b) Robot Emotion

No	Human behavior
0	sleep
1	get up
11	walk
12	launder
13	open refrigerator
14	get in toilet
15	sit down

(c) Human behavior

No	Human gesture
0	No gesture recognition
1	Hello
2	Come here
3	Swing
4	Stop
5	Bye bye

(d) Human gesture

No	Robot action
0	No action
1	Hello
2	Come here
3	Swing
4	Bye bye
5	Information support

(e) Robot Action

Fig. 16. Conversation contents parameters.

communication process between the robot partner and the person in the experiment. In the experiment, as the environment brightness is high, the robot partner is in the state of happiness (a,c,e)

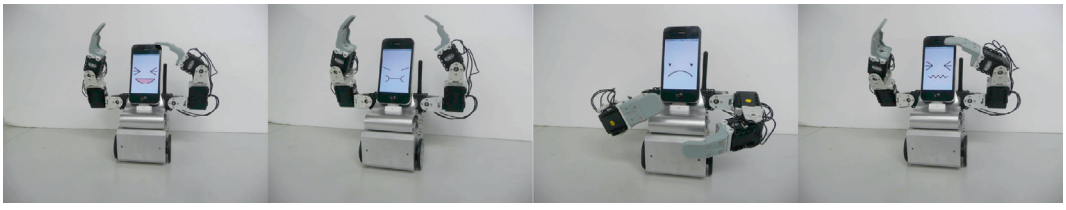
Table 6
Contribution parameters from perception input to feeling.

	Pleasure	Sadness	Fear	Anger
People detected	0.1	−0.1	−0.2	−0.2
Distance to human	0.3	0	0.1	0
Human activity	0.2	0	0.1	0
Brightness	0.4	−0.2	−0.1	−0.1
Darkness	−0.2	0.2	0.4	0.1
No people detected	−0.1	0.2	0.03	0.01

as depicted in Fig. 19. In addition, when the person started to make some action (moving action), the happiness value increases faster than before, especially when the person walked near the desk (d,f). Meanwhile, the fear value increases when the room changed to dark (b). From here, we can notice that linkage between the emotion model and environmental changes. This is a very crucial aspect as the robot partner need to be very sensitive to the environmental changes, when communicating with the human.

8.2. Case study 2

In the second experiment, gesture recognition is considered in the communication with the robot partner. Through gesture recognition, the robot partner is expected to understand the human behavior using non-verbal communication. On contrary with the first experiment, in this experiment we added some contribution parameters for gesture as shown in Table 7. In Fig. 20, we can see the environmental conditions including the changing of the room brightness and the human gesture recognition process. Fig. 21 depicts the experiment results through a graph



(a) Pleasure (b) Anger (c) Sadness (d) Fear

Fig. 17. Facial and gestural expressions.

Fig. 18. Snapshots of the first experiment.

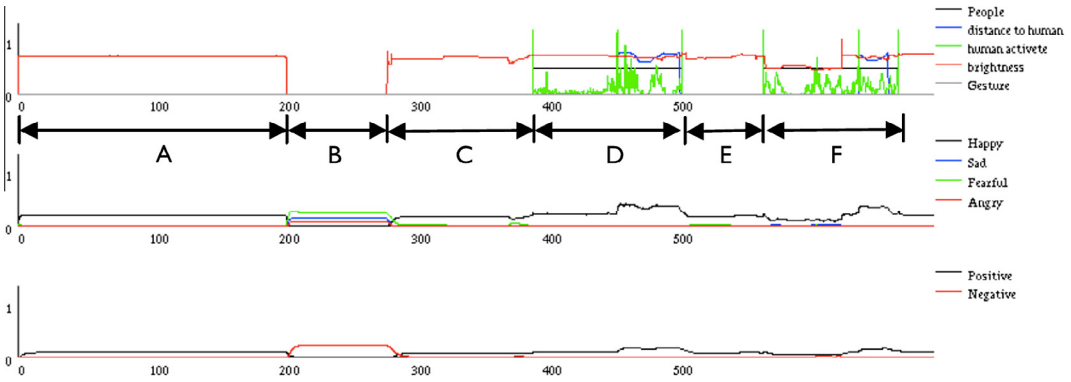


Fig. 19. The change of states in the first experiment.

Table 7
Contribution parameters from perception input to feeling and gesture .

	Pleasure	Sadness	Fear	Anger
People detected	0.1	−0.1	−0.2	−0.2
Distance to human	0.3	0	0.1	0
Human activity	0.2	0	0.1	0
Brightness	0.4	−0.2	−0.1	−0.1
Darkness	−0.2	0.2	0.4	0.1
No people detected	−0.1	0.2	0.03	0.01
Gesture (Bye Bye)	−0.1	0.5	0.1	0
Gesture (Hello)	0.4	0	−0.1	−0.1
Gesture (Stop)	−0.3	0.05	0.03	0.5
Gesture (Swing)	0.4	0	−0.1	−0.1

representation. In this graph, the fear value increased when the room was dark (a). The happiness value increases and the fear value decreases when the person comes into the room and turns on the light (b). When the person entered into the view, the happiness value increased (c), however when the robot partner detected the [stop] gesture, the anger value increased, on the contrary the happiness value started to decrease (d).

8.3. Case study 3

In the case studies 1 and 2, based on the changes of the environmental condition and human gesture in informationally structured space, we realized the robot partner's emotion building process. Through the input data for the emotional model, the robot partner can express its emotions based on the current condition. In this experiment, as an information support system we realized multimodal communication between the robot partner and the human based on environmental conditions. As shown in Fig. 22 we built an environmental model of elderly people's house for the experiment. In this room, we installed SunSPOT in chair, toilet, and refrigerator. In addition, we also installed wireless optical oscillo-sensor in the bed and Microsoft Kinect above the bookshelf.

The experiment is conducted to investigate the information support by the robot partner based on human behavior. Starting from getting up from the bed, sitting on the chair, going to the toilet and preparing breakfast. The snapshot of the experiment can be seen in Fig. 23. The human behavior record which is saved in the human state database is expressed in Fig. 24, where it can be seen

Fig. 20. Snapshots of the second experiment.

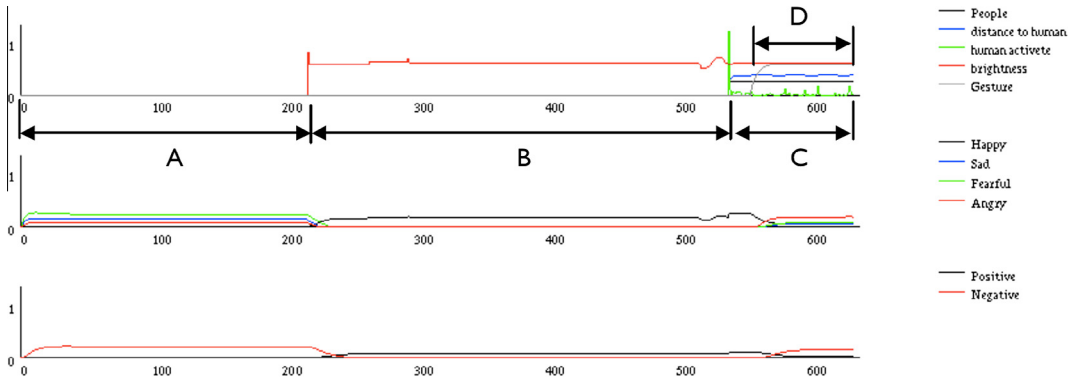


Fig. 21. The change of states in the second experiment.

Fig. 22. Experimental room.

(a) Sleeping on bed (b) Getting up from bed (c) Walking around (d) Sitting on chair

Fig. 23. Snapshots of the third experiment.

that at (a) shows human when sleeping on the bed, (b) shows human when getting up from the bed, (c) shows human when walking around the room, and (d) shows human when sitting on the chair. The robot partner uses this information as a reference in conducting communication with the human. Fig. 25 shows the sample of human behavior database. The information in this figure is used to recognize the human's state. Besides human state, human behavior, human gesture, and robot emotion as the main

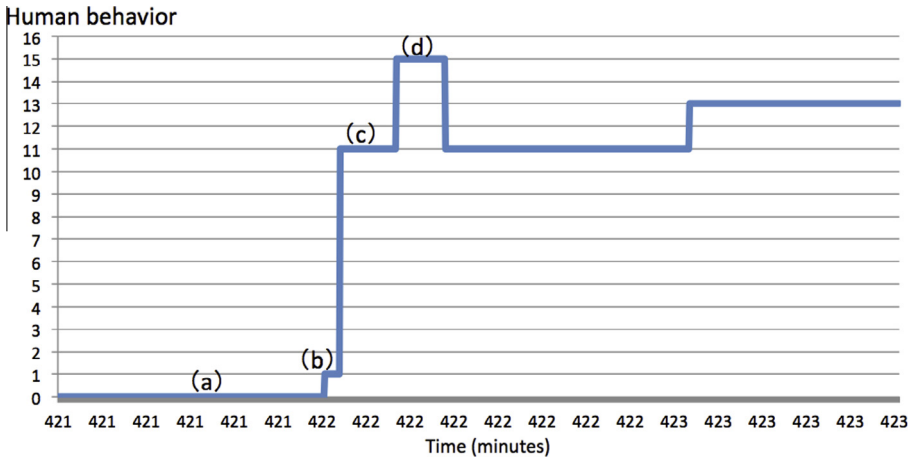


Fig. 24. Graph of human behavior log.

id	timestamp	h_state	h_behavior	h_gesture	r_emotion	season	year	month	week	day	time
8282	2014/6/21 7:01	1	0	0	0	2	2014	6	6	21	421
8283	2014/6/21 7:01	1	0	0	0	2	2014	6	6	21	421

Fig. 25. Human behavior log.

attributes, and the other attributes such as season, year, month, day, week, and time (minutes) are used to recognize personal life time and build the personal model of the human.

Fig. 26 shows data saving in the conversation log database during the experiment. In order to conduct learning process in the conversation process, human and environmental state connected with time become the main attribute. For the time attribute, we use season, year, month, day, week, and time (minutes). Meanwhile, human state, human behavior, human gesture, robot motion, and robot action are used as the attributes to decide utterances. For classifying the conversation contents, we define mode, node1, node2, and content id for the attributes. The conversation contents of the conversation log database is depicted in Fig. 27.

8.4. Discussions

Through the experimental results, we can conclude that, in the case study 1, feeling is affected dominantly by the brightness of the room. As the perception parameter value on brightness is high, intuitively we could guess, that the brightness as well as the darkness of the room will give important effect on the robot partner's feelings. On the other hand, although human movement and distance also give some effect to the feeling, but still could not replace the dominance of room brightness effect.

In the case study 2, basically the robot partner's feelings are affected by the room's brightness. However, when the robot partner recognized a particular human gesture, the dominance of the room's brightness is replaced by human gesture resulting in the changing of the robot partner's feeling. In the case study 3, the realization of informationally structured space is conducted. Here, we took the experiment for several days to investigate the human life cycle activity. The information acquired from and outputted into environment is saved in the database, and used for recognizing the state and learning in the conversation process. In the result, after recognizing the situation, the robot partner could conduct suitable conversation with the human.

As mentioned in the introduction, there are some research methods that aim similar goals as we also aim in this research. For example Takemura and Ishii (2011) was focusing their research on cognitive environment, while Shahdi and Bakar (2012) was

timestamp	season	year	month	day	week	time	h_state	h_behavior	h_gesture	r_motion	r_action	mode	node1	node2	contents_id
2014/6/21 7:02	2	2014	6	21	6	421	1	1	0	1	1	1	101	100	11011001007
2014/6/21 7:02	2	2014	6	21	6	421	1	15	0	1	0	1	101	101	11011011013
2014/6/21 7:02	2	2014	6	21	6	420	1	15	0	1	5	2	100	100	21001001000
2014/6/21 7:03	2	2014	6	21	6	422	1	13	0	1	0	1	100	101	11001011007

Fig. 26. Conversation log in database.

	Conversation	Human Behavior
Robot partner:	Hi!	Get up from bed
Human:	Morning!	Get up from bed
Robot partner:	Where are you going today?	Sit on chair
Human:	I haven't decided yet	Sit on chair
Robot partner:	How about today's weather	Sit on chair while asking weather information
Robot partner:	Today is cloudy	Sit on chair while asking weather information
Robot partner:	Let's have breakfast together	Open the refrigerator

Fig. 27. Conversation log.

Table 8
Comparison to other methods.

Methods	Cognitive environment	Human recognition	Considering emotions
Takemura and Ishii (2011)	√	–	–
Shahdi and Bakar (2012)	–	√	–
Botzheim et al. (2012)	–	√	–
Böhme et al. (2003)	–	√	–
Bien and Lee (2007)	–	√	√
Botzheim et al. (2013)	√	√	–
Proposed method	√	√	√

focusing theirs in human recognition technology. Meanwhile Bien and Lee (2007) conducted their research which related to human recognition and emotional consideration. In contrary, our proposed method including the cognitive environment, human recognition, and emotional consideration realizes the robot partner's reaction through various input data. For example, when the human gets up from the bed, the robot partner with its environmental cognitive ability for recognizing human condition gives some reactions such as utterance and particular gesture to reinforce the utterance contents. Moreover, with the combination of gesture recognition and emotional model as we can see in the experimental result, the communication between the robot partner and the human can be realized in a more natural and interesting way. More details about the comparison of the methods can be seen in Table 8.

9. Conclusion and future works

This paper discussed the actualization of natural communication between a robot partner and a human based on the application of relevance theory in the mutual recognition space. In order to realize this, we built the architecture in informationally structured space as the basis of our research along with the database system, which supports the informationally structured space in transferring data to and from the environment. While discussing informationally structured space, we also explained the elements constructed our system such as sensor network, web system, and robot partner including gesture recognition and emotional model. In further, the conversational architecture which allows the natural communication to be realized was deeply discussed. This includes the discussion of database structure and conversation selection algorithm.

In terms of the theoretical contribution, the proposed system is built based on four theories as follows: (a) Relevance Theory – Realizing natural communication between human and robot is the main focus of this paper. Natural communication can be realized when robot can understand human intention or thought. We implemented the theory of relevance proposed by Sperber and Wilson (1995) to build mutual cognitive environment between human and robot partner into our system, which called Informationally Structured Space to handle this problem. According to Sperber and Wilson (1995), relevance theory is very useful to discuss the multimodal communication, where each person has his or her own cognitive environment that make their communication restricted. Therefore, usually humans use their utterances or gestures to expand their cognitive environment by extracting person's attention into specific target object, event, or person. When human's cognitive environment became wider, they can share each other intention or thought. The implementation of this theory into our system can be observed in the structure of database. (b) Rasmussen's Behavior Theory – For conducting daily conversation with human, robot partner has to have enough knowledge and contents. However, to get enough knowledge and contents, the conversation system contents and task will become bigger. At this state, flexibility and simplicity of the system become a new issue. To deal with this, we proposed new conversation system architecture based on Rasmussen's behavior model. Using this model, we divided the conversation into three types based on complexity. (c) Computational Intelligence (Soft Computing) – In the gesture recognition, we implemented structured learning involving growing neural gas for information extraction and spiking neural network to recognize the gesture, and finally (d) Boltzmann Selection – For acquiring non-monotonic or lively conversation between human and robot partner we used Boltzmann selection by controlling the value of temperature to perform word selection when the robot partner communicates with the human.

In terms of practical implementation, the development of sensor network installed in nursing home to support caregivers in conducting monitoring for elderly people has been increased. However, in order to give information support and encourage elderly people for social activities, mutual cognitive environment between the human and the robot must be built in order to realize natural communication. The implementation of relevance theory and Rasmussen's model into this system has been discussed in the previous section. The experimental results explained the capability of the proposed method to be applied in the real world. In addition, since we introduced a low cost robot partner using iPhone as a mainframe in the proposed system, this system can be applied in the real world in a short time as an advanced elderly people monitoring system.

There are few possible developments of our system to be upgraded or adding some new features in order to improve it. Currently, we developed the communication between robot partner and human using only English as user language. One of the possible future work is extending the user language into several languages, since we also conduct the research about cultural comparison in Botzheim, Yusuf, Kubota, and Yamaguchi(2013). Secondly, we build mutual cognitive environment named informationally structured space based on relevance theory. In the informationally structured space, the change of environment and human states effect the robot partner's emotion and conversation system. In the future work, in order to understand human behavior, we intend to build mutual cognitive environment focused on human. The collection data of human such as human state, life log and human preferences will become important aspects. Moreover, we will also extend the capability of sensor and robot partner. Beside these, additional information such as location and furniture information will be included. Using these extended information, we expect new features such as (a) Elderly people can check their living condition using visualization system built by using life log data in informationally structured space. This information can also be shared and used by the elderly people's family and caregivers as a remote monitor, (b) The upper feature leads to a feature in early detection of the unusual life pattern or unusual behavior of elderly people to acquire immediately assistance, (c) We will conduct the integration and fusion between sensors to extend sensor capability. With this method, the sensor measurement error can be avoided, which will raise the robustness, accuracy and efficiency of the sensor. Additionally, using the visualization system based on the sensor information, we can conduct the maintenance process easily. Finally, with the extension of robot partner capability, besides supporting in daily live conversation, robot partner can also be applied as a recommender and reminder system.

Acknowledgments

This work was partially supported by MEXT Regional Innovation Strategy Support Program: Greater Tokyo Smart QOL (Quality of Life) Technology Development Region; and Chee Seng Chan is supported in part by The Hitachi Research Fellowship, from The Hitachi Scholarship Foundation, Japan and High Impact MoE Grant

UM.C/625/1/HIR/MoE/FCSIT/08, H-22001-00-B00008 from the Ministry of Education Malaysia.

References

Alexandre, D. S., & Tavares, J. M. R. S. (2010). Introduction of human perception in visualization. *International Journal of Imaging and Robotics, 4*, 60–70.

Bien, Z. Z., & Lee, H-E. (2007). Effective learning system techniques for human–robot interaction in service environment. *Knowledge-Based Systems, 20*, 439–456.

Böhme, H. J., Wilhelm, T., Key, J., Schauer, C., Schröter, C., Gross, H., et al. (2003). An approach to multi-modal human–machine interaction for intelligent service robots. *Robotics and Autonomous Systems, 44*, 83–96.

Botzheim, J., & Kubota, N. (2012). Growing neural gas for information extraction in gesture recognition and reproduction of robot partners. In *Proc. of the 23rd international symposium on micro-nanomechatronics and human science* (pp. 149–154).

Botzheim, J., Obo, T., & Kubota, N. (2012). Human gesture recognition for robot partners by spiking neural networks and classification learning. In *Proc. of the 6th international conference on soft computing and intelligent systems and the 13th international symposium on advanced intelligent systems* (pp. 1954–1958).

Botzheim, J., Yusuf, B., Kubota, N., & Yamaguchi, T. (2013). Computational intelligence for gestural communication using emotional model. In *Proc. of the 3rd international workshop on advanced computational intelligence and intelligent informatics*.

Botzheim, J., Tang, D., Yusuf, B., Obo, T., Kubota, N., & Yamaguchi, T. (2013). Extraction of daily life log measured by smart phone sensors using neural computing. *Procedia Computer Science, 22*, 883–892.

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: A review. *Gerontechnology, 8*, 94–103.

Cerpa, A., Elson, J., Estrin, D., Girod, L., Hamilton, M., & Zhao, J. (2001). Habitat monitoring: application driver for wireless communications technology. *Computer Communication Review, 31*, 20–41.

Chan, C. S., Liu, H., & Brown, D. J. (2007). Recognition of human motion from qualitative normalised templates. *Journal of Intelligent and Robotic Systems, 48*, 79–95.

Chernbumroong, S., Cang, S., Atkins, A., & Yu, H. (2013). Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications, 40*, 1662–1674.

Cook, D. J., Augusto, J. C., & Jakkula, V. R. (2009). Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and Mobile Computing, 5*, 277–298.

Costa, A., Castillo, J. C., Novais, P., Caballero, A. F., & Simoes, R. (2012). Sensor-driven agenda for intelligent home care of the elderly. *Expert Systems with Applications, 39*, 12192–12204.

Daosodsai, N., & Maneewarn, T. (2013). Fuzzy based emotion generation mechanism for an emoticon robot. In *Proc. of the 13th international conference on control, automation and systems* (pp. 1073–1078).

Forkan, A., Khalil, I., & Tari, Z. (2014). Cocamaal: A cloud-oriented context-aware middleware in ambient assisted living. *Future Generation Computer Systems, 35*, 114–127.

Forkan, A. R. M., Khalil, I., Tari, Z., Foufou, S., & Bouras, A. (2015). A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living. *Pattern Recognition, 48*, 628–641.

Fritzke, B. (1992). Growing cell structures – A self-organizing network in *k* dimensions. *Artificial Neural Networks, 2*, 1051–1056.

Fritzke, B. (1995). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems, 7*, 625–632.

Fritzke, B. (1996). Growing self-organizing networks – why? In *Proc. of the european symposium on artificial neural networks* (pp. 61–72).

Hebb, D. O. (1949). *The organization of behavior*. New York, USA: Wiley and Sons.

Iengo, S., Rossi, S., Staffa, M., & Finzi, A. (2014). Continuous gesture recognition for flexible human–robot interaction. In *Proc. of the 2014 IEEE international conference on robotics and automation* (pp. 4863–4868).

Ingelrest, F., Barrenetxea, G., Schaefer, G., Vetterli, M., Couach, O., & Parlange, M. (2010). Sensorscope: Application-specific sensor network for environmental monitoring. *ACM Transactions on Sensor Networks, 6*, 1–32.

Jeng, T. (2009). Toward a ubiquitous smart space design framework. *Journal of Information Science and Engineering, 25*, 675–686.

Jitviriya, W., & Hayashi, E. (2014). Design of emotion generation model and action selection for robots using a self organizing map. In *Proc. of the 11th international conference on electrical engineering/electronics, computer, telecommunications and information technology* (pp. 1–6).

Johanson, B., Fox, A., & Winograd, T. (2002). The interactive workspaces project: Experiences with ubiquitous computing rooms. *IEEE Pervasive Computing, 1*, 67–74.

Kim, J., Kim, Y., & Lee, K. (2004). The third generation of robotics: Ubiquitous robot. In 2nd *International conference on autonomous robots and agents*.

Kim, H. -G., Yang, J. -Y., & Kwon, D. -S. (2013). Episodic memory system of affective agent with emotion for long-term human–robot interaction. In *Proc. of the 10th international conference on ubiquitous robots and ambient intelligence* (pp. 720–722).

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer.

Kubota, N. (2005). Computational intelligence for structured learning of a partner robot based on imitation. *Information Sciences, 171*, 403–429.

Kubota, N. (2008). Cognitive development of partner robots based on interaction with people. In *Proc. of the joint 4th international conference on soft computing and intelligent systems and 9th international symposium on advanced intelligent system* (pp. 820–825).

Kubota, N., & Yorita, A. (2009). Topological environment reconstruction in informationally structured space for pocket robot partners. In *Proc. of the 2009 IEEE international symposium on computational intelligence in robotics and automation* (pp. 165–170).

Kubota, N., Tang, D., Obo, T., & Wakisaka, S. (2010). Localization of human based on fuzzy spiking neural network in informationally structured space. In *Proc. of the IEEE world congress on computational intelligence* (pp. 2209–2214).

Kubota, N., Toda, Y., Botzheim, J., & Tudjarov, B. (2013). Multi-modal perception for human-friendly robot partners with smart phones based on computational intelligence. In *Proc. of the fifth international conference of south-west university, faculty of mathematics & natural sciences* (pp. 17–25).

Kubota, N., & Nishida, K. (2006). Cooperative perceptual systems for partner robots based on sensor network. *International Journal of Computer Science and Network Security, 6*, 19–28.

Kubota, N., Nojima, Y., Kojima, F., & Fukuda, T. (2006). Multiple fuzzy state-value functions for human evaluation through interactive trajectory planning of a partner robot. *Soft Computing, 10*, 891–901.

Kubota, N., & Wakisaka, S. (2010). Emotional model based on computational intelligence for partner robots. In T. Nishida (Ed.), *Modeling machine emotions for realizing intelligence* (pp. 89–108). Berlin Heidelberg: Springer-Verlag.

Lee, J. H., Lee, H., Kim, M. J., Wang, X., & Love, P. E. (2014). Context-aware inference in ubiquitous residential environments. *Computers in Industry, 65*, 148–157.

Lim, M. K., Tang, S., & Chan, C. S. (2014). iSurveillance: Intelligent framework for multiple events detection in surveillance videos. *Expert Systems with Applications, 41*, 4704–4715.

Liu, J., Pasupat, P., Cyphers, S., & Glass, J. (2013). Asgard: A portable architecture for multilingual dialogue systems. In *ICASSP'13* (pp. 8386–8390).

Liu, Z., Hu, F., Luo, D., & Wu, X. (2014). Visual gesture recognition for human robot interaction using dynamic movement primitives. In *Proc. of the 2014 IEEE international conference on systems, man and cybernetics* (pp. 2094–2100).

Lopes, J., Eskenazi, M., & Trancoso, I. (2015). From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech and Language, 31*, 87–112.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. In *Proc. of the 1st ACM international workshop on wireless sensor networks and applications* (pp. 88–97).

Martinetz, T. M., & Schulten, K. J. (1991). A 'neural gas' network learns topologies. *Artificial Neural Networks, 1*, 397–402.

Montes, A. F., Ortega, J., Venzala, J. S., & Abril, L. G. (2014). Software reference architecture for smart environments: Perception. *Computer Standards and Interfaces, 36*, 928–940.

Nesselrath, R. (2013). Towards a cognitive load aware multimodal dialogue framework for the automotive domain. In *Proc. of the 9th international conference on intelligent environments* (pp. 266–269).

Preuveneers, D., den Bergh, J., Wagelaar, D., Georges, A., Rigole, P., Clerckx, T., et al. (2004). Towards an extensible context ontology for ambient intelligence. *Lecture notes in computer science* (Vol. 3295, pp. 148–159). Springer.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. Canada: Wiley Interscience Publication.

Römer, K., Schoch, T., Mattern, F., & Dübendorfer, T. (2004). Smart identification frameworks for ubiquitous computing applications. *Wireless Networks, 10*, 689–700.

Rueangsirarak, W., Atkins, A., Sharp, B., Chakpitak, N., & Meksamoot, K. (2012). Fall-risk screening system framework for physiotherapy care of elderly. *Expert Systems with Applications, 39*, 8859–8864.

Shahdi, S. O., & Bakar, S. A. R. A. (2012). Face recognition: Robust approach under varying and low resolution head poses. *International Journal of Imaging and Robotics, 7*, 70–87.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Blackwell Publishing.

Takemura, Y., & Ishii, K. (2011). Auto color calibration algorithm using neural networks and its application to robocup robot vision. *International Journal of Artificial Intelligence, A11*, 368–383.

Tan, W. R., Chan, C. S., Yogarajah, P., & Condell, J. (2012). A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics, 8*, 138–147.

Tang, D., Botzheim, J., Kubota, N., & Yamaguchi, T. (2013). Estimation of human transport modes by fuzzy spiking neural network and evolution strategy in informationally structured space. In *Proc. of the IEEE international workshop on genetic and evolutionary fuzzy systems* (pp. 36–43).

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior, 38*, 75–84.

Wu, L.-H., & Hsu, P.-Y. (2011). Embodied cognition of information visualization: Human–computer interaction with six degrees of freedom in movement of an information space. *Expert Systems with Applications, 38*, 10730–10736.

Xiao, Y., Yuan, J., & Thalmann, D. (2013). Human-virtual human interaction by upper body gesture understanding. In *Proc. of the 19th ACM symposium on virtual reality software and technology* (pp. 133–142).

Yorita, A., Botzheim, J., & Kubota, N. (2013). Emotional models for multi-modal communication of robot partners. In *Proc. of the 2013 IEEE international symposium on industrial electronics*.

Zhang, L., Jiang, M., Farid, D., & Hossain, M. A. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications, 40*, 5160–5168.