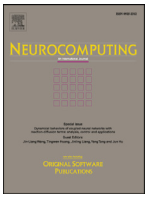
Neurocomputing 397 (2020) 464–476

Contents lists available at ScienceDirect

Neurocomputing

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



New telecare approach based on 3D convolutional neural network for estimating quality of life

Satoshi Nakagawa*, Daiki Enomoto, Shogo Yonekura, Hoshinori Kanazawa, Yasuo Kuniyoshi

University of Tokyo, Tokyo, Japan

ARTICLE INFO

Article history:
Received 15 May 2019
Revised 23 August 2019
Accepted 11 September 2019
Available online 31 March 2020

Keywords:
Human agent interaction
Communication system
Deep learning
Quality of life estimation
Welfare

ABSTRACT

Quality of life (QoL) is an effective index of well-being, including physical health, aspect of social activity, and mental state of individuals. A new approach that uses a deep-learning architecture to estimate the score of a user's QoL is presented. This system was built using a combination of a 3D convolutional neural network and a support vector machine for multimodal data. In order to evaluate the accuracy of the estimation system, three experiments were conducted. Before these experiments, ten hours of audio and video data were collected from healthy participants during a natural-language conversation with a conversational agent we implemented. In the first experiment, the QoL question-answer estimation experiment, the accuracy of "Physical functioning," which is one of the eight scales that constitute QoL, reached 84.0%. In the second experiment, the QoL-score-regression experiment, in which the scores of each scale were directly estimated, the distribution of the difference between the actual score and the estimated results, known as error, was investigated. These results imply that the features necessary for QoL estimation can be extracted from audio and video data, except for the "Mental Health" domain. One of the reasons why it was difficult to estimate the "Mental Health" scale may be that the learning framework could not extract an appropriate feature for estimation. Therefore, we estimated "Mental Health" by focusing on eye movement. From the result, it was proven that estimation is possible, and the proposed system using multimodal data demonstrated its effectiveness for estimation for all eight scales that constitute QoL and for extracting high-dimensional information regarding the QoL of a human, including their satisfaction level towards daily life and social activities. Finally, suggestions and discussions regarding the plausible behavior of the estimation results were made from the viewpoint of human–agent interaction in the field of elderly welfare.

1. Introduction

Elderly individuals are increasing in number. The global population of those who are 60 years old and above is about 962 million in 2017, which is more than twice the population in 1980. This number is expected to increase to 2.1 billion by 2050 [1]. It is reported that, on average, one in five women and one in ten men live alone. It is considered challenging to prevent incidents such as illness, accidents, and injury among elderly individuals who live alone and do not visit healthcare facilities because of problems such as shortage of caregivers or medical staff. Furthermore, should an incident occur, taking immediate action and appropriate measures may be difficult. Therefore, in recent years, monitoring systems for elderly individuals have been implemented both by local governments and in the private sector. The introduction of monitoring robots is an example of this trend. Estimating the well-being of the subjects is indispensable for human–agent interaction (HAI) with robots for the welfare of elderly individuals. Here, quality of life (QoL) is useful as an indicator, not only for measuring human physical suffering, but also for managing mental and social activities in a comprehensive manner. Many studies in the field of welfare aim to improve QoL; however, a number of problems concerning QoL exist.

It is often reported that the introduction of monitoring systems such as robots improves qualitative QoL such as happiness, which is not assessed quantitatively. For instance, improvement in QoL is attributed either to improvements in the operability of the instrument or to an increase in the number and/or duration of the conversations. This is despite the existence of an evaluation method that treats QoL quantitatively. Although important, there has, thus far, been no discussion on the relationship between

* Corresponding author.
E-mail address: nakagawa@isi.imi.i.u-tokyo.ac.jp (S. Nakagawa).

https://doi.org/10.1016/j.neucom.2019.09.112

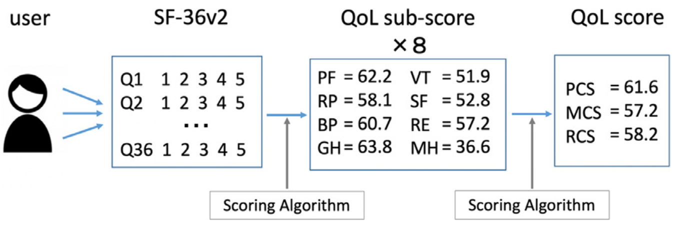


Fig. 1. Method of measuring QoL using SF-36v2.

HAI and each element of QoL. Furthermore, given that current QoL evaluations use questionnaires, the evaluations are not easy, and real-time property is not guaranteed.

Therefore, the goal of this research is to construct a QoL estimation system through natural interaction between human and agents, based on the idea that the requirement specification for HAI includes the following two points: first, QoL estimation must be performed via noninvasive and natural interactions; and second, information on QoL must be extracted and used to modify the agent's behavior based on the QoL state of the user. In this study, we refer to SF-36v2 [2] for the QoL index. According to this index, QoL consists of eight subscales: "Physical functioning (PF)," "Role physical (RP)," "Body pain (BP)," "General health (GH)," "Vitality (VT)," "Social functioning (SF)," "Role emotional (RE)," and "Mental Health (MH)." The questionnaire provided in SF-36v2 consists of 36 questions; each question is answered by choosing one out of three to five choices. The answers from the questionnaire are fed into a scoring algorithm, with the score of each scale calculated as shown in Fig. 1. Furthermore, the scores of these scales can be used to calculate the Physical component summary score (PCS), Mental component summary score (MCS), and Role / Social component summary score (RCS). The calculated QoL score can also be compared with the national standard value.

In order to realize natural-language conversation in HAI, we construct a communication system with a conversational agent using voice input/output that includes Open JTalk, Chat API, and morphological analysis. The agent replies in response to the user's remark and gathers video and audio of the user's upper body. The information that is extracted through the natural-language conversation is used for QoL estimation. In order to estimate QoL, deep learning is used on one of the models for three-dimensional convolutional neural network (CNN), Convolutional 3D (C3D) is applied to the videos, and support vector machine (SVM) is applied to the audio data. For accuracy evaluation of the model, the leave-one-out method [3] is applied. In this method, one sample from all the data is assigned to be test data, and the remaining data samples are used as training data. In this study, three types of estimation experiments are conducted. The first experiment is an estimation of each answer from the QoL questionnaire, and the second one is a score-estimation experiment on the eight subscales. An accuracy evaluation is performed on each experiment. From these two experiments, it is found that it is difficult to estimate "Mental Health," which is one of the scales that constitute QoL, and it is hypothesized that one of the causes for this is that features related to eye movement cannot be extracted by the framework. Therefore, for the last experiment, we focus on eye movement and estimate "Mental Health" by improving the CNN architecture and learning the time-series data.

In this paper, we propose QoL estimation through natural-language dialogue as a method of understanding the user through interaction. We conclude that discussions on relevant matters, such as providing appropriate support in accordance with the QoL of the user, will begin for the first time from this estimation result, assuming that these results are sufficient for judging the development of interactions and behavior selection based on user understanding. The novelty of this research is that we select QoL estimation instead of emotion estimation for estimating human condition. QoL is an index that comprehensively deals with a variety of aspects, and because the various scales can be quantified, QoL can output more information, when compared with emotion recognition via facial expression analysis. In addition, because we conduct experiments with consideration for actual social operations, the data are not collected deliberately; rather, information is obtained from natural situations such as conversations with the agent, making our method more practical. Therefore, there is also novelty in our proposal of a QoL estimation method that can be used in practical situations. Furthermore, in the field of welfare for the elderly, a number of watching robots have been researched, developed, and utilized. However, only qualitative QoL improvement has been claimed. It is also unclear which scales change and how HAI works with individual QoL scales. Through the introduction of this QoL estimation system to robots that are active in the welfare for the elderly, discussing the relationship of the individual scales with the interaction of the monitoring agents becomes possible and becomes a starting point for discussions on raising overall QoL, which is the overall goal of the field of welfare for the elderly.

The rest of the paper is organized as follows. Section 2 describes other research related to monitoring systems and methods for estimating emotions, such as a framework using 3D convolutional neural networks. Section 3 describes our new approach, which consists of a natural-language dialogue system and interpersonal experiments that are conducted before QoL estimation. Section 4 introduces two learning and estimation experiments with a framework implemented for QoL estimation. We also describe the results for these experiments, along with discussions of these results. Section 5 introduces an experiment for the estimation of "Mental Health," which is otherwise difficult for the two methods described in Section 4. We estimate "Mental Health" by extracting eye-movement information. Section 6 summarizes this research from the perspective of deep learning and elderly welfare and describes future directions for this research.

2. Related works

In the field of HAI, there are already many robots that interact with humans. The seal-type robot Paro [4] is a therapy robot certified by the Guinness Book of World Records as the most therapeutic robot in the world. Responding to people's speech is expected to promote human mental health. While Paro performs one-way interaction, robots that can utilize advanced communication abilities are able to promote communication with other people and potentially improve the user's social life functions. Examples of such advanced communication robots include Papero [5] and Palro [6]. Papero is a robot that realizes relationship building via two-way interaction. Palro focuses on the creation of natural dialogueues and realizes more intimate conversations between robots and people.

Furthermore, in the field of HAI, it is necessary to estimate the human's state correctly. Regarding human-pose estimation, Yao et al. [7] proposed a mutual context model to jointly model objects and human poses in human–object interaction activities and showed that their model succeeded in detecting objects, estimating human poses, and classifying human–object interaction activities. Cao et al. [8] proposed real-time multi-person 2D human pose estimation using part affinity fields, which is a set of 2D vector fields encoding the position and orientation of limbs in the input image.

With regard to emotion estimation, there are vision-based, audio-based, text-based [9,10], EEG-based [11–13], and multimodal approaches [14]. Because the face plays an important role in both emotional expression and perception during communication, many

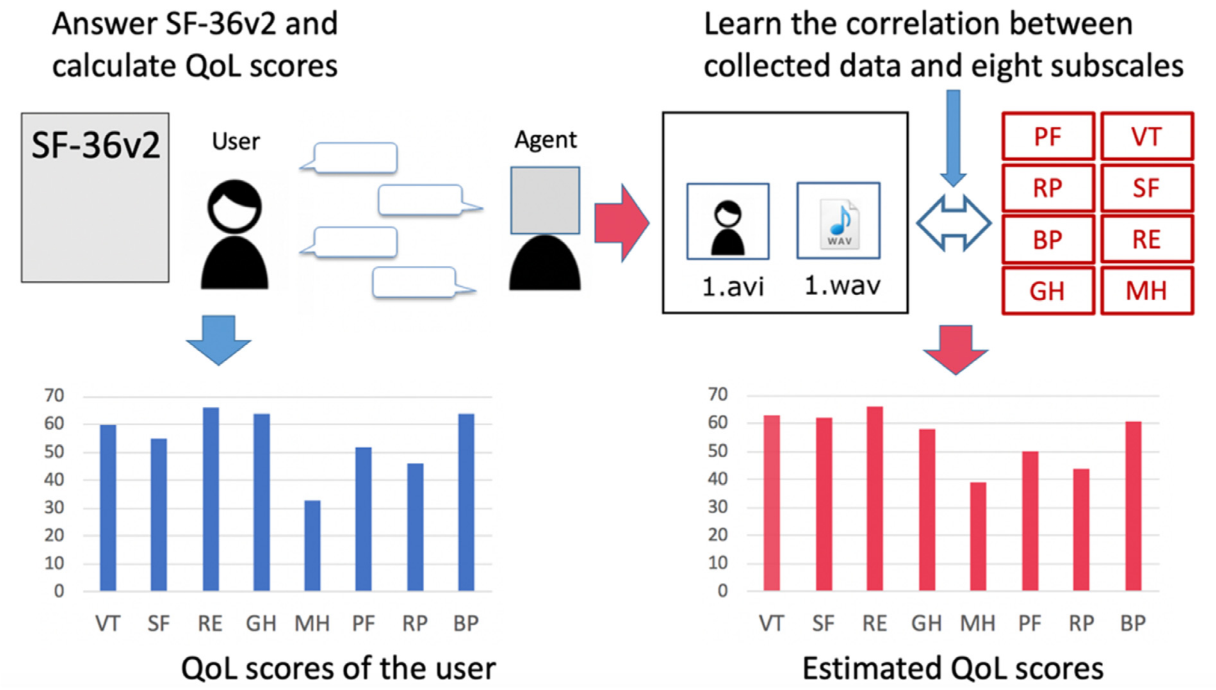


Fig. 2. QoL of the individual can be calculated by answering SF-36v2. Through communication with the individual, the agent gathers video and audio data and learns about the correlation between the data and the eight subscales that constitute QoL. The agent will try to estimate the QoL scores after the learning process.

vision-based studies focus on facial-expression analysis. Emotion estimation via face analysis primarily uses two methods. One method uses 2D spatiotemporal facial features, and the other uses 3D spatiotemporal facial features. Concerning 2D spatiotemporal facial features, Chang et al. [15], Pantic and Bartlett [16], and Kotsia and Pitas [17] proposed geometric-feature-based methods. Ren and Huang [18] proposed an automatic facial-expression learning method by extracting feature points on the face. Barakova et al. [19] proposed a genetic programming (GP) approach and dealt with the interpretation of emotions identified from facial expressions in context. As for 3D spatiotemporal facial features, Chang et al. [20] and Wang et al. [21] used 3D expression data for facial-expression recognition.

Recently, Tran et al. [22] found that 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets and proposed a type of 3D convolutional network called Convolutional 3D (C3D). C3D is an approach that performs convolution in three dimensions (combining two spatial dimensions and one temporal dimension). Fan et al. [23] presented a video-based emotion recognition system combining C3D and recurrent neural networks (RNN).

With regard to audio-based emotion recognition, Scherer [24] and Kwon et al. [25] studied the effect of emotion on voice and speech and found that certain emotions affect pitch, energy, speech rate, etc. Vidrascu and Devillers [26] showed the relationship between emotion and speech disfluencies such as silent pauses.

Furthermore, for agents that interact with users, it is important to estimate emotions that occur spontaneously at the time of communication, rather than emotions expressed intentionally. Tian et al. [27] argued that most of the existing research on facial-expression recognition is based on intentional and often exaggerated facial expressions. Cohn [28] investigated the difference between spontaneous and intentional facial expressions. In our research, we perform estimation using spontaneous expressions of emotion.

There are also several methods of communication aimed at collecting video and audio data, such as human–human conversation [29] and Wizard of Oz scenarios [30]. However, in HAI, robots need to communicate autonomously. Therefore, this study emulates an autonomous communication situation by using a chat API to realize a complete human–agent conversation.

3. Proposed approach

In this paper, we propose a new QoL evaluation approach. Therefore, before conducting QoL estimation, we implement a conversation agent with a natural-language dialogue system and collect data, to be used for machine learning, through an interpersonal experiment. In the experiment, the agent records a process in which an individual has a natural-language conversation with the agent and outputs the features obtained from the movements of the upper body as an .avi file and the audio data as a .wav file. An .avi file and a .wav file are created for each conversation. Ultimately, we aim to allow agents to learn the relationship between the aforementioned information and the eight elements that make up QoL, so that they can be estimated. With regard to the training data for learning, the results of the answers from the QoL indicator SF-36v2, which are obtained just before the user communicates with the natural-language conversation agent, are used. In QoL measurement, it is possible to calculate the score using a scoring algorithm that applies the Japanese national standard value. Fig. 2 shows the entire configuration of the data-collection part. The following paragraphs provide details about the natural-language dialogue system and interpersonal experiment, which are important in the proposed approach. Fig. 3 shows a flowchart of the overall frame design of our research.

3.1. Natural-language dialogue system

First, we create a conversational agent on the computer. The agent consists of the following four main parts: a voice-input part, for recording the user's voice; a voice-recognition part, for converting the recorded voice into a character; a conversation part, for generating appropriate responses to the user; and a sound-output part, for outputting the reply via speech synthesis.

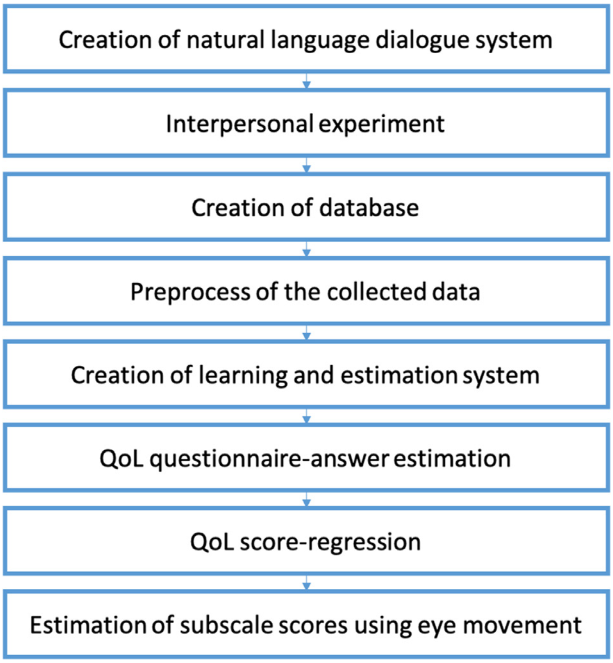

Creation of natural language dialogue system
Interpersonal experiment
Creation of database
Preprocess of the collected data
Creation of learning and estimation system
QoL questionnaire-answer estimation
QoL score-regression
Estimation of subscale scores using eye movement

Fig. 3. Flowchart of overall frame design.

3.1.1. Voice-input part

For the voice-input part, assuming a one-to-one conversation between the agent and the user, the recording time is set to start at the end of the agent's speech and to finish at the end of the user's speech. Therefore, the recorded content consists of the entire natural-conversation speech by the user. Hence, both the time of utterance and silence at the time of utterance can be collected. In the speech analysis, it is possible to utilize both the prosody obtained via frequency analysis of the speech data obtained here and the prosodic feature for silence.

3.1.2. Voice-recognition part

Regarding the speech recognition part, we use the speech recognition API provided by docomo developer support [31]. This speech recognition API is able to achieve a high recognition rate by using a substantial voice database and an advanced recognition algorithm. In this part, text transformation of the user's speech is performed in real time.

3.1.3. Conversation part

In order to achieve a smooth dialogue with the user, the agent does not only need to return an appropriate response to the utterance from the user; it also needs to continue the conversation. Therefore, this time, we include a chat dialogue API [31] into the agent, making it possible to include the latest information into the conversation.

3.1.4. Sound-output part

For the purpose of realizing communication similar to a conversation between individuals, a mechanism voicing out the agent's responses is used. In order to output the character string as voice, it is necessary to synthesize the speech. To generate the voice output, Open JTalk, which is a hidden Markov model text-to-speech synthesis system, is used. This system can freely change the quality and pitch of the voice based on specific Japanese sentences. In this experiment, we adopt a format that uses Open JTalk to generate the chat-dialogue-API responses to user utterances.

The system diagram of the conversation agent constructed in this study is shown in Fig. 4.

3.2. Interpersonal experiment

3.2.1. Setting

The subjects for the experiment are 14 adults (seven males and seven females). All subjects participate after receiving explanations regarding the experiment beforehand and after confirming consent based on free will. After responding to SF-36v2, the subject sits in front of the computer in which the natural-language dialogue system is implemented, in order to start the experiment. When we run the natural-language dialogue system, the agent first greets the subject with a "Hello." After that, the subject and the agent take turns to speak. (For example, "Hello," "Hello," "It has been getting warmer recently. What are you planning to do today?" "I am going to have lunch with my friend after this.") During the experiment, there are no restrictions against the content of conversation and the total conversation time, except for the condition that the number of conversations needs to be 30 or more, with each conversation being limited to approximately 10 s.

Fig. 5.

3.2.2. Collecting data

Through free communication with the user, the agent gathers the data, as shown in Fig. 6, from the subject. By collecting and learning the time-series data from videos rather than images, we expect to observe the change in the information projected on the face of the user with respect to time, which cannot be observed from only a single still picture, while simultaneously be able to use the data for learning.

The collected answer results to SF-36v2 are subjected to the following scoring algorithm. The score for each of the 8 scales is calculated. First, the scores for selected answers to questions belonging to each scale are summed. Let x be the raw score of the sum of questions belonging to the scale, x_{min} be the minimum possible score that the scale can take, and x_{range} be the expected raw-score range. The score is then normalized to a scale ranging from 0 to 100, as follows.

$$score\ of\ scale = (x - \ x_{min}) /\ x_{range}$$

From the aforementioned process, we obtain scores for the 8 scales, which are referred to as PF, RP, BP, GH, VT, SF, RE, and MH. With SF-36v2, scoring based on national standard values is possible. Therefore, the adjusted score for each scale is converted, so that the average value representing the national standard values is 50 points, and the standard deviation is 10 points. This is achieved through the following computation. The average of the national standard value is subtracted from the score, in the range of 0 to 100 points, and divided by the standard deviation.

$$PF\_Z = (PF - 89.13446)/13.85045$$

$$RP\_Z = (RP - 89.24007)/18.80773$$

$$BP\_Z = (BP - 73.77098)/22.39818$$

$$GH\_Z = (GH - 62.91007)/18.76562$$

$$VT\_Z = (VT - 62.82787)/19.46255$$

$$SF\_Z = (SF - 86.38347)/19.40441$$

$$RE\_Z = (RE - 87.84637)/20.01521$$

$$MH\_Z = (MH - 71.60598)/18.62983$$

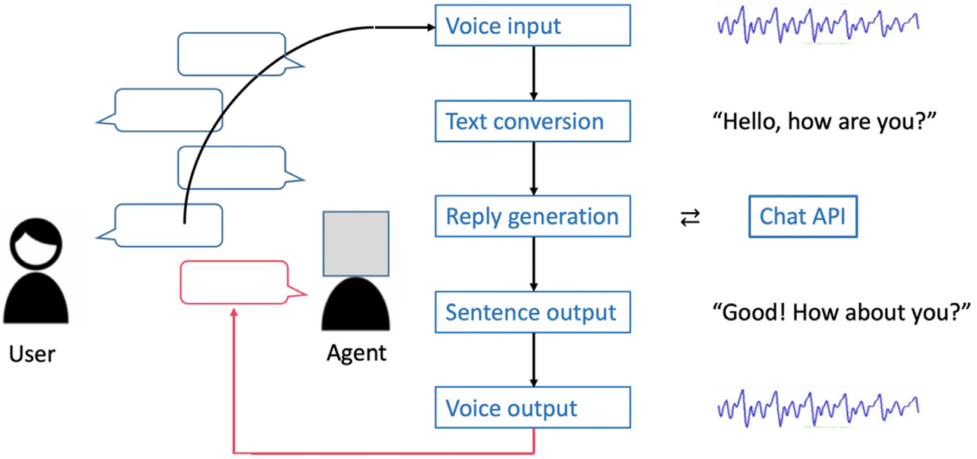


Fig. 4. System diagram of the conversation agent.

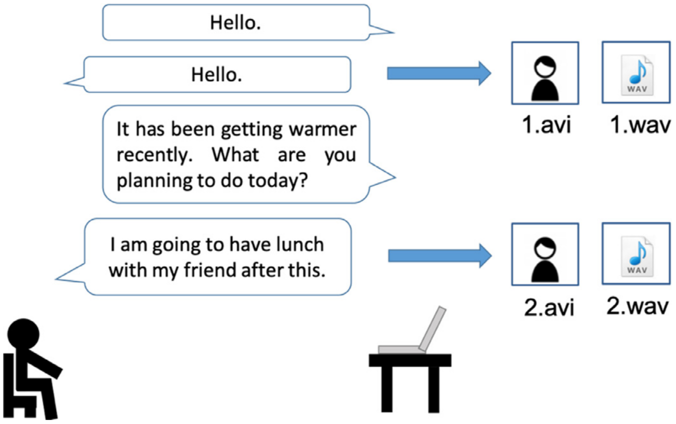

Hello.
Hello.
It has been getting warmer recently. What are you planning to do today?
I am going to have lunch with my friend after this.
1.avi
1.wav
2.avi
2.wav

Fig. 5. System diagram of the interpersonal experiment. Every time the subject speaks, the natural-language dialogue system which is implemented in the computer collects video and audio data.

Conversion is made so that the average is 50 points and the standard deviation is 10 points.

$$\begin{aligned}&score\ of\ scale\ based\ on\ national\ standard\ value\\&\quad=(scale\_Z*10)+50\end{aligned}$$

With this process, the scores of the eight scales that constitute QoL are calculated.

3.2.3. Preprocessing of audio data

For each question, the subjects who provide the same answers are classified as a group; this group is labeled with the answer number. Regarding speech recognition and estimation, while methods using neural networks and methods using support vector machine (SVM) both exist, established methods do not. Therefore, we classify speech features, which are extracted using Talkbox Scikit, using SVM, which is a method of constructing two-class pattern discriminators by using the simplest linear threshold element as a neuron model. From the training data, the parameters of the linear threshold element are learned for maximizing the margin. Classification is not limited to linear classification of two classes; it is also possible to combine multiple SVMs or to use a kernel learning method to construct a nonlinear discriminator (i.e., multiclass classification).

3.2.4. Preprocessing of video data

In this study, three-dimensional data (longitudinal and lateral directions of the image, and time) are collected and used as input data for machine learning, where C3D is implemented. For the input data, we convert the obtained video data into an appropriate

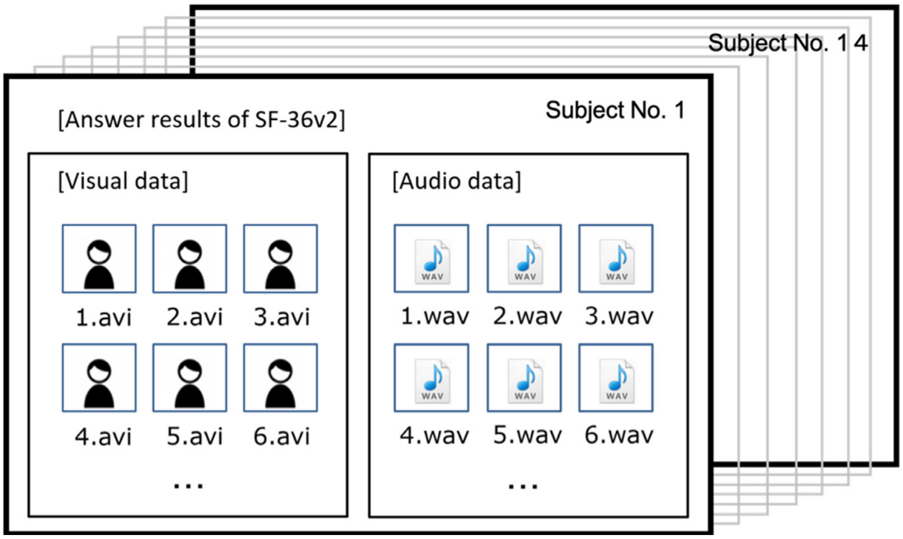


Fig. 6. Collected data. Includes the answers to SF-36v2, visual data, and audio data for each subject.

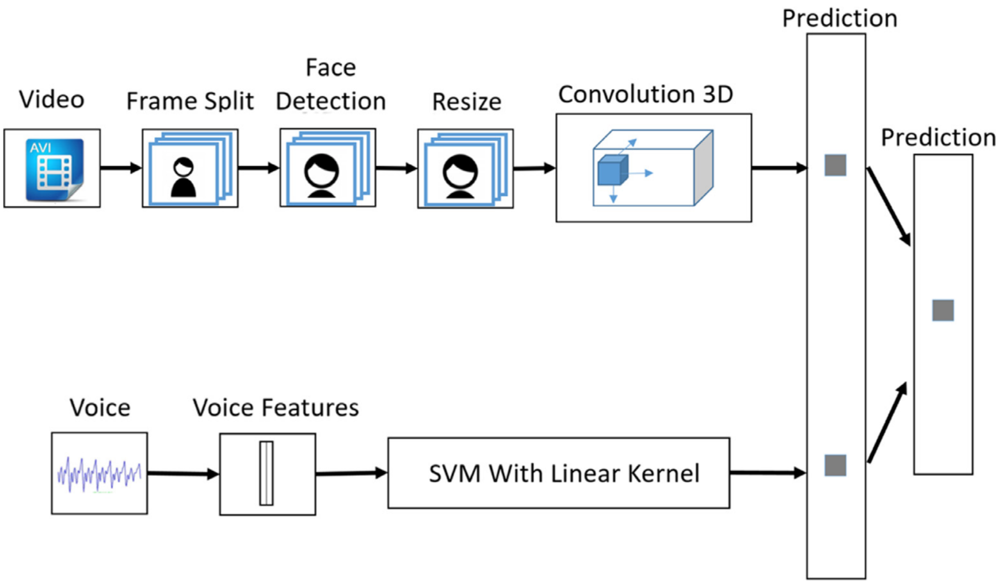


Fig. 7. Learning and estimation system.

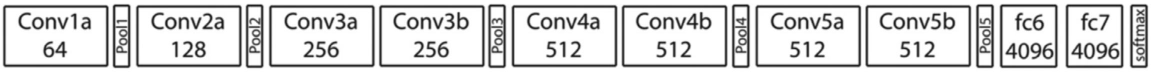


Fig. 8. Implemented architecture for the learning part, using C3D [22].

format as follows. To use the frames, which follow the time series, as input data, we divide the moving image into frames.

In addition, to observe the significance of the facial expressions of the user, only the facial part of the subject is extracted at the time we divide the data into frames. Given that the learning time is proportional to the frame size, we use a 50 × 50 pixels frame size. Furthermore, the lengths of the .avi files are unified to 100 frames for all subjects, and each pixel is assigned three values corresponding to RGB. Therefore, for each subject, we prepare and create 100 × 50 × 50 × 3 array data as input data. With regard to learning, Convolutional 3D is applied. Convolutional 3D is a convolutional neural network that applies a three-dimensional convolution of vertical dimension, horizontal dimension, and depth to three-dimensional input data. This means that the kernel itself that is used for convolution is three-dimensional. As a result of using this method, three-dimensional data holding time-series information of input data are obtained as output.

4. Learning and estimation experiment using 3D convolutional neural network

4.1. Learning and estimation system

The learning and estimation system is built as shown in Fig. 7. Subsequent to preprocessing the visual and audio data, each estimated result is integrated into a single output.

Fig. 8 shows the implemented C3D architecture. Three-dimensional convolution is performed in a total of eight layers. The max pooling layer is provided after a total of five convolutional layers: the first, second, fourth, sixth, and eighth layers. fc6 and fc7 are fully connected layers. Finally, the output layer is placed. At this time, all three-dimensional convolution kernels sizes are 3 × 3 × 3, which is reported to be the size with the highest accuracy in a deep-learning experiment using C3D [22]. The size of the first layer of the pool kernel is 1 × 2 × 2, while the size of the others is 2 × 2 × 2. A total of 4096 units can be output in the fc7 layer, but the final layer is implemented so that it has 5 dimensions, corresponding to the number of choices for each question in SF-36v2, or 8 dimensions, corresponding to the score outputs for the 8 scales of QoL.

Two kinds of learning experiments are conducted as follows.

4.2. QoL questionnaire-answer estimation

For estimating responses to questionnaires, experiments are set up for each question in SF-36v2 as follows. In order to classify which of the five options is a response, each set of video and audio data is labeled with a five-dimensional vector. For example, when the user chooses the 2nd option as the answer, the label for their video and audio data is [0, 1, 0, 0, 0], wherein 1 is substituted for the second element of the vector, and 0 is substituted for the other elements. We prepare 36 such labels, corresponding to the total number of questions in SF-36v2, and perform the learning process. After the learning process, wherein the estimation system learns the correlation between the labels and video and audio data, the system will try to output vectors with five elements, such as [0, 0.92, 0.08, 0, 0], against the test data; each element represents the possibility that its corresponding option is the answer. For example, [0, 0.92, 0.08, 0, 0] means the second option is most likely to be the correct answer.

4.3. QoL-score regression

To directly estimate the eight subscale scores, we design the system such that the output is an 8-dimensional vector corresponding to the eight subscale scores constituting the QoL. For example, one user's output set of data is [PF, RP, BP, GH, VT, SF, RE, MH] = [58.5, 55.7, 61.7, 62.8, 40.2, 57.0, 56.1, 43.8]. For the training data, such labels are attached to every set of data collected from the user. In this experiment, the system tries to output the scores for the eight subscales at once in the shape of vectors with 8 elements each. The challenge of estimating the answer to the questionnaire is a classification problem; however, in this experiment, given that the scale score is directly learned and output, it

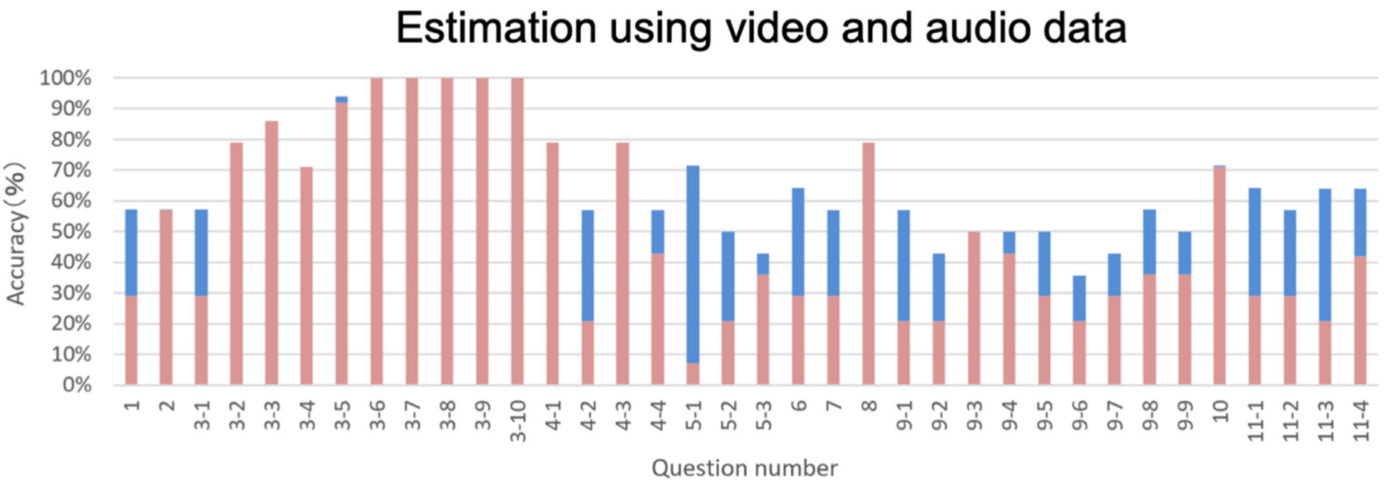


Fig. 9. Results of calculating the correct-answer rate of the estimation result for each question.

Table 1
Subscales with corresponding numbers of items and question numbers from SF-36v2.

Subscale	Number of items	Question number
Physical functioning	10	3–1 – 3–10
Role physical	4	4–1 – 4–4
Bodily pain	2	7, 8
General health	5	1, 11–1 – 11–4
Vitality	4	9–1, 9–5, 9–7, 9–9
Social functioning	2	6, 10
Role emotional	3	5–1 – 5–3
Mental health	5	9–2, 9–3, 9–4, 9–6, 9–8

is reduced to a regression problem. In evaluating the estimation accuracy of this model (for both answers and subscale scores), a leave-one-out method [3] is used.

4.4. Results and discussion of QoL questionnaire-answer estimation

4.4.1. Results

In the case of a classification problem, wherein the evaluation is based on which items are "correctly classified" or "incorrectly classified," it is possible to calculate the correct-answer rate of the evaluation, in order to estimate the accuracy of the model. The same operation was repeated for all 36 questions in SF-36v2. Each of these 36 questions were categorized into one of the eight subscales, as listed in Table 1. Question 2 was not included in the calculation of the eight subscale scores and was treated individually.

The graph in Fig. 9 shows the calculated correct-answer rate of the estimation result for each question. The vertical axis represents the correct-answer rate (%). The horizontal axis corresponds to the question number, and there are 36 questions in total. The red part of the graph represents the correct-answer rate when the questionnaire's average answer is taken as the estimation result. The blue part of the graph shows the degree of increase in the correct-answer rate from using this system.

The probabilities of the correct values being output from the motion-picture and audio data for the eight subscales using the estimation system are shown in Table 2.

4.4.2. Discussion

For question 9–6, which had the worst answer rate according to Fig. 9, the question that was asked was "Are you depressed?" which was included in the "Mental Health" scale of the eight subscales.

Table 2
Accuracy of subscales estimated from the motion-picture and audio data.

Subscale	Accuracy (%)	Subscale	Accuracy (%)
PF	84.0	VT	51.3
RP	68.0	SF	68.0
BP	67.8	RE	54.8
GH	62.3	MH	47.0

With regard to the reason for why the accuracy for this question was low, considering the existence of research on emotion estimation using speech [32,33], video [34], and both [23], it is likely that information about depression affects the voice and expressions of the user and can be observed by others.

Considering the studies on the measurement of heart rate from facial expressions using photoelectric volumetric pulse wave recording [35,36], on the relationship between heartbeat and head vibration [37,38], and on the relationship between emotion and heartbeat [39,40], it is possible to measure the psychosomatic state from the physical characteristics of an individual. In fact, the Emotion API provided by Microsoft Azure is already widely used in systems that classify facial expressions into eight basic emotions.

Furthermore, there were studies [41,42] on performing emotional analyses by analyzing vocal features, so these features are considered useful for classifying or estimating human emotions.

4.5. Results and discussion of QoL-score regression

4.5.1. Results

Fig. 10 shows a graph of the distribution of errors (absolute values), which visualizes the error between the estimation by the system and the actual result calculated from scoring algorithm of SF-36v2. When the "Mental Health" scale (median: 9.9, first quartile: 5.4, third quartile: 26) is compared with the other scales (median: 3.9, first quartile: 1.6, third quartile: 8.3), the "Mental Health" bar at the right end of the graph is particularly pronounced. Estimation of "Mental Health" is found to be the most difficult.

4.5.2. Discussion

As a result of the QoL-score regression experiment, 7 out of 8 scales can be estimated with small errors (median: 3.9, first quartile: 1.6, third quartile: 8.3). Because the learning-estimation architecture we implemented used time-series data, it contained more features than two-dimensional convolution using still images does.

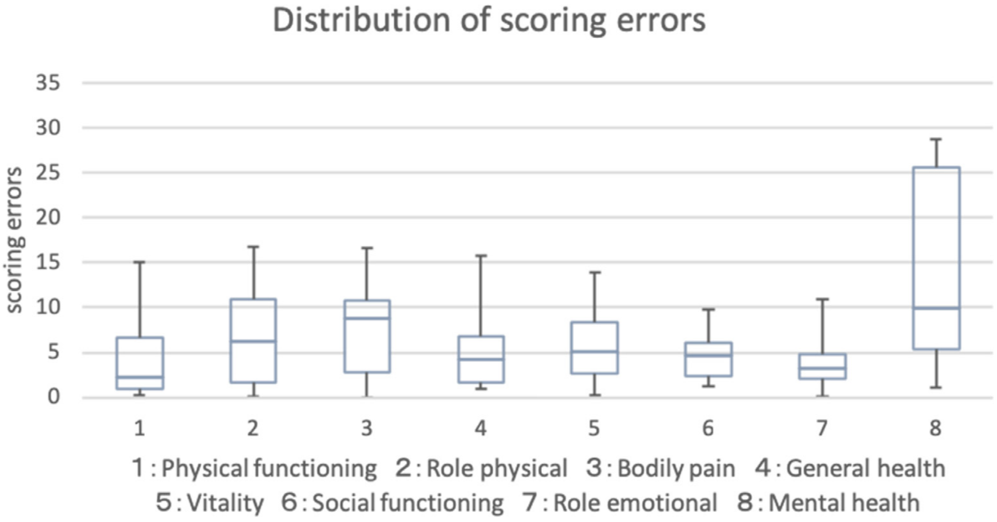


Fig. 10. Distribution of the errors (absolute value) between the actual scores and estimated scores.

This time, we extracted mainly the face part. Therefore, extraneous information such as head fluctuation was lost; features on the face surface were collected instead. From this experimental result, features that are effective for QoL estimation were included in the face, and our implemented learning-estimation architecture using C3D succeeded in extracting high-dimensional information regarding QoL and in estimating the scores for QoL. Furthermore, while the collected videos were made to be 50 × 50 in size this time, estimation was shown to still be sufficiently possible. We were able to reduce the learning cost by making the image size smaller, and at the same time, succeeded in estimating with higher accuracy by using time-series data. This experiment demonstrated the effectiveness of applying three-dimensional convolution to time-series data when estimating QoL, which is one of the measurements of human state. This approach can also be applied to and show effectiveness in estimations of other human states, such as facial-emotion recognition (FER). Therefore, we can say that the experimental results reveal the effectiveness of our new telecare approach based on 3D convolutional neural network in estimating the quality of life.

In the regression problem, as in the classification problem, it was found that estimation of “Mental Health” was the most difficult. The following four explanations can be cited as possible reasons.

- Failure of the framework we implemented to extract features that are efficient for the estimation for the “Mental Health” scale
- Discrepancy between self-evaluation of emotion and evaluation by others
- Discrepancy between subjective emotion in the questionnaire response before the experiment and emotion expressed during the natural-language conversation with conversational agent
- Changes in emotion

Regarding the first reason, because there are studies that estimate depression from facial expressions, information on emotions such as depression is considered to be expressed in human expressions. Several studies [43] have argued that gaze patterns are one of the useful features for estimating human state. However, this time, we directly used as input data the video data of the subject talking to the agent. Therefore, although this video contained information on eye movement, it also contained many other variables, so there is a possibility that it failed to extract features that were necessary for estimating “Mental Health.”

Regarding the second hypothesis, question 9–6 asked the individual for his/her current emotion; thus, it is possible to think that it was only concerning this part where the problem was the same as with emotion estimation. In general, the evaluation method in emotional analysis can be divided into two types: one is self-assessment, which is a subjective evaluation by the individual himself / herself, and the other is objective evaluation by a third party. According to the result in [44], in which the basic emotions were labeled both via self-evaluation and via objective evaluation of the voice including emotion, and in which the ratios of disagreement between them were compared, the error of emotional classification via evaluating others is larger than that via self-evaluation, for sadness, anxiety, and distress, which are considered to be emotions similar to a “depressed mood.” In the interpersonal experiment described in this paper, the subjects themselves were asked to answer the questionnaire SF-36v2 beforehand; subsequently, an estimation was performed using the data collected by the conversational agent during the conversation. Therefore, the data collected in the first half of the experiment was from self-evaluation, while the data analysis in the second half of the experiment was akin to evaluation by third party. Moreover, the natural-language conversation was made with a conversational agent. It is inferred that the error between self-evaluation and third-party evaluation became even larger than the result of a previous study [44]. This is because our method is different from the conventional method of emotion estimation, in which specific sentences and emotions are presented to subjects, who express themselves in a way that they are most likely to express the given feeling. Hence, it is considered difficult to correctly estimate “depressed mood” because of the discrepancy in the nature of emotion evaluation in this analysis. For the same reason, it is possible to explain why the correct-answer rate for problem 9–2, “Quite nervous or not,” is less than 50% and is the second lowest. This can be reduced to a problem equivalent to emotion estimation, for the same reason.

Next, we discuss the third hypothesis, which is regarding the discrepancy between subjective emotions when answering the questionnaire and the emotions expressed during the natural-language conversation with the conversational agent. The reason behind this discrepancy may be that this experiment used a new method, which is not performed in daily life, causing the subjects to feel resistance towards the experiment. There was also the influence caused by the HAI. Regarding the creation of relationships in HAI, Ono et al. [45] used psychological experiments to confirm that having an interaction with an anthropomorphic agent on a

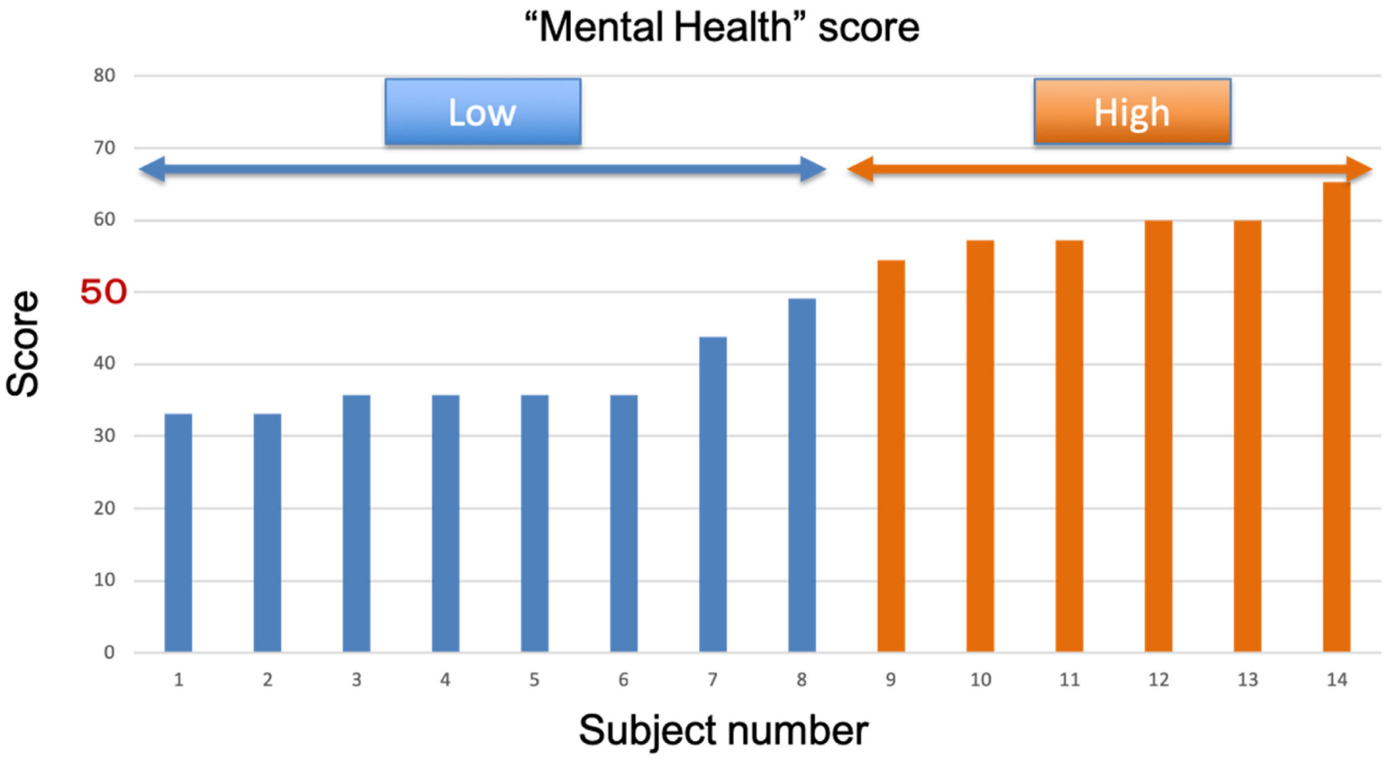


Fig. 11. "Mental Health" scores of 14 subjects.

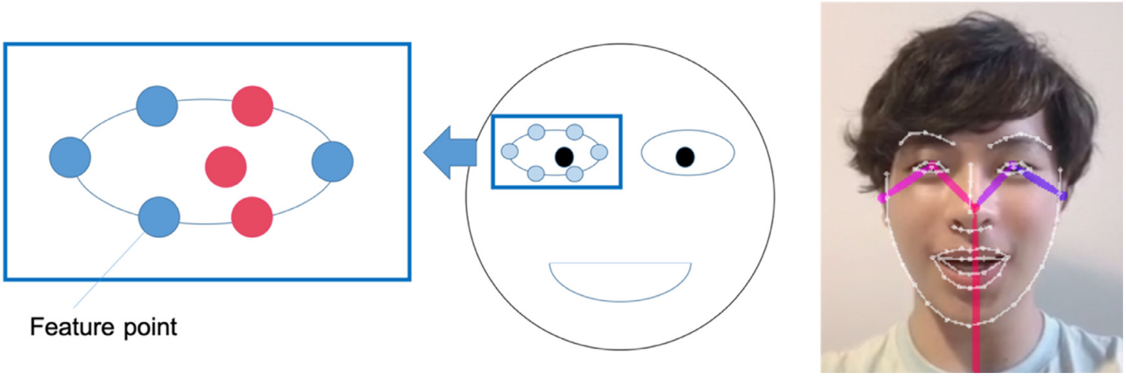


Fig. 12. Feature points that OpenPose face estimation extracts. The three points in red are the points that we used for the learning and estimation experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

screen causes individuals to gain confidence and affinity for the agent. Medley and Mitzel [46] showed that in familiar places, not only does the group's ability to solve problems increase, but also does the relationship between people and self-evaluation become higher, which implies that the environment where the experiment is done affects the potential and states of the subjects. It is inferred that in our experiment, there is a possibility of tracing the process in which an affinity field was created, and self-evaluation results were high as a result of having a sense of trust and affinity, due to the formation of a friendly relationship between the natural-language conversational agent and the subjects. Therefore, we cannot deny the possibility that the feelings that existed before the experiment began had changed and were different from the emotions during natural-language conversation with the conversational agent.

Before explaining the fourth reason, we explain how objective observation becomes possible as a result of the body receiving sensory input and processing. Information on emotion is conveyed to the amygdala and then further to the hypothalamus. The hypothalamus is the center of autonomic nervous function and hormone secretion and causes fluctuations in gastrointestinal movement and heartbeat. With regard to fear, information is transmitted from the amygdala to the midbrain, and as a result, responding behavior such as trembling is caused. These movements caused by the hypothalamus and the midbrain are involuntary and also affect vocal-cord oscillations. However, because individuals tend to have diverse interactions with the outside world due to their physicality and are also subjects themselves at the same time, organs in the body constantly receive the influence of environmental changes and physical changes, and feelings are not always constant. While it can be said that the feelings of people are always fluctuating, by maintaining the endocrine system and the nervous system, the homeostasis of emotion is also maintained.

From the previous discussion, we can conclude that emotions are constantly fluctuating, and because of the influence of both fluctuations of the emotions themselves and internal processes attempting to maintain the internal conditions of the individual, we infer that the characteristics of voice and expressions, that are in

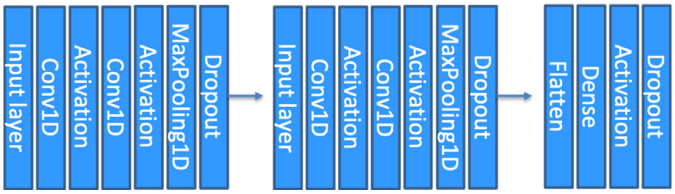


Fig. 13. Learning and estimation system.

a form from which observations can be made, are causing similar fluctuations.

4.6. Summary of QoL-estimation experiment

4.6.1. Summary

From both of the experiments, QoL questionnaire-answer estimation and QoL-score regression, we concluded that it is difficult to estimate the “Mental Health” scale score. For this reason, we produced four hypotheses, which are explained in Chapter 4.5.2.

4.6.2. Comparison between classification problem and regression problem

In our approach, for the purpose of evaluating the estimation accuracy, we conducted two kinds of experiments. One experiment was to estimate answers to questions in QoL questionnaire SF-36v2 and calculate eight types of scores of a lower scale that constitute QoL using scoring algorithms based on the results, and the other was to estimate the score of each subscale of QoL. The former can be called an indirect estimation because it went through the step of estimating the answer to the questionnaire before calculating the score using a scoring algorithm, while the latter can be called a direct estimation because it directly estimated the score from the input data. In the indirect estimation, the amount of information was dropped to 36 5-dimensional vectors in the middle stage, such that information loss, which should be useful on a lower scale, occurred.

Nevertheless, both experiments unexpectedly output similar results. This means that there is a possibility to perform behavior equivalent to a case where all information is directly output as estimation results, despite reducing the amount of information. Although it is not easy to argue from imagination about what each estimation system was focusing on in the moving-image and voice data, it is necessary to determine which one can output the score more accurately, or whether both can output the score after the number of subjects and the amount of data are sufficiently increased in the future. It is also necessary to create discussions based on actual results and theory.

5. Mental-health estimation using eye movement

5.1. Estimation of subscale scores using eye movement

A feature concerning QoL could have been overlooked when only video data was directly used as input for the two experiments, QoL questionnaire-answer estimation and QoL-score regression. Therefore, eye movement information was individually extracted, and QoL estimation was performed. Here, we examined one of the QoL subscales, “Mental Health” scale. Fig. 11 shows the “Mental Health” scores of 14 subjects. The national standard value for the QoL score was adjusted to 50 and the standard deviation to 10. Therefore, a score of 50 could be treated as a cutoff value separating the subjects into two groups, a low-scoring group and a high-scoring group. The reason for this classification was that it would be difficult to estimate the scores directly without classifying the groups.

In this experiment, OpenPose [8] was used in order to extract time-series patterns of eye movement. Relative positions (x-coordinate and y-coordinate) from the face center of the three points shown in the Fig. 12 were extracted, and their time-series data were created. With regard to the eye movement, when the behaviors of these three specific points were observed, it was expected that both the pattern of the direction to which the user was looking and the blink pattern could be extracted.

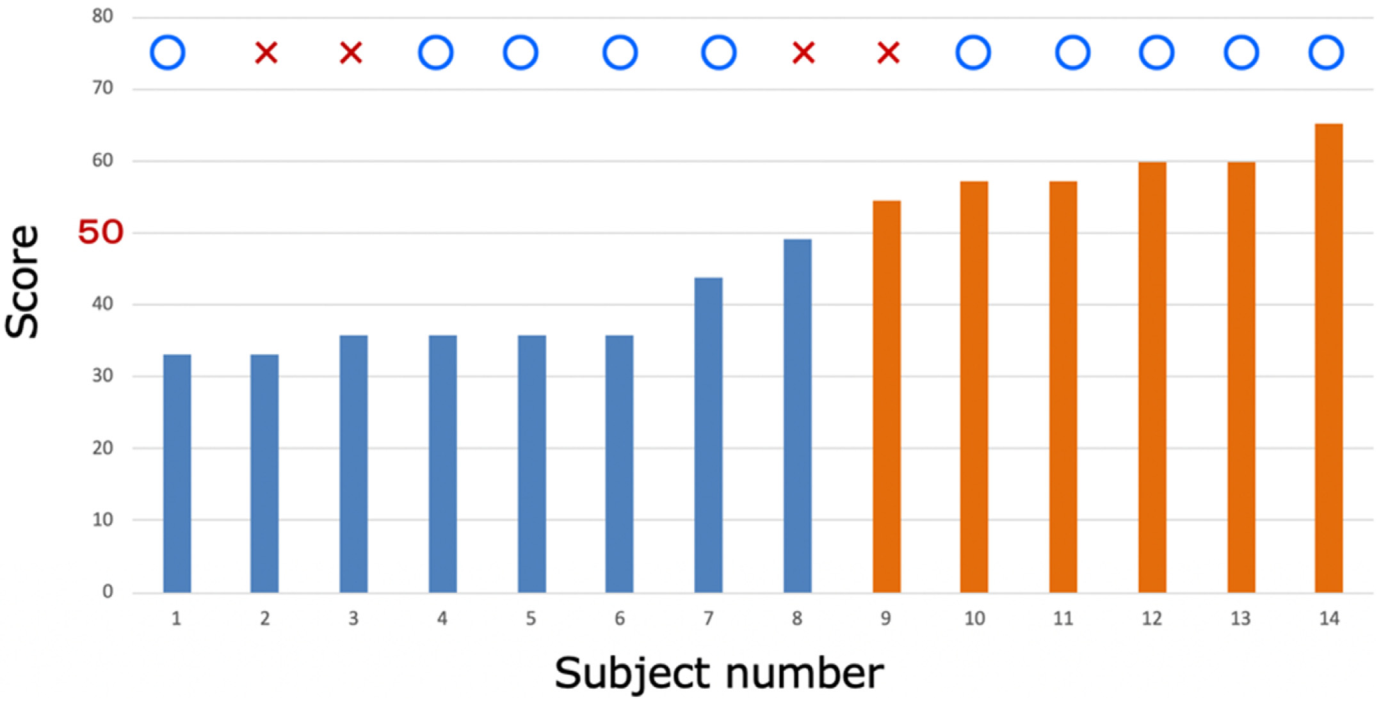


Fig. 14. Result of the estimation using eye movement. Circles and crosses above the graph indicate correct and incorrect estimations, respectively.

Table 3
Accuracy and precision of estimation for “Mental Health” using eye movement.

Accuracy	Precision
0.714	0.833

Input data were each at 60 frames in length and about 3 s in duration. We prepared 10 sets of data per person, created the CNN model shown in Fig. 13, and conducted the learning and estimation experiment. The leave-one-out method was used to evaluate the accuracy of the learning and classification system.

5.2. Results and discussion of the learning and estimation experiment using eye movement

5.2.1. Results

Fig. 14 shows the result of estimation. In the figure, a circle is drawn when the estimation result and the actual score coincide, while a cross is drawn otherwise. From this result, it was found that the subjects who judged that their “Mental Health” score was bad had a poor score, except for one subject who had a score close to the cutoff value. The accuracy and precision of the estimation are as shown in Table 3.

5.2.2. Discussion

The accuracy of the estimation based on eye movement was 71.4%. This result indicates that it is possible to estimate the “Mental Health” score from eye movement. In addition, from this result, it can be concluded that a reason why it was difficult to estimate the “Mental Health” scale in the two estimation experiments, QoL questionnaire-answer estimation and QoL score-regression, was that the scales for changes other than eye movement were large when moving images were directly input. Therefore, it was possible that information useful for QoL estimation could not be extracted properly. Implementing a framework for extracting eye movement would make it possible to estimate scores for all eight scales that constitute QoL and demonstrate the effectiveness of the proposed system.

6. Conclusions

In this study, we performed estimation of QoL of users through natural-language dialogue with a conversational agent implemented in a computer. The conversational agent learned spatiotemporal features using moving images and speech with mixed C3D and SVM structure. In the accuracy evaluation, we conducted a comparison between the actual questionnaire results and the system-estimated questionnaire responses, and with the scores for eight subscales as directly estimated by the system. As a result, the highest correct-answer rate for the former answer estimation was 84.0%; however, the estimation accuracy for the question regarding “Mental Health” was the lowest. With regard to the latter estimation, when the scores for the eight subscales were directly estimated, both the estimation error and the error variance were found to be small for the seven subscales, which excluded the “Mental Health” scale. As with the former answer estimation, “Mental Health” was also the most difficult to predict. Regarding the reasons why “Mental Health” could not be estimated, four points were considered in the discussion section of the previous chapter, along with related studies and solution proposals. Although all of these are hypotheses, we believe that it is possible to estimate with higher precision by improving on the problem associated with each of these hypotheses.

In order to test the first hypothesis, we used OpenPose to extract only the time-series data on eye movement from the video data collected during the communication with the conversational agent and learned the correlation between eye movement and the level of “Mental Health” using CNN. As a result, the accuracy of the estimation reached 71%. From this result, it was found that the moving images collected from the interpersonal experiment certainly included the aforementioned information on “Mental Health,” and it became possible to extract them by improving the learning-estimation structure. The first of the four hypotheses turned out to be correct, and the implementation of a method that enabled estimation of all eight scales that constitute QoL was achieved. As a conclusion, experimental results show the effectiveness of our approach.

In this study, we created a QoL estimation system. QoL is an effective index for the evaluation of physical health and mental state of individuals and societies. From the results, we concluded that it is possible to extract high-dimensional information regarding the QoL of a human, including their satisfaction level towards daily life and social activities, from multimodal data.

Through the introduction of this system to the monitoring robot system and communication robot, a new judgment criterion for deciding the action specialized for individual property and personality has been made, in addition to already developed systems, such as human detection and speech detection via constantly grasping the mind and body condition of the user. In addition, through the construction of a long-term relationship between a person and a robot, it is possible to estimate the current demand of the user via learning the patterns of mind and body, to allow the agent to autonomously learn what kinds of behavior cause what kinds of physical and psychological changes in the user, and to think about selecting the best action via prediction. We argue that the goal that HAI should aim for is to develop such temporal development in human interaction, which is important in the field of cognitive robotics [47], and in enriching the human mind, which is an important factor in the field of mental engineering [48]. Finally, numerous monitoring robots have been researched and developed and are widely used; however, only the improvement of QoL is claimed. Here, QoL, which should be evaluated quantitatively, is currently being regarded qualitatively as the meaning closest to happiness. Furthermore, the relationship between HAI and each scale has not been discussed.

The QoL-estimation system developed in this study is sufficiently useful and is the only method for various welfare-related robots to be evaluated via comparing how they function on each scale constituting QoL through human interaction. We hope that this topic will not be limited to monitoring robots and will be used as a common evaluation method for all robots that build interactive relationships with humans. Simultaneously, we also expect that this study will be the starting point of discussions to collectively raise all measures of QoL. In this study, we focused on understanding elderly welfare and conducted QoL estimation using 3D convolutional neural network, SVM. We hope that henceforth, approaches to HAI will consider the states and appropriate behavior. We look forward to discussions and research on how to understand users and on how to behave adaptably according to the user’s character and behavior.

Acknowledgment

This work was supported in parts by Next Generation Artificial Intelligence Research Center, The University of Tokyo and a donation from NVIDIA Corp.

References

[1] World population ageing 2017, United Nations, 2017. www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf (accessed August 22, 2019).

[2] S. Fukuhara, S. Bito, J. Green, A. Hsiao, K. Kurokawa, Translation, adaptation, and validation of the SF-36 health survey for use in Japan, J. Clin. Epidemiol. 51 (11) (1998) 1037–1044, doi:10.1016/S0895-4356(98)00095-X.

[3] E. Kocaguneli, T. Menzies, Software effort models should be assessed via leave-one-out validation, J. Sys. Softw. 86 (7) (2013) 1879–1890, doi:10.1016/j.jss.2013.02.053.

[4] T. Shibata, Research on interaction between human and seal robot, Paro, J. Robot. Soc. Jpn. 29 (1) (2011) 1–31.

[5] S. Ohnaka, T. Ando, T. Iwasawa, The introduction of the personal robot papero, IPSJ SIG Notes 37 (7) (2001) 37–42.

[6] H. Ninomiya, Introduction of the communication robot Palro and efforts in robot town Sagami, J. Robot. Soc. Jpn. 33 (8) (2015) 607–610.

[7] B. Yao, L. Fei-Fei, Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses, IEEE T. Pattern Anal. 34 (9) (2012) 1691–1703, doi:10.1109/TPAMI.2012.67.

[8] Z. Cao, T. Simon, S.E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[9] F. Ren, K. Xin, C. Quan, Examining accumulated emotional traits in suicide blogs with an emotion topic model, IEEE J. Biomed. Health 20 (5) (2016) 1384–1396, doi:10.1109/JBHI.2015.2459683.

[10] C. Quan, F. Ren, Sentence emotion analysis and recognition based on emotion words using REN-CECPS, Int. J. Adv. Intell. 2 (1) (2010) 105–117.

[11] Y.P. Lin, C.H. Wang, T.P. Jung, T.L. Wu, et al., EEG-based emotion recognition in music listening, IEEE T. Biomed. Eng. 57 (7) (2010) 1798–1806, doi:10.1109/TBME.2010.2048568.

[12] R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from EEG, IEEE T. Affect. Comput. 5 (3) (2014) 327–339, doi:10.1109/TAFFC.2014.2339834.

[13] F. Ren, Y. Dong, W. Wang, Emotion recognition based on physiological signals using brain asymmetry index and echo state network, Neural Comput. Appl. (2018) 1–11, doi:10.1007/s00521-018-3664-1.

[14] T. Bänziger, D. Grandjean, K.R. Scherer, Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT), Emotion 9 (5) (2009) 691–704, doi:10.1037/a0017088.

[15] Y. Chang, C. Hu, R. Feris, M. Turk, Manifold based analysis of facial expression, Image Vision Comput. 24 (6) (2006) 605–614, doi:10.1016/j.imavis.2005.08.006.

[16] M. Pantic, M.S. Bartlett, Machine analysis of facial expressions, in: K. Delac, M. Grgic (Eds.), Face Recognition, IntechOpen, Vienna, Austria, 2007, pp. 377–416.

[17] I. Kotsia, I. Pitas, Facial expression recognition in image sequences using geometric deformation features and support vector machines, IEEE T. Image Process 16 (1) (2006) 172–187, doi:10.1109/TIP.2006.884954.

[18] F. Ren, Z. Huang, Automatic facial expression learning method based on humanoid robot XIN-REN, IEEE T. Hum. Mach. Syst. 46 (6) (2016) 810–821, doi:10.1109/THMS.2016.2599495.

[19] E.I. Barakova, R. Gorbunov, M. Rauterberg, Automatic interpretation of affective facial expressions in the context of interpersonal interaction, IEEE T. Hum. Mach. Syst. 45 (4) (2015) 409–418, doi:10.1109/THMS.2015.2419259.

[20] Y. Chang, M. Vieira, M. Turk, L. Velho, Automatic 3D facial expression analysis in videos, in: W. Zhao, S. Gong, X. Tang (Eds.), International Workshop on Analysis and Modeling of Faces and Gestures, Springer-Verlag Berlin Heidelberg, Heidelberg, Germany, 2005, pp. 293–307.

[21] J. Wang, L. Yin, X. Wei, Y. Sun, 3D facial expression recognition based on primitive surface feature distribution, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 1399–1406, doi:10.1109/CVPR.2006.14.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision, 2015, pp. 4489–4497.

[23] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction. Association for Computing Machinery, New York, New York, 2016, pp. 445–450.

[24] K.R. Scherer, Vocal communication of emotion: a review of research paradigms, Speech Commun. 40 (1–2) (2003) 227–256, doi:10.1016/S0167-6393(02)00084-5.

[25] O.W. Kwon, K. Chan, J. Hao, T.W. Lee, Emotion recognition by speech signals, in: Proceedings of the Eighth European Conference On Speech Communication and Technology, International Speech Communication Association, Baixas, France, 2003, pp. 125–128.

[26] L. Vidrascu, L. Devillers, Detection of real-life emotions in call centers, in: Proceedings of the Ninth European Conference on Speech Communication and Technology, Baixas, France, International Speech Communication Association, 2005, pp. 1841–1844.

[27] Y.L. Tian, T. Kanade, J.F. Cohn, Facial expression analysis, in: S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition, Springer, New York, 2005, pp. 247–275.

[28] J.F. Cohn, K. Schmidt, The timing of facial motion in posed and spontaneous smiles, in: J.P. Li, J. Zhao, J. Liu, N. Zhong, J. Yen (Eds.), Active Media Technology, World Scientific Publishing, Singapore, 2003, pp. 57–69.

[29] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 568–573, doi:10.1109/CVPR.2005.297.

[30] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy method, Neural Netw 18 (4) (2005) 423–435, doi:10.1016/j.neunet.2005.03.004.

[31] DOCOMO developer support, NTT DOCOMO. (n.d.). https://dev.smt.docomo.ne.jp (accessed May 15, 2019).

[32] M. Shuzo, T. Yamamoto, M. Shimura, F. Monma, S. Mitsuyoshi, I. Yamada, Construction of natural voice database for analysis of emotion and feeling, Trans. Inf. Proc. Soc. Jpn. 52 (3) (2011) 1185–1194 http://id.nii.ac.jp/1001/00073614/.

[33] F. Ren, K. Matsumoto, Semi-automatic creation of youth slang corpus and its application to affective computing, IEEE T. Affect. Comput. 7 (2) (2015) 176–189, doi:10.1109/TAFFC.2015.2457915.

[34] Emotion API, Microsoft Azure. (n.d.). https://azure.microsoft.com/ja-jp/services/cognitive-services/emotion/ (accessed May 15, 2019).

[35] G. Cennini, J. Arguel, K. Akşit, A. van Leest, Heart rate monitoring via remote photoplethysmography with motion artifacts reduction, Opt. Express 18 (5) (2010) 4867–4875, doi:10.1364/OE.18.004867.

[36] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 4264–4271.

[37] G. Balakrishnan, F. Durand, J. Guttag, Detecting pulse from head motions in video, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013, pp. 3430–3437.

[38] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, T. Moriyama, Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior, in: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2004, pp. 610–616, doi:10.1109/ICSMC.2004.1398367.

[39] W.B. Cannon, D. de la Paz, Emotional stimulation of adrenal secretion, Am. J. Physiol. 28 (1) (1911) 64–70, doi:10.1152/ajplegacy.1911.28.1.64.

[40] B.M. Appelhans, L.J. Luecken, Heart rate variability as an index of regulated emotional responding, Rev. Gen. Psychol. 10 (3) (2006) 229–240, doi:10.1037/1089-2680.10.3.229.

[41] K. Shibasaki, S. Mitsuyoshi, Evaluation of emotion recognition from intonation: evaluation of sensibility technology and human emotion recognition, IEICE Tech. Rep. 105 (291) (2005) 45–50.

[42] O.W. Kwon, K. Chan, J. Hao, T.W. Lee, Emotion recognition by speech signals, in: Proceedings of the Eighth European Conference on Speech Communication and Technology, Baixas, France, International Speech Communication Association, 2003, pp. 125–128.

[43] M.G. Calvo, P.J. Lang, Gaze patterns when looking at emotional pictures: motivationally biased attention, Motiv. Emot. 28 (3) (2004) 221–243, doi:10.1023/B:MOEM.0000040153.26156.ed.

[44] M. Shuzo, T. Yamamoto, M. Shimura, F. Monma, S. Mitsuyoshi, I. Yamada, Construction of natural voice database for analysis of emotion and feeling, Trans. Inf. Proc. Soc. Jpn. 52 (3) (2011) 1185–1194.

[45] T. Ono, M. Imai, T. Etani, R. Nakatsu, Construction of relationship between humans and robots, Trans. Inf. Proc. Soc. Jpn. 41 (1) (2000) 158–166.

[46] D.M. Medley, H.E. Mitzel, Some behavioral correlates of teacher effectiveness, J. Educ. Psychol. 50 (6) (1959) 239–246, doi:10.1037/h0039475.

[47] M. Asada, H. Ishiguro, Y. Kuniyoshi, Toward cognitive robotics, J. Robot. Soc. Jpn. 17 (1) (1999) 2–6.

[48] F. Ren, C. Quan, K. Matsumoto, Enriching mental engineering, Intl. J. Innov. Comput. Inf. Control 9 (8) (2013) 3271–3284.

Satoshi Nakagawa received the bachelor's degree and master's degree from the University of Tokyo, Tokyo, Japan, in 2018 and 2020 respectively. He is currently a Ph.D. candidate in the Department of Mechano-Informatics, Graduate School of Information Science and Technology, the University of Tokyo. His research interests include human robot interaction, deep learning and elderly welfare.

Daiki Enomoto received B.S. (2012) in agriculture from Hokkaido Univ., Hokkaido, Japan. and M.S. (2019) in Interdisciplinary Information Studies from the University of Tokyo, Tokyo, Japan. Since 2012, he has been a member of LITALICO Inc. He is responsible for the research and development of support systems and services for people with disabilities.

Shogo Yonekura received Ph.D. in 2017 from the Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo. His research interests involve stochastic spiking neural network, robot optimal control, and modeling robot emotion.

Hoshinori Kanazawa is a Project Research Associate at the University of Tokyo, Tokyo Japan. He graduated from Faculty of Human Health Science, Kyoto University Graduate School of Medicine, and received Ph.D. degree in Medicine from Kyoto University, Kyoto, Japan. His main research interests include early human development of sensorimotor systems and cognitive systems.

Yasuo Kuniyoshi received Ph.D. from The University of Tokyo in 1991 and joined Electrotechnical Laboratory, AIST, MITI, Japan. From 1996 to 1997 he was a Visiting Scholar at MIT AI Lab. In 2001 he was appointed as an Associate Professor and then full Professor in 2005 at The University of Tokyo. He is also the Director of RIKEN CBS-Toyota Collaboration Center since 2012, the Director of Next Generation Artificial Intelligence Research Center of The University of Tokyo since 2016, and an affiliate member of International Research Center for Neurointelligence (IRCN) of The University of Tokyo since 2018. He published over 300 refereed academic papers and received IJCAI 93 Outstanding Paper Award, Gold Medal &quot;Tokyo Techno-Forum21&quot; Award, Best Paper Awards from Robotics Society of Japan, IEEE ROBIO T.-J. Tarn Best Paper Award in Robotics, Okawa Publications Prize, and other awards. He is a Fellow of Robotics Society of Japan, President of the Japan Society of Developmental Neuroscience, and a member of IEEE, Science Council of Japan (affiliate), Japan Society of Artificial Intelligence, Information Processing Society of Japan, Japanese Society of Baby Science. For further information about his research, visit http://www.isi.imi.i.u-tokyo.ac.jp/ and http://www.ai.u-tokyo.ac.jp/