Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 3, June 2020, pp. 1220~1228
ISSN: 2302-9285, DOI: 10.11591/eei.v9i3.2035

The study of attention estimation for child-robot interaction scenarios

Muhammad Attamimi[1], Takashi Omori[2]
[1]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia
[2]Graduate School of Engineering, Tamagawa University, Japan

Article Info

Article history:

Received Aug 6, 2019
Revised Nov 3, 2019
Accepted Dec 4, 2020

Keywords:

Attention estimation
Child-robot interaction
Features extraction
Multimodal information

ABSTRACT

One of the biggest challenges in human-agent interaction (HAI) is the development of an agent such as a robot that can understand its partner (a human) and interact naturally. To realize this, a system (agent) should be able to observe a human well and estimate his/her mental state. Towards this goal, in this paper, we present a method of estimating a child's attention, one of the more important human mental states, in a free-play scenario of child-robot interaction (CRI). To realize attention estimation in such CRI scenario, first, we developed a system that could sense a child's verbal and non-verbal multimodal signals such as gaze, facial expression, proximity, and so on. Then, the observed information was used to train a model that is based on a Support Vector Machine (SVM) to estimate a human's attention level. We investigated the accuracy of the proposed method by comparing with a human judge's estimation, and obtained some promising results which we discuss here.

Corresponding Author:

Muhammad Attamimi,
Department of Electrical Engineering,
Institut Teknologi Sepuluh Nopember,
Raya ITS, Sukolilo 60111, Surabaya, Indonesia.
Email: attamimi@ee.its.ac.id

1. INTRODUCTION

Interaction between humans and agents such as computers, robots, etc., has long been studied in the field of human-agent interaction (HAI). In human-computer interaction (HCI), a great deal of effort has been made in developing computer interfaces [1-3]. The purpose of those studies is to make better interfaces that can interact with humans in a natural (that is, more human-like) manner. As methods in robotics have advanced, robots have become more commonplace in humans' daily lives. This progress has led to increased interest in the field of human-robot interaction (HRI) towards the development of robots that can interact naturally with humans. By "natural interaction," we usually refer to interfacing by oral and gesture commands and reports. However, an essential part of natural interaction is deeply related to aspects of human behavior driven by human intention and other mental states.

In considering human-to-human interaction, humans respond according to their partner's actions during interaction. Verbal and non-verbal signals can be observed. Therefore, it can be said that information is exchanged during the interaction. Several investigations have addressed this issue, including studies of verbal and/or nonverbal communication as well as their application to robots [4-6]. However, the communication signals implemented in these studies were rather artificial and therefore did not provide sufficient data to understand more natural human-robot interactions. To realize human-like, mental-state based, interaction, we need to estimate the mental states of human based on the observable

information. Such information is normally acquired by our sensory systems such as vision, audition and so on. We strongly believe that the same information can be applied to a natural HRI scenario. Thus, in order to create such scenarios, we needed to solve two problems. First, the systems (including the robot itself), needed to be able to observe a set of features of human behavior, i.e., utterance, gaze, distance, and so forth, that might represent the human mental state. Second, we needed a classifier able to estimate human mental states such as arousal, attention, mood, etc. from the observed signals.

Additionally, it is possible to roughly segregate target HRI groups into children, adults, and elderly according to the subject of interaction. Although in general, similar considerations can be made for all three groups, we focused on child-robot interaction (CRI) in this study, because unlike adults, children tend not to feign their feelings and observations would be easier. We consider that this approach could be advantageous for evaluation in this field. In CRI, there are several related studies [7-10]. Methods of controlling robots able to act as friends or playmates have been discussed in [7-8]. And although the results were encouraging, the scenarios were overly controlled. In [9-10], studies were conducted on the interaction between robots and children in more natural settings, or "in the wild," as it is said. Although these efforts were ground-breaking and related to our research, their action decision analysis was based on direct physical observation and not on the mental state, thus limiting its application to the specific tasks of the experimental design.

Therefore, in our study, we focused on attention estimation of children playing freely with a robot as a simple but typical example of mental state-based interaction. To this end, we first developed a system including sensor networks and a remotely operated robot to enable free-interaction with the children. Over 23 3-6 year-old children participated in our interaction experiment and a total of 60 minutes interaction was recorded. In a pilot study, we investigated one of the subjects which interacted for five to six minutes with the robot. Figure 1 shows some scenes from the free-play interaction. We processed the data and extracted the features of proximity, eye gaze, emotion in facial expression, and behaviors, as well as a set of attention values labelled by human judges. We then utilized a Support Vector Machine (SVM) to estimate the attention level. We determined the accuracy of the proposed method by comparing with the expert judgments, and some promising results were obtained which we report here.

Figure 1. Child-robot interaction scenes: free-play

Several other studies relate to ours [11-22]. An interesting discussion related to gaze estimation has been elaborated in [11], and an alternative solution to gaze estimation was proposed in [12]. A model for adaptive emotion expression was developed in [13], and the work in [14] closely relates to our study. However, although the gaze is an effective signal for the attention estimation, gaze alone is insufficient for the estimation of emotion or intention. What distinguishes our study from these others is the use of information acquired through free-play in the CRI scene. In [15-16], vision-based attention information was proposed. However, those studies employed tasks different to those we applied here. The works that discussed the visual recognition system implemented on the robot [17-20] and computer vision [21-31] were important to obtain a better sensing in this study. Moreover, the scenarios proposed in [32-33] were also interesting to be implemented in our system. The remainder of this paper is organized as follows: An overview of the proposed child-robot interaction framework followed by the details of its components is described in the next section. Experimental results are discussed in section 2 followed by a discussion in section 4. Finally, section 5 concludes this study.

2. PROPOSED METHOD

The proposed child-robot interaction (CRI) framework is depicted in Figure 2. As illustrated in Figure 2(a), the proposed CRI framework incorporates global and local sensing. The proposed framework consists of: (1) sensor networks, (2) a robot platform, (3) robot remote operation, (4) database, and (5) modeling. The details of each part are discussed as follows.

2.1. Sensor networks

To provide global sensing, we developed a sensor network, shown in Figure 2(b). Sensor networks consisted of four Kinect sensors and four stereo microphones used to capture audio information during the interaction. From the Kinect sensor, color and depth information was collected at 30 fps, which was used to detect human information including position, face, etc. as well as acquiring information about the robot. These cues were used for feature extraction (see section 2.5).

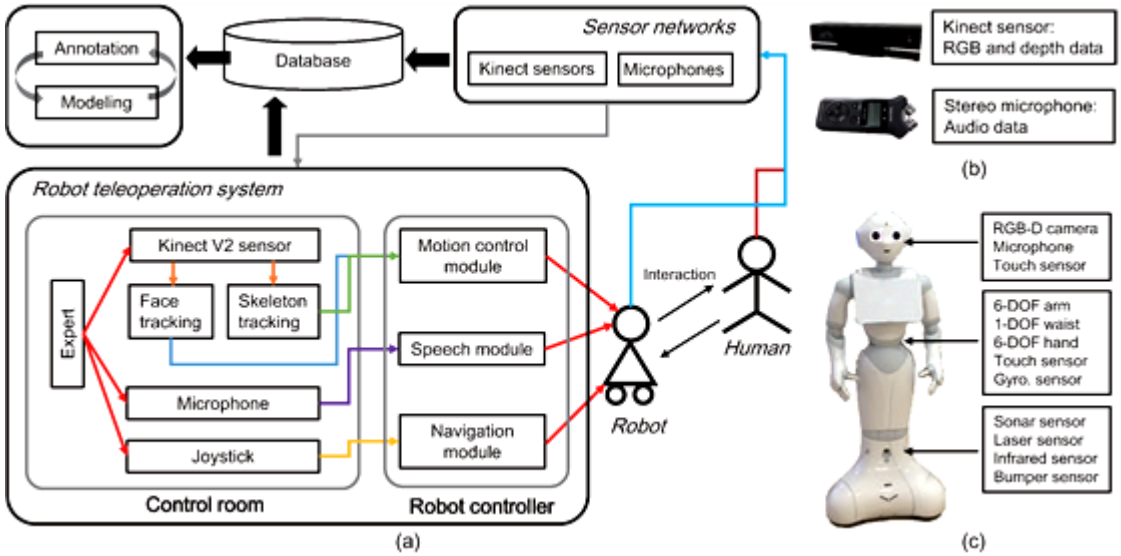

Figure 2. Overview of proposed child-robot interaction framework, (a) Child-robot interaction framework, (b) Sensors used in this study, and (c) "Softbank Robotics' Pepper": robot platform

2.2. Robot platform

In this study, the commercially-available "Pepper" robot, shown in Figure 2 (c) developed by Softbank Robotics was used. This robot can be roughly divided into three parts, i.e., head, upper body, and lower body. The head consists of one depth camera and two RGB cameras that are calibrated, four microphones, and three touch sensors. Using the RGB and depth cameras, this robot is able to detect and track humans faces by exploiting the API provided by the robot platform. The touch sensors built into the robot's head can be used for touch-interaction. The upper body of the robot consists of two arms with six degrees of freedom (DOF) each, a two-DOF neck, and a one-DOF waist. This body structure enables the robot to move freely as well as perform human-like motions such as bowing. The lower body is based on omnidirectional wheels, two sonar sensors, six laser sensors, three bumper sensors, and two infrared sensors. Using these sensors, the robot is not only able to navigate freely but also can detect obstacles and avoid collisions. This safety function is rather important in CRI scenarios. A computer built into the robot is used for computation and communication between the modules. The robot is also capable of wireless communication, which can be used to transmit acquired data for multipurpose application.

2.3. Robot teleoperation system

In the first step of our study, we implemented a teleoperation system in the robot to enable human-like responses. This ability is important in dealing with children, especially during the first meeting, to encourage longer interaction. Moreover, it is quite difficult to apply autonomous action decisions to the robot for a natural interaction on the first try. To tackle such a problem, we collected data in this first step using a teleoperation system and created a model based on the captured data to enable more natural child-robot interaction.

The teleoperation system consisted of: (1) a motion control module, (2) a speech module, and (3) a navigation module. In the control room, a Kinect sensor, one microphone, and a joystick were used to control the robot. An expert kindergarten teacher who was also a children's education specialist teleoperated the robot from the control room. Thanks to the sensor networks, the CRI could be observed in real time. First, we used a face-tracking API and skeleton-tracking API from the Microsoft Kinect API to capture the head movements and body movements of the operator, respectively. We then integrated these signals in order to guide the robot's movements. Next, for the speech module, we used Google's speech

recognition. The utterances of the operator were recognized, and the results were used to generate the robot's voice. It should be noted that this process is necessary since the robot's voice should differ from the operator's. Finally, we applied a simple module to control the robot's base using a joystick.

2.4. Multimodal data collection

We collected a multimodal dataset which can be roughly divided into (1) the data from the robot via local sensing as well as its internal state, (2) the environment including the human and robot, and (3) the operation records. These data mainly consisted of color and depth information and audio information. We also annotated the data according to the expert's judgments.

2.5. Attention estimation

The captured multimodal data was processed to extract features for input of the attention estimation model. To model attention, we considered gaze, utterance, behavior, proximity, and emotion of the child interacting with the robot as well as the input data. In a preliminary effort, we considered a simple Support Vector Machine (SVM) to categorize the child's attention based on the observed data. An SVM [34] learns by mapping the training data to a higher dimensional feature space using a kernel function and constructing a separating hyperplane to achieve maximum margin between classes. In this study, a publicly available library LIBSVM [35] was used because it provides a probabilistic estimation which is fast and easy to implement. Each of the features described below is normalized to one for a better depiction of multimodal data input.

2.5.1. Manually extracted features

One of the purposes of this study was to realize natural CRI as human beings do. It is straightforward to study and compare the features that come from the human to ones that are calculated by machines. Thus, we extracted the gaze, utterance, and other behaviors by asking the judges what their considerations were when judging the scenes. The gaze was set to zero when the child was considered not to be looking at the robot, and otherwise, it was set to one. For utterances, when the child was talking to the robot, it was set to one, and otherwise, to zero. We also asked the judges to rate the behavior of the child towards the robot on a 0-3 scale based on interest. It should be noted that features were extracted on a second-by-second basis.

2.5.2. Automatically extracted features

Given the set of calibrated color and depth information as shown in Figure 3(a), Figure 3(b), Figure 3(c), and Figure 3(d). We performed a 3D segmentation using the publicly available library PCL [36] to separate human and robot signals.

Figure 3. Examples of scene segmentation and face processing, (a) Input color image (1920 x 1080) with detected face illustrated in red rectangle, eyes, and mouth, (b) Depth image (512 x 424), (c) Segmented scene (512 x 424), and (d) Extracted robot and child (512 x 424). Each image is calibrated and the rectangles as well as their center positions in (a) and (c) correspond

The idea of the 3D segmentation is (1), to detect the plane and isolate point clouds which belong to the plane, and (2), to cluster the remaining points. These steps are depicted in Figure 3(c) given Figure 3(b) as an input depth image. After extracting the plane, we use connected-component labeling [37] to reduce the noise caused by the Kinect sensors. We then calculated the position of detected objects by considering the center of gravity of the cluster, and the Euclidean distance between the robot and the human could then easily be calculated.

For face detection, we applied image processing provided by [38-40]. Thanks to [38] and [40], the parts of the face including the eyes and mouth could be detected as well as estimates of facial expression including neutrality, happiness, surprise, anger, and sadness could be taken. Given the position of the eyes and mouth, we were able to draw a triangle as shown in Figure 3(a) in color space. Corresponding points were then calculated in the depth space. We then calculated the normal vector of the triangle plane. The robot's head position could also easily be obtained. Here, gaze was defined as the inner product of the robot's normal vector and child's normal vector. The greater this value, the greater the tendency the robot and the child were looking at each other.

3. EXPERIMENTS

We implemented the CRI framework and conducted experiments. The objectives of these experiments were as follows. First, we wanted to investigate the features used by the experts and compare these with the automatically extracted ones. Second, we wanted to test the attention estimation model given the proposed features. The experiments were conducted at a kindergarten and 23 children (13 boys and 10 girls, aged three to six years) participated.

3.1. Experimental setup

A control room and free-play room were used as experimental environments. Figure 4(a) and Figure 4(b) show some examples of scenes taken during the experiments. Before the children entered the play room (see Figure 4 (b)), they were encouraged to play in a playground as shown in Figure 4(a). Considering that the problem-setting in this study was free-play, no explicit instructions were given to the children. They were just told that there was a robot in the free-play room. After entering the play room, the children interacted freely with the robot as illustrated in Figure 4(b).

An expert, who is a kindergarten teacher and analyzes children's education operated the robot as shown in Figure 4(c). Our teleoperation system greatly facilitated the manipulation of the robot. Some examples of skeleton data, facial data, and the robot's movements are illustrated in Figure 4(c). From the figure, we could see that the robot could follow the operator. At the same time, the operator observed the play room and responded to the children's questions or asked questions. It should be noted that children were not told that the robot was remotely operated. The quantitative evaluation of the overall system will remain work in progress.

Figure 4. Examples of scenes taken during the experiments: (a) Play ground, (b) Free-play room, and (c) Control room. In the control room, the expert (on the left, in (c)) operated the robot. The upper panel in (c) depicts the skeleton data and facial data of the operator, whereas the bottom panel illustrates the corresponding movements of the robot

All the subjects participated in an interaction experiment and a total of 60 minutes interaction was recorded. Several interesting scenes occurred, such as when children asked the robot to engage in "make-believe play," made physical contact by touching the robot's head, and so forth. It should be noted that such scenes occured purely and spontaneously, without control by the robot. In preliminary work,

we investigated one of the subjects which had interacted for five to six minutes with the robot (hereafter, referred as a "target subject"). The data collected over all subjects, though, is extensive enough that it will take time to fully analyze. However, we found our results promising enough to constitute a good first approach in the analysis of the free-play situation in CRI and should be valuable in additional studies of human-agent interaction. To evaluate the proposed method, we asked six experts to view the video of the target subject and make judgments every second, resulting in 303 seconds of annotated videos. It should be noted that the labels most selected by the experts were used. We also asked our judges what they took into consideration to determine those labels. Based on these responses, the experts were later asked to score the gaze, utterance, and behavior for each second of a video.

3.2. Feature extraction and comparison

The collected multimodal data, consisting of visual (color and depth) information and audio information was processed. Figure 5 shows the features that were extracted manually (left) and automatically (right) (see section 2.5.1 and 2.5.2), respectively. From the figure we can see the feature variations over time. Examples of interaction scenes are shown in Figure 6. It can be seen that at first, the child was eager to interact with the robot. This fact was also supported by high gaze and utterance values. The subject wanted to talk with the robot so she stomped her foot to draw the robot's attention. When the robot responded to the subject, she approached the robot and began a dialogue. The system was also able to estimate the facial expression of "happiness" at that time. After a while, the subject became bored and showed signs of wanting to retreat. We can see that the distance increased and the gaze value diminished.

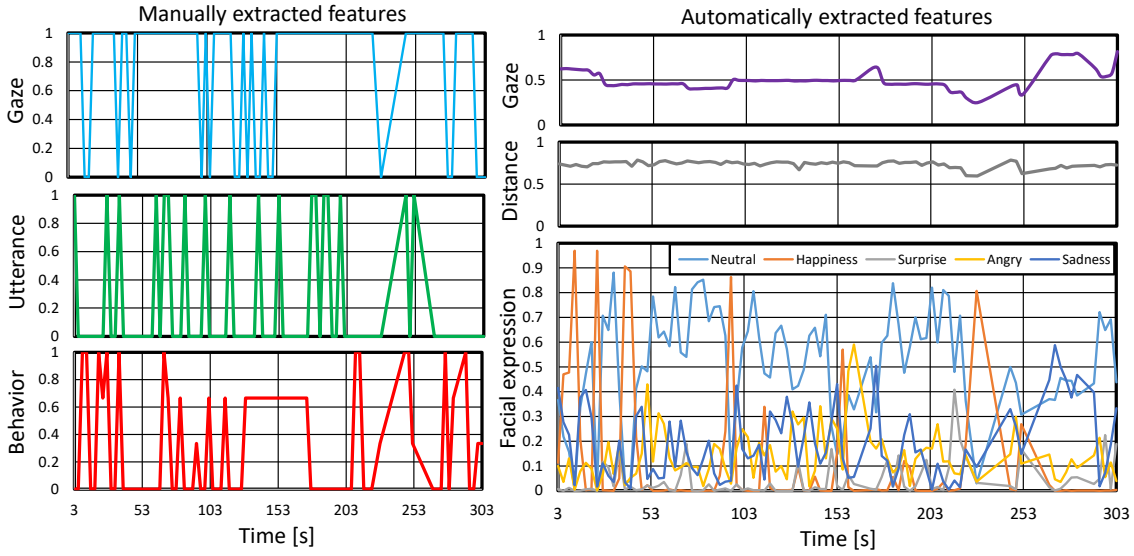

Figure 5. Features used in this study

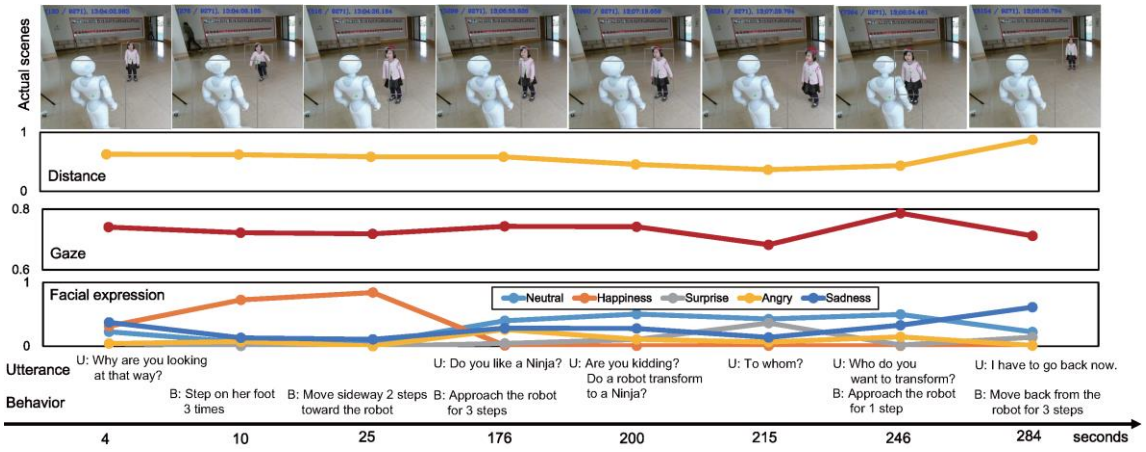

Figure 6. Examples of interaction scenes with corresponded features

3.3. Evaluation of attention estimation

To validate the proposed method, we performed leave-one-out validation (LOOV) on the collected data. On occasions such as when the subject approached and hid behind the robot, it was impossible to process the data due to occlusion and thus could not be segmented automatically. Excluding these data, LOOV was done on 258 data sets. Here, one data set consisted of manually extracted features (gaze: one-dimensional feature vector, utterance: one-dimensional feature vector, behavior: one-dimensional feature vector) and automatically extracted features (gaze: one-dimensional feature vector, proximity: one-dimensional feature vector, facial expressions: five-dimensional feature vector). Here, we tested our proposed method using two- and three-classes of SVM.

In this study, two-classes indicate that the subject was “uninterested” or “interested” in the robot; whereas three-classes indicate that the subject was “uninterested” or “less interested” or “interested” in the robot. We have tested several combinations of features as listed in Table 1 and Table 2. We can see that both manually- and automatically-extracted features reached their highest rates when all the features were included as input except when method number 5 in Table 2 was used due its poor extraction of gaze. In addition, the best results of manual and automatic features for each class were written in bold. The results proved interesting because in three-classes of attention estimation, the features calculated automatically by machine outperformed those of the experts. We address this interesting phenomenon in the next section.

Table 1. Manually extracted atures used for attefention estimation and their corresponding classification rates

Features used	Classification results (two-classes) [%]	Classification results (three-classes) [%]
Gaze	80.34	54.65
Utterance	66.67	47.29
Behavior	76.92	52.71
Gaze + Utterance	80.34	58.53
Gaze + Behavior	79.49	58.14
Utterance + Behavior	77.78	57.36
Gaze + Utterance + Behavior	82.05	58.91

Table 2. Automatically extracted features used for attention estimation and their corresponding classification rates

Features used	Classification results (two-classes) [%]	Classification results (three-classes) [%]
Gaze	75.21	54.26
Utterance	64.96	53.86
Behavior	76.92	62.40
Gaze + Utterance	70.94	57.75
Gaze + Behavior	76.92	63.57
Utterance + Behavior	77.78	62.40
Gaze + Utterance + Behavior	78.63	63.57

3.4. Discussion

We can see from Table 1 and Table 2 that the attention estimation results of two-classes for manually extracted features were higher than the automatic ones. However, the difference in accuracy was less than 4%. It can be said that the proposed automatically-extracted features were similarly predictive as compared to the manually extracted ones. Hence, we argue that automatically-extracted features can approximately represent the attention level judgments of humans. Next, the results of three-class feature extraction both of manually and automatically extracted features were poorer as compared to the two-class ones.

However, a notable point of these results was the slightly better results obtained by automatic feature extraction as opposed to manual. We emphasize that a combination of several features proved better for attention estimation than a single one. Although we used a simple machine learning algorithm in this study, the same thing can be said for the complex one. In the future, we plan to further exploit multimodality. Attention estimation involving multiple classes is not an easy task even for humans. This is one of the reasons that machine learning could outperform human estimation results based on human-generated cues.

4. CONCLUSION

We have proposed a method of estimating children's attention, an important human mental state. In this study, a free-play scenario of child-robot interaction was considered. To estimate attention level, we first proposed a framework for child-robot interaction based on local and global sensing using robots and sensor networks. We developed the system ourselves and implemented a teleoperation system for the robot. We have also investigated several features that were manually and automatically extracted and compared their effectiveness. We found that the combination of all features worked better than those based on subsets. Although the proposed automatically-extracted features performed well for higher level classification, additional work is required to understand which features and conditions are most predictive of successful human-robot interaction. Our goal, therefore is to develop better models that can more fully exploit the multiple modalities of the captured data.

REFERENCES

[1] G. Salvendy and M. J. Smith, "Human-computer interaction: Software and hardware interfaces," *5th International Conference on Human-Computer Interaction, Elsevier Science*, pp. 140-145, 1993.

[2] T. Clemmensen, V. Kaptelinin, and B. Nardi, "Making HCI theory work: An analysis of the use of activity theory in HCI research," *Behaviour & Information Technology*, vol. 35, no. 8, pp. 608-627, 2016.

[3] Y. Rogers, "HCI theory: Classical, modern, and contemporary," *Synthesis Lectures on Human-Centered Informatics*, vol. 5, no. 2, pp. 1-129, 2012.

[4] T. Belpaeme et al., "Multimodal child-robot interaction: Building social bonds," *Journal of Human-Robot Interacion*, vol. 1, no. 2, pp. 33-53, 2013.

[5] O. W. Kwon et al., "Emotion recognition by speech signals," *European Conference on Speech Communication and Technology*, pp. 125-128, 2003.

[6] V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Proceedings of artificial neural networks in engineering*, vol. 710, 1999.

[7] S. Shahid, E. Krahmer, and M. Swerts, "Child-robot interaction: playing alone or together?," *Extended Abstracts on Human Factors in Computing Systems*, pp. 1399-1404, 2011.

[8] M. Attamimi, K. Abe, A. Iwasaki, T. Nagai, T. Shimotomi, and T. Omori, "Robots that can play with children: What makes a robot be a friend," *International Conference on Neural Information Processing*, pp. 377-386, 2013.

[9] J. de Greeff, O. B. Henkemans, A. Fraajie, l. Solms, N. Wignor, and B. Bierman, "Child-robot interaction in the wild: Field testing activities of the ALIZ-E project," *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 148-149, 2014.

[10] R. Ros, M. Nalin, R. K. Wood, and P. Baxter, "Child-robot interaction in the wild: Advice to the aspiring experimenter," *13th International Conference on Multimodal Interfaces*, pp. 335-342, 2011.

[11] P. Baxter, J. Kennedy, A-L. Vollmer, J. de Greeff, and T. Belpaeme, "Tracking gaze over time in HRI as a proxy for engagement and attribution of social agency," *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 126-127, 2014.

[12] J. Kennedy, P. Baxter, and T. Belpaeme, "Head pose estimation is an inadequate replacement for eye gaze in child-robot interaction," *Int. Conference on Human-Robot Interaction Extended Abstracts*, pp. 35-36, 2015.

[13] M. Tielman, M. Neerincx, J-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 407-414, 2014.

[14] S. Yildirim and S. Narayanan, "Recognizing child's emotional state in problem-solving child-machine interactions," *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pp. 1-4, 2009.

[15] L. Li et al., "Vision-based attention estimation and selection for social robot to perform natural interaction in the open world," *7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 183-184, 2012.

[16] D. van der Pol, R. H. Cuijpers, and J. F. Juola, "Head pose estimation for real-time low-resolution video," *28th Annual European Conference on Cognitive Ergonomics*, pp. 353-354, 2010.

[17] M. Attamimi, T. Araki, T. Nakamura, and T. Nagai, "Visual recognition system for cleaning tasks by humanoid robots," *International Journal of Advanced Robotic Systems*, vol. 10, no. 11, pp. 1-14, 2013.

[18] M. Attamimi, T. Nagai, and D. Purwanto, "Particle filter with integrated multiple features for object detection and tracking," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 16, no.6, pp. 3008-3015, 2018.

[19] M. Attamimi et al., "Real-time 3D visual sensor for robust object recognition," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4560-4565, 2010.

[20] M. Attamimi, T. Nakamura, and T. Nagai, "Hierarchical multilevel object recognition using Markov model," *21st International Conference on Pattern Recognition*, pp. 2963-2966, 2012.

[21] T-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.

[22] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," *IEEE Xplore*, pp. 9308-9316, 2018.

[23] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4203-4212, 2018.

[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *ArXiv: 1804.02767Comment: Tech Report*, pp. 1-6, 2018.

[25] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517-6525, 2017.
[26] J. Dai et al., "Deformable convolutional networks," *IEEE Xplore*, pp. 764-773, 2017.
[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.
[28] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," *Conference on Computer Vision and Pattern Recognition*, pp. 2846-2854, 2016.
[29] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "SoftNMS improving object detection with one line of code," *International Conference on Computer Vision*, pp. 5562-5570, 2017.
[30] Z. Cai and N. Vasconcelos, "Cascade RCNN: Delving into high quality object detection," *Conference on Computer Vision and Pattern Recognition*, pp. 6155-6162, 2018.
[31] K. Chen et al., "Hybrid task cascade for instance segmentation," *Conference on Computer Vision and Pattern Recognition*, pp. 4974-4983, 2019.
[32] M. Attamimi, Y. Katakami, K. Abe, T. Nakamura, and T. Nagai, "Modeling of honest signals for human robot interaction," *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 415-416, 2016.
[33] K. Abe, C. Hieida, M. Attmimi, and T. Nagai, "Toward playmate robots that can play with children considering personality," *2nd International Conference on Human-Agent Interaction*, pp. 165-168, 2014.
[34] B. Scholkopf, R. C. Williamson, and P. L. Barlett, "New support vector algorithms," *Journal Neural Computation*, vol. 12, no. 5, pp. 1207-1245, 2000.
[35] C.C. Chang and C-J. Lin, "LIBSVM: a library for support vector machines," *Journal ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.
[36] R. B. Rusu and C. Steve, "3D is here: point cloud library (PCL)," *IEEE International Conference on Robotics and Automation*, pp. 1-4, 2011.
[37] L. He, Y. Chao, and K. Suzuki, "A run-based two-scan labeling algorithm," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 749-756, 2008.
[38] T. Yamashita, "Human sensing technology OKAO vision," *IHS Interaction Technology Summit*, 2013.
[39] K. Kinoshita, Y. Konishi, S. Lao, and M. Kawade, "Facial feature extraction and head pose estimation using fast 3D model fitting," MIRU 2008, pp. 1325-1329, 2008.
[40] Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade, "Real-time estimation of smile intensities," *Interaction*, vol. 4, pp.47-48, 2008.

BIOGRAPHIES OF AUTHORS

Muhammad Attamimi received his BE, ME, and DE degrees from the University of Electro-Communications in 2010, 2012, and 2015, respectively. He received scholarship from Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT) for his BE and ME courses. From 2012 to 2015, he was also a Research Fellow (DC1) of Japan Society for the Promotion of Science (JSPS). He was with a postdoctoral researcher at the department of Mechanical Engineering and Intelligent Systems, The University of Electro-communications (UEC) from April 2015 to December 2015. Since January 2016, he was with Tamagawa University Brain Science Institute as a commission researcher for six months. Currently, he is a lecturer at Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research interests are computer vision, visual recognition, machine learning, multimodal categorization, artificial intelligent, probability robotics, intelligent systems, intelligent robotics.

Takashi Omori completed the graduate program at the Graduate School of Engineering, University of Tokyo, in 1980 and became a research associate in 1981. He served as a lecturer from 1987, as an associate professor from 1988, and as a professor from 1998 in the Department of Electronic and Electrical Engineering, Tokyo University of Agriculture and Technology. He became a professor since May 2000 in the Graduate School of Engineering, Hokkaido University, and is now a professor of Faculty of Engineering, Tamagawa University and Brain Research Institute of Tamagawa University from 2006. He started his research field from a self-organizing model of NN, and it moved to higher cognition process from vision, memory and development. His current interest is in a computational modeling of mental process between two interacting agent. He served as a president of Japanese Cognitive Science Society from 2015 to 2016, and Japanese Neural Notwork Society from 2017 to 2018.