# Supporting Human–Robot Interaction Based on the Level of Visual Focus of Attention

Dipankar Das, Md. Golam Rashed, Yoshinori Kobayashi, and Yoshinori Kuno, *Member, IEEE*

***Abstract*—We propose a human–robot interaction approach for social robots that attracts and controls the attention of a target person depending on her/his current visual focus of attention. The system detects the person's current task (attention) and estimates the level by using the "task-related contextual cues" and "gaze pattern." The attention level is used to determine the suitable time to attract the target person's attention toward the robot. The robot detects the interest or willingness of the target person to interact with it. Then, depending on the level of interest of the target person, the robot generates awareness and establishes a communication channel with her/him. To evaluate the performance, we conducted an experiment using our static robot to attract the target human's attention when she/he is involved in four different tasks: reading, writing, browsing, and viewing paintings. The proposed robot determines the level of attention of the current task and considers the situation of the target person. Questionnaire measures confirmed that the proposed robot outperforms a simple attention control robot in attracting participants' attention in an acceptable way. It also causes less disturbance and establishes effective eye contact. We implemented the system into a commercial robotic platform (Robovie-R3) to initiate interaction between visitors and the robot in a museum scenario. The robot determined the visitors' gaze points and established a successful interaction with a success rate of 91.7%.**

***Index Terms*—Gaze pattern, human–robot interaction, task-related contextual cues, visual focus of attention (VFOA).**

## I. Introduction

FOR robots to interact effectively with humans in service applications or in collaborative work scenarios, they should be perceived as social actors and exhibit social intelligence and awareness [1]. This social role awareness involves the ability to behave in a socially correct manner, the ability to communicate with proper timing, according to the situation, and the feelings of the interactive partners as humans do with other humans. We propose an intelligent robotic method of attracting a target person's attention and establishing a communication channel with her/him based on her/his level of visual focus of attention (LVFOA). The visual focus of attention (VFOA) is the behavioral and cognitive process that indicates where and at what a person is looking, and that can be determined by eye gaze and head pose dynamics [2]. LVFOA refers to how much concentration is given to a particular VFOA and is classified into discrete levels: low, high, or medium [3]. If the robot needs to start communication urgently such as during an emergency, it does not need to consider the current situation of the person. Otherwise, the robot should observe the person to know at what/who she/he is looking (VFOA) and how attentively she/he is doing so (LVFOA). Then, it should find a proper timing to attract her/his attention so that it does not interfere with her/his current work. We propose a system in which the robot interacts with the target person intelligently and in a socially acceptable manner so that it can interact by considering her/his current VFOA as well as other persons in the environment. Researchers in human–robot interaction have been interested in developing models inspired by human cognitive processes because such models result in natural interaction behaviors [4], [5]. Providing robots with skills that make the interaction intelligent and intuitive supports a high level of satisfaction for interacting humans.

The VFOA is an important cue for attracting attention and initiating interaction because: 1) it helps with understanding what the person is doing, and 2) it indicates addressee-hood (who is looking at whom). For instance, if the target person's VFOA is toward the robot, the robot can immediately establish a communication channel through eye contact. If the target person is involved in some task, the robot should wait to find a proper timing to attract her/his attention and establish a communication channel. In this research, the proper timing is determined by detecting the level of attention of the target person on her/his current task. In a scenario such as reading, writing, or browsing, the robot should initiate interaction with the target person when her/his level of attention is low. In other settings, such as at a museum, the robot may need to consider people's high level of attention to help with objects of interest.

We use visual cues such as gaze pattern, and the task context of the target person to recognize the VFOA and its level. Visual cues such as gaze pattern and head pose can be used as an approximation for VFOA [2], [6]. The task context plays an important role to relate the set of circumstances in which the task takes place [7]. Context is also relevant to derive a precise understanding of task behavior. Knowledge of context can be used to make decisions such as when to interrupt the target person. For instance, if the target person is involved in "reading," contextual cues such as "turn page" or "change in the tilt angle

Manuscript received March 1, 2014; revised August 22, 2014, November 14, 2014, and February 26, 2015; accepted May 16, 2015. Date of publication July 7, 2015; date of current version November 12, 2015. This work was supported by the Japan Science and Technology Agency, Core Research for Evolutionary Science and Technology. This paper was recommended by Associate Editor C. M. Lewis.

D. Das is with the Department of Information and Communication Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh (e-mail: dipankar@cv.ics.saitama-u.ac.jp).

M. G. Rashed and Y. Kuno are with the Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan (e-mail: golamrashed@cv.ics.saitama-u.ac.jp; kuno@cv.ics.saitama-u.ac.jp).

Y. Kobayashi is with the Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan, and also with the Japan Science and Technology Agency, Saitama 332-0012, Japan (e-mail: yosinori@cv.ics.saitama-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2015.2445856

of head" can be used to determine the loss of current VFOA. In [8], we used head pose information to determine the level of attention of the target human in a controlled experimental environment. Here, we use head pose and eye gaze information to determine the level of attention. We evaluate the proposed system with a commercial robot in a museum scenario.

## II. Related Work

### A. Visual Focus of Attention

VFOA is often highly correlated with behavior or activity and is estimated by eye gaze. However, in some cases, the scale of the scene does not permit estimation of eye gaze directly because it could require either the movement of the subject to be constrained or high-resolution images of the eyes, which may not be practical [9], [10]. In [11], Stiefelhagen *et al.* use a hidden Markov model to detect a user's focus of attention from an observed sequence of gaze estimates. They only consider head pose as the indicator of the gaze and showed that VFOA can be derived by head pose in many cases. For estimating the degree of VFOA, Asteriadis *et al.* [12] use information from head rotation and eye gaze estimation. In estimating head pose, the authors use Bayesian modality fusion of both local and holistic information, while for eye gaze, they use a methodology that calculates eye gaze directionality, removing the influence of head rotation. In [13], the authors use low-cost camera images for estimating VFOA using head rotation, as well as fuzzy fusion of head rotation and eye gaze estimates, in a fully automatic manner, without the need for any special hardware or a priori knowledge regarding the user, the environment or the setup. However, in [13], the user needs to maintain a frontal pose to the camera at start-up and there is no appropriate mappings between 2-D projections and head/eye gaze analysis to certain points on a target plane. We use a low-cost camera for estimating VFOA that combines both eye and head movement effects in a human–robot interaction scenario.

For recognizing a human's level of VFOA, researchers use techniques based on active sensing using infrared light [14]. Although accurate, these methods are invasive and restrictive. Researchers try to estimate the level and orientation of attention by relying on exterior attributes [15]. Movement analysis, head pose estimation, and eye gaze measurement are used to investigate the VFOA [16], [17]. These approaches require technology applicable to laboratory environments (e.g., wearing devices like binoculars or using multimirrored setups and projections). Ba and Odobez [7] proposed a multiperson VFOA approach using head pose and contextual information. They propose to recognize the participants' visual attention in order to introduce context-dependent interaction models that relate group activity and social dynamics of communication. However, we use the task and task-related behavior pattern to measure the level of attention.

### B. Initiating Interaction

We consider when a human and a robot may not face each other and the robot initiates an interaction in a socially acceptable way depending on her/his level of current VFOA. Human gaze supported better human–agent interfaces in [18]. Adaptive gaze patterns have been used for interaction with human and artificial agents [19]. Moment-by-moment eye gaze plays an important role in human–agent interaction and collaboration. Mutlu *et al.* [20] investigated the role of eye gaze in a story telling robot and showed that the participants can better recall the story when the robot looked at them more often while telling the story. The robot's gaze directly influences where people look in the scenes, and this affects people's comprehension of the robot's utterance [21]. The eye-tracking data produce different patterns of human eye gaze depending on the robot's gaze and speech. We use gaze pattern to estimate VFOA to find a suitable time of interaction. Researchers have used gaze behaviors as a tool to study human cognitive processes including reading, viewing pictures or videos, and driving [22]–[24]. Although most state-of-the-art gaze trackers are accurate and reliable, they require complex hardware (helmet with a mounted camera) and/or the user has to be in a fixed position (e.g., with a chin chest). Few studies use gaze patterns for initiating interaction in HRI. Johansson *et al.* [25] used head pose patterns in multiparty human–robot team-building interaction. They presented a data collection apparatus for exploring turn-taking in three-party human–robot interaction involving objects competing for attention. Weiddenbacher *et al.* [26] showed that the combined information of head pose and eye gaze provides more effective information.

### C. Establishing Communication Channel

Attention attraction (AA) can produce observable behavioral responses such as eye movements, head movements, or body orientation. If the target person is attracted by the robot behaviors, the target person will turn toward the robot, supporting eye contact. Several robotic systems establish eye contact by gaze crossing [27], [28]. Studies show, however, that the gaze crossing action alone may not be enough to establish eye contact. Gaze awareness is also necessary for humans to feel that they have made eye contact [29]. Therefore, robots need not only to detect human gaze but also to accurately display their gaze awareness for human interpretation. Even if a robot has noticed that a human is looking at it, eye contact may not be established if the human is not aware of this fact. The computational agent should be able to display its awareness explicitly through some actions (e.g., facial expression, eye blinking, and waving) [30]. Eye blinking by an on-screen agent gives participants a stronger feeling of being looked at [31]. Herein, we show the effectiveness of awareness generation by eye blinking action in making eye contact.

## III. Human Visual Focus of Attention Analysis

Our main objective is to estimate VFOA and its level for a given target person when she/he is involved in a task. While the VFOA is defined by 3-D eye gaze direction, people tend to look at target objects which are of immediate interest [32]. We define the VFOA of a target person involved in a task, $T_i$ ($T = \{reading, writing, browsing, viewing\ painting\}$), where $i = 1 \ldots 4$, is an element belonging to a finite set of viewable

TABLE I
SPAN OF VFOA IN MINUTES

| | Reading | Writing | Browsing | Viewing |
|---|---|---|---|---|
| Average | 2.50 | 3.25 | 5.25 | 2.52 |
| Maximum | 3.25 | 4.50 | 6.00 | 3.00 |
| Minimum | 1.50 | 2.00 | 3.50 | 1.00 |

targets, $L_i$. The set $L_i$ is composed of different target object(s) for different tasks. For example, $L_1 = \{book\}, L_2 = \{notebook\}, L_3 = \{display, keyboard, mouse\}$, and $L_4 = \{paintings\}$ for reading, writing, browsing, and viewing painting tasks, respectively. We define the loss of attention when the target person diverts her/his VFOA from the specified target object. We also measure the span or duration of the VFOA of the target persons when she/he is involved in a task.

Humans cannot continue attending to some task and looking at the same target object and may need to divert their attention. There may be occasions when they avert their eyes from the current target. For example, in reading, they may not concentrate on looking at the pages when they turn the pages. These occasions are considered to be opportunities for the robot to try AA. The span or duration for humans to keep their VFOA may depend on the task. Thus, we performed observation experiments.

### A. Participants and Procedure

We videotaped 18 participants (14 males, mean age 28 years, standard deviation 4.9) doing four tasks: reading (four participants), writing (four), browsing (six), and viewing paintings (fixing attention on a painting in the room; four). The instructions were to concentrate on the task. The average recorded length for each person for reading, writing, browsing, and viewing paintings were 9, 9, 8, and 8 min, respectively.

### B. Data Collection

Our observation focused on measuring the span of VFOA on a task and finding the task-related contextual information. To measure the span of VFOA, we watched the recorded video data and manually annotated (using pause and restart) the period when a participant produces a consistent result on a task without loss of attention. Loss of attention was detected when the participant changed her/his current VFOA to another direction. For reading and writing, participants lost attention when "turning over the pages" and "stopping writing," respectively. For reading, writing, browsing, and viewing paintings, we detected 14, 10, 9, and 12 losses of attention, respectively. From the duration of these occasions, we estimated the span of VFOA for each task (see Table I).

From videos, we observed how head direction changed at the time of loss of attention for different tasks. To measure the head pose, we used the Seeing Machine faceAPI [33]. The minimum deviation of the head orientation can be used as a clue to detect the loss of attention. When humans lost attention from reading or writing, they mostly changed the tilt angle of their head and then changed the pan angle. The minimum deviation of tilt angle of the head for reading and writing were $14^\circ$ and $18^\circ$, respectively. However, when browsing, people normally changed the pan angle of their head to shift their attention to another direction. In this case, the minimum deviation was $17^\circ$. When participants attended to a painting, the loss of attention was detected using either the pan or tilt angle of the head. In such a case, the minimum deviation of pan and tilt angles were $14^\circ$ and $9^\circ$, respectively.

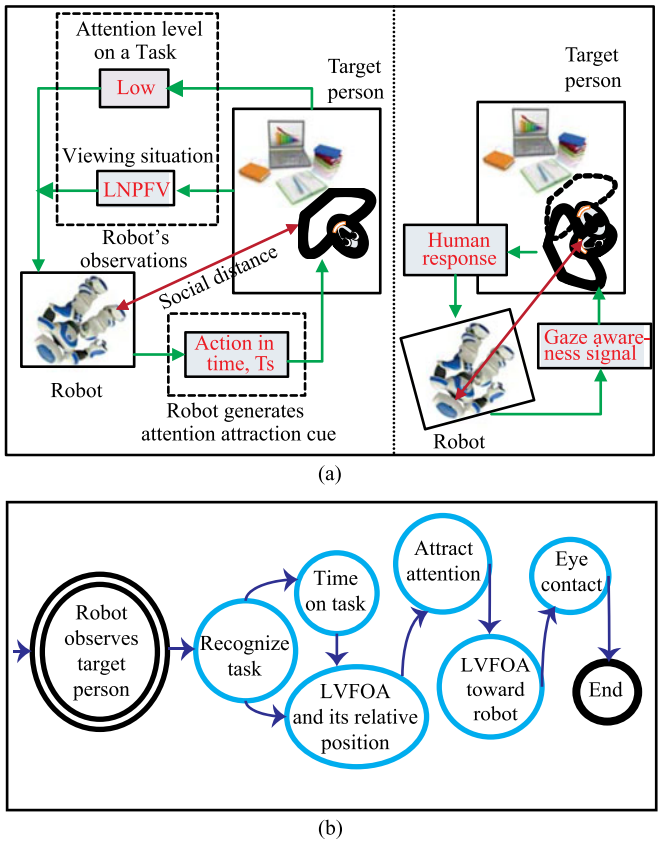



Fig. 1. (a) Abstract view of the proposed approach. (b) Basic steps of the proposed approach.

## IV. PROPOSED APPROACH

The proposed approach is illustrated in Fig. 1. In the *initiating interaction module* (see Fig. 1(a), left), the robot recognizes and tracks the target person's VFOA. If they are initially face-to-face, the robot generates an awareness signal and makes eye contact with the target person. Otherwise, the robot tries to attract the target person's attention by recognizing her/his current task. The robot detects the level of current VFOA until $T_s$ (where $T_s$ is the maximum span of sustained VFOA). We use the maximum values in Table I in the later experiments. The robot uses either a low or high level of current VFOA (depending on the person's current task) at time $t$ to generate an AA signal (weak or strong) depending on the viewing situation of her/his shifted VFOA. A person's field of view (FOV) is divided into central and peripheral visions. We represent the viewing situation (relation between the target person's gaze (face) direction and the

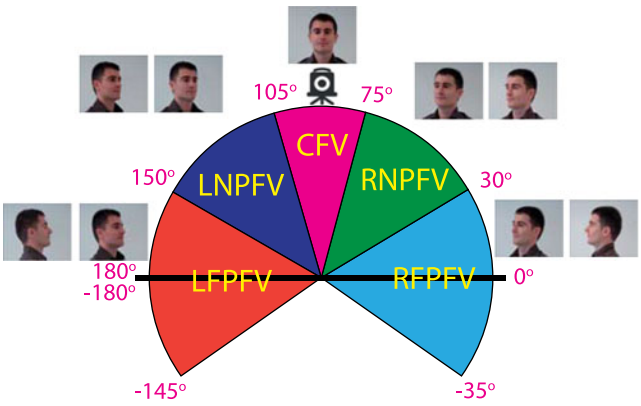



Fig. 2. Classification of head orientation into five angular regions. The faces shown from the GTAV Face Database [37].

robot position) by where the robot is seen in the FOV of the target person. We classify it into the three regions [34]–[36].

1) *Central Field of View (CFV)*: This FOV exists at the center of the human FOV. This zone is set to a 30° cone-shaped area (75° to 105° in Fig. 2).
2) *Near Peripheral Field of View (NPFV)*: It is defined as the 45° fan-shaped area on both sides of the CFV zone. On the right side of the CFV (30° to 75° in Fig. 2) it is defined as the right near peripheral field of view (RNPFV) and on the left side (105° to 150° in Fig. 2) the zone is called left near peripheral field of view (LNPFV).
3) *Far Peripheral Field of View (FPFV)*: This FOV exists on both sides at the edge of the human FOV. The right side of the RNPFV (−35° to 30° in Fig. 2) is known as right far peripheral field of view (RFPFV) and on the left side of the LNPFV (−145° to 150° in Fig. 2) it is the left far peripheral field of view (LFPFV).

If the VFOA is detected in CFV/LNPFV/RNPFV, then the robot generates a head turning action (weak signal). If the detected VFOA is in LFPFV or RFPFV, then the robot generates a head shaking action (strong signal). Fig. 2 illustrates these classified regions when the camera is placed in the CFV region. We define the angular regions based on the detected frontal and profile faces. For example, in the CFV region, we detect a frontal face only. In the other regions, we may detect two face patterns such as a half-pose right profile face and a full-pose right profile face in the LFPFV region.

When the robot succeeds in attracting the target person's attention, the *communication channel establishment module* [right part of Fig. 1(a)] tries to establish a communication channel with her/him. The robot determines the level of shifted attention toward it and generates an awareness signal toward the target person to indicate that it wants to communicate with her/him. The robot makes eye contact through eye blinking.

### A. Recognition of Visual Focus of Attention and Its Level

We are interested in detecting: *sustained attention* and *focused or shifted attention*. Focused or shifted attention is a short-term response to a stimulus or any other unexpected occurrence. The span or length of this attention is brief [38], and after a few seconds, it is likely that the person will look away, return to the previous task, or think about something else. Sustained attention is the level of attention that produces consistent results on a task over time. The duration of sustained attention depends on the task. We use the following cues to recognize VFOA and estimate its level.

*1) Visual Cues:*

a) *Head pose*: We use the Seeing Machines's faceAPI to detect and track the head pose, $h_p$ of the target person. Here, we classify the detected head poses into five angular regions: $h_p^{cfv}$, $h_p^{lnpfv}$, $h_p^{lfpfv}$, $h_p^{rnpfv}$, and $h_p^{rfpfv}$ if they are detected in the CFV, LNPFV, LFPFV, RNPFV, and RFPFV areas, respectively. The pan, and tilt angles of head poses are denoted by $h_p^p$, and $h_p^t$, respectively.

b) *Head movement*: To detect the head movement, $h_m$ we use the optical flow feature [39]. We generate a rectangular window circumscribing pixels with large flow values. If the total flow value in the window exceeds a threshold, we consider that a head movement ($h_m = 1$) cue is detected.

c) *Overlapping face window*: If a face is detected and overlap with the most recent head movement window, $h_m$ is more than 50%, we consider that an overlapping face window, $o_f$ is detected ($o_f = 1$). This detection means that the target person is turning her/his face toward the robot. Faces are detected using the Viola-Jones AdaBoost Haar-like face detector [40].

*2) Gaze Pattern:* A person's gaze pattern indicates her/his object of interest [41]. In general, human gaze patterns are classified into three viewing categories, distinguished by context [42]. *Spontaneous viewing* occurs when a person views the scene without any specific task in mind, i.e., when she/he is "just seeing" the scene. *Task or scene-relevant viewing* appears when a person views the scene with a particular question or task in mind (e.g., she/he may be interested in a particular painting in the museum). *Orientation of thought viewing* occurs when the subject is not paying much attention to where she is looking, but is attending to some "inner thought." We consider the former two. The gaze pattern indicates the pattern that is constructed by considering the effect of both head movements and eye gaze. We classify gaze pattern using the support vector machine (SVM) classifier [43].

*a) Iris center detection:* We use a multistage approach for detection of the center of the iris. First, the 3-D head tracker [44] detects the head position $H_t$ and its rectangular area in the image. Then, based on the head location and its area, we detect and track the facial feature points using the active shape model (ASM) [45] [see Fig. 3(b)]. The facial feature points are used to roughly estimate the eye regions on the face [see Fig. 3(c)]. The vector field of the image gradients (VFIG) within the eye regions are used to detect the iris center [red points in the Fig. 3(d)]. Although the facial feature points detected by the ASM model include the eye center points, their accuracy is not sufficient for iris center detection as in Fig. 3(b). We propose the VFIG iris center detection method to detect the iris center [see Fig. 3(d)] in the eye image as follows.

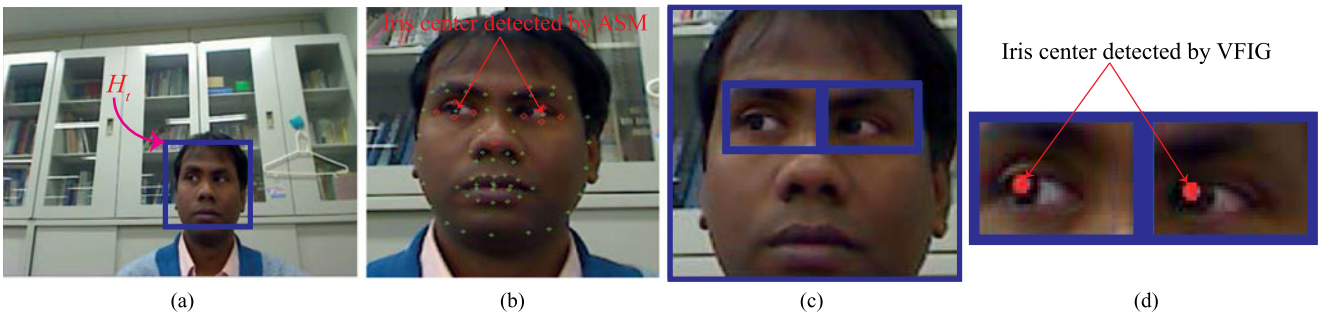



Fig. 3. (a) Detected head and its location in the image. (b) Extracted facial feature points. (c) Estimated eye regions based on facial feature points. (d) Detected iris center within eye area.

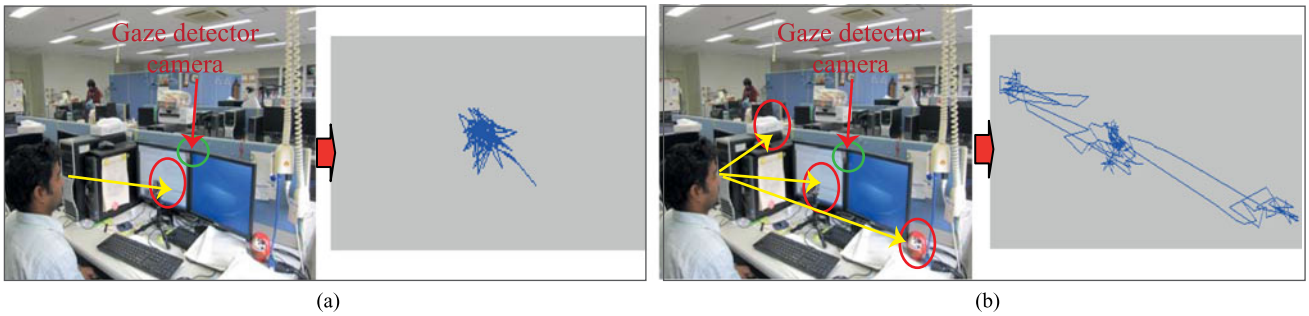



Fig. 4. Gaze pattern: (a) Task or scene-relevant viewing. (b) Spontaneous viewing.

Let $I_c$ be the possible iris center and $I_{g_i}$ the gradient vector at position $I_{x_i}$. If $I_{d_i}$ is the normalized displacement vector, then it should have the same absolute orientation as the gradient $I_{g_i}$. We can determine the optimal center ${I_c}^*$ of the iris (darkest position of the eye) by computing the dot products of $I_{d_i}$ and $I_{g_i}$ and finding the global maximum of the dot product over the eye image:

$$ {I_c}^* = \underset{I_c}{\operatorname{argmax}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( {I_{d_i}}^T I_{g_i} \right)^2 \right\} \tag{1} $$

where

$$ I_{d_i} = \frac{(I_{x_i} - I_c)}{(||I_{x_i} - I_c||_2)} \tag{2} $$

$i = 1, \ldots, N$, and $\forall i : ||I_{g_i}||_2 = 1$. The displacement vector $I_{d_i}$ is scaled to unit length in order to obtain an equal weight for all pixel positions in the image.

*b) Gaze pattern detection:* To obtain the gaze pattern $G_p$, we consider the translation movement of the head in the image and the position change of the iris in the eye. Let $H_0$ indicates the initial head position and $E_t$ be the eye gaze position (relative iris position in the eye) for $t$th frame. If $T_{H_t}$ indicates the translation vector of the head movement from $H_0$, the gaze point, $Q_t$ for the $t$th frame is determined as follows:

$$ Q_t = E_t + (H_0 + T_{H_t}). \tag{3} $$

Then, $G_p = \{Q_0, Q_1, \ldots, Q_{L-1}\}$ denotes the gaze pattern for $L$ frames. The gaze pattern of a person viewing a particular point in the scene (task or scene-relevant viewing) is in Fig. 4(a) and viewing three different points (spontaneous viewing) is in Fig. 4(b).

*c) Feature representation and classification:* The feature vector is represented from the gaze pattern data through normalization using the center of gravity for the pattern. Let us assume $C_m$ to be the center of gravity and $r_t$ to be the Euclidean distance from $C_m$ to the gaze point $Q_t$:

$$ r_t = |Q_t - C_m| \tag{4} $$

where $t = 0, 1, \ldots, L-1$. We sort the distance values $r_t$ into descending order and construct the feature vector for the classifier. We use the multiclass SVM classifier to classify the gaze patterns. To train, we collect gaze data and construct gaze patterns for scene-relevant and spontaneous viewing. The learned SVM model classifies the gaze pattern into either *spontaneous viewing*, $S_l$ or *task or scene-relevant viewing*, $T_l$.

To evaluate the performance, we collected 80 gaze pattern data (40 for the *task or scene-relevant viewing* and 40 for the *spontaneous viewing*) from ten people (from 0.5 to 2 m distances between the people and the camera). For the *task or scene-relevant viewing*, the participants looked at any specific object in the scene. For the *spontaneous viewing*, the participants looked around aimlessly. Our system automatically collected 150 frames per gaze pattern data. We randomly selected 40 samples (20+20) for training the classifier and the rest were used for testing. Thirty-seven of 40 samples were recognized correctly (recognition rate of 92.5%).

*3) Task Context:* The task context is determined by recognizing the task in which the target person is involved. For instance,

if the target person is involved in a "reading" task, then the contextual cue such as "downward head" indicates that her/his attention is toward the book. However, the "page turn over," or "upward the head" indicates that the person loses her/his attention.

*a) Task recognition:* Given a video sequence, we extract the histogram of orientation gradient (HOG) feature [46] for each frame. The HOG features are combined for ten consecutive frames to build an HOG feature pattern, $\mathrm{HOG}_P$. Thus

$$\mathrm{HOG}_P = F_0 + \sum_{i=1}^{9} |F_{i-1} - F_i| \tag{5}$$

where $F_0$ and $F_i$ are the HOG features of the first and $i$th frames, respectively. The first frame captures the human appearance features involved in a task, and the rest of the HOG feature frames indicate the change of behavior pattern in the task. Thus, the $\mathrm{HOG}_P$ feature captures the appearance of the task and the task related behavior. Each bin in the histogram represents the number of edges with orientation within a given angular range. The angular range is set to $20^\circ$ and we use unsigned gradients. Thus, the bin size $= 180/20 = 9$. With this bin size, we create the $\mathrm{HOG}_P$ feature vector of size 90. A multiclass SVM [43] is learned using our $\mathrm{HOG}_P$ feature. In detection, we use the SVM classifier in the recognition mode. We use the dataset as described in Section III-A to divide training and test videos for the different tasks. The classification system generates training/test samples from the training/test videos by taking ten successive frames for each sample. The classifier is trained using 8270 samples for the four different tasks. To evaluate the performance, we use 10275 test samples from the same dataset for the four different tasks. Among them, 9601 samples are correctly detected with 93.4% accuracy.

*b) Contextual cues:* After recognizing the task (or current VFOA) of the target person, we use the related contextual cues of the task to recognize the level of attention. For each task, we use the task related VFOA span ($T_s$) to determine how long the robot should wait or within which period of time the robot interacts with the target person. We also define some task specific cues to determine the level of attention. With reading, we use the page turn over, $P_t$, and deviation in tilt angle, $d_{h_p^t}$ cues to measure the LVFOA. For writing, LVFOA is estimated using stop writing, $W_s$, and $d_{h_p^t}$ cues. For browsing, we use deviation in pan angle, $d_{h_p^p}$ to determine the LVFOA. For viewing paintings, $d_{h_p^t}$, and $d_{h_p^p}$ cues are used for estimating LVFOA. Page turn over, and stop writing cues are detected using a threshold value for the resultant magnitude of the optical flow pattern. The positions of these cues are determined with respect to the relative position of the person's body. The body-tracking system is described in [47]. A threshold value of 100 is set to detect the page turn over. If the resultant magnitude is approximately 0 (in the experiment we set it 2) for ten consecutive frames, then the stop writing behavior is detected. We consider ten consecutive frames because people may stop their "writing motion" for a moment without shifting attention. The threshold values are set experimentally on a trial-and-error basis from the dataset in Section III-A. We evaluated the system on this dataset to detect the loss of attention (low VFOA) using task context. The loss of attention period in the video data are annotated manually with the start and end frame numbers. We assumed the loss of attention as positive samples. Among 45 cases of loss of attention in the four different tasks, our system detected it 37 times correctly with a detection rate of 82.2%. We calculated the true negative, false positive, and false negative of the system. Here the true negative indicates a high LVFOA on a task. Table II shows the performance of the system for determining the positive and negative samples.

TABLE II
PERFORMANCE OF THE SYSTEM FOR RECOGNIZING THE LEVEL OF VFOA (CORRECTLY RECOGNIZED SAMPLES/TOTAL NUMBER OF SAMPLE)

| | Loss of VFOA | High VFOA | False positive | False negative |
|---|---|---|---|---|
| Reading | 11/14 | 15/15 | 0 | 3 |
| Writing | 8/10 | 10/11 | 1 | 2 |
| Browsing | 8/9 | 9/11 | 2 | 1 |
| Viewing | 10/12 | 12/13 | 1 | 2 |
| **Avg. accuracy** | **82.2%** | **92.0%** | **11.1%** | **16.0%** |

### B. Level of Sustained Visual Focus of Attention

The level of VFOA is classified into two categories (low or high) based on the contextual cues, and gaze pattern. When the level of attention goes low, the system assumes that a loss of VFOA is detected. For different tasks, the attention level is detected as follows:

$$SA_{L,\text{read}} \leftarrow S_l \vee P_t \vee (d_{h_p^t} \geq 14^\circ) \tag{6}$$

$$SA_{L,\text{write}} \leftarrow S_l \vee W_s \vee (d_{h_p^t} \geq 18^\circ) \tag{7}$$

$$SA_{L,\text{browse}} \leftarrow S_l \vee (d_{h_p^p} \geq 17^\circ) \tag{8}$$

$$SA_{L,\text{viewing}} \leftarrow S_l \vee (d_{h_p^p} \geq 14^\circ) \vee (d_{h_p^t} \geq 9^\circ). \tag{9}$$

$S_l$ indicates spontaneous viewing. If spontaneous viewing is detected then it is assumed that the person has no particular attention on a task. Thus, a low attention level is detected. For reading and writing tasks, in addition to head pose changes, we consider the "page turn over," $P_t$, and "stop writing," $W_s$, behaviors for detection of low attention level. For (6)−(9), if the specific head pose changes and stability is greater than or equal to three frames, then the level of attention is low for the corresponding task. Otherwise, the attention level is high, and the current attentional focus remains on the task. The threshold values for different head poses in (6)−(9) are determined through human-based experiment (see Section III-B).

### C. Detection of Focused/Shifted Attention

Focus/shifted attention is detected in two phases. First, to attract the target person's attention, the robot detects *focus/shifted attention from sustained VFOA*. Second, after sending an AA signal, the robot needs to detect *focus/shifted attention toward it*.

*1) Shifted Attention from Sustained VFOA:* To initiate a polite social interaction, the robot should attract the target person's

attention depending on her/his current sustained VFOA. The robot first detects the loss of sustained VFOA of the target person using one of (6)−(9). In some cases, such as reading, writing, and browsing, the robot attracts the target person's attention when her/his sustained LVFOA is low. However, for viewing paintings the robot attracts attention when her/his sustained LVFOA is high. After attracting attention, the robot detects the shifted VFOA of the target person. Depending on the environmental factors and the target person's mental focus, the sustained VFOA can shift into one of the five regions: CFV, LNPFV, LFPFV, RNPFV, and RFPFV. The shifted VFOA region is detected using the pan angle of head pose, $h_p^p$.

*2) Focused/Shifted Attention Toward the Robot:* The detection of focused/shifted attention toward the robot is an important cue for the robot to make eye contact with the target person. If the robot and the target person are not facing each other, then the robot sends some AA signal and waits for her/his attention toward it. When the target person shifts or turns her/his attention toward the robot, it needs to detect focused/shifted attention toward it. To make successful eye contact, the robot classifies the level of focused/shifted attention into three categories: Low, Medium, and High. The robot sends an AA signal toward the target person and analyzes the input video images frame-by-frame to detect whether the target person is moving toward it. If the target person is turning to look at the robot from her/his current focus of attention, then some contiguous $h_m$ windows will be detected surrounding the head. Depending on the detected visual cues (see Section IV-A), the level of focused/shifted VFOA is classified according to the following.

When none of the visual cues are detected except for head movement as in (10), we assume that the focused/shifted attention level is low, $FA_L$:

$$FA_L \leftarrow ((N_{h_m} \geq 1) \wedge (o_f = 0) \wedge (N_{f_s} \leq 1) \wedge (h_p^p \neq \text{CFV})) \quad (10)$$

where $N_{h_m}$ is the number of contiguous head movement windows in the subsequent frames (in frames), $o_f$ indicates whether any overlapping window is detected (1) or not (0), $h_p^p$ is the estimated pan angle of head pose, and $N_{f_s}$ is the face stability detection result in the subsequent frame (in frames) after detection of the overlapping window.

If the head movement is detected with an overlapping window of face within the contiguous head movement area, the level of attention is medium, $FA_M$:

$$FA_M \leftarrow ((N_{h_m} \geq 5) \wedge (o_f = 1) \wedge (N_{f_s} \leq 1) \wedge (h_p^p = \text{CFV/LNPFV/RNPFV})). \quad (11)$$

When the visual cues are successfully detected and stable, we assign the high level of attention, $FA_H$:

$$FA_H \leftarrow ((N_{h_m} \geq 5) \wedge (o_f = 1) \wedge (N_{f_s} \geq 5) \wedge (h_p^p = \text{CFV})). \quad (12)$$

When all conditions on the right-hand side of (10)−(12) are satisfied, the corresponding level of attention is detected. The detected level of attention is used in the subsequent awareness generation, and making successful eye contact. Based on [3], different threshold values for parameters $N_{h_m}$, $o_f$, and $N_{f_s}$ in the (10)−(12) are fixed manually.

### D. Initiating Interaction Based on Visual Focus of Attention

In polite social interaction, humans usually raise or turn their head first toward the person with whom they would like to communicate. However, if the target person's attentional focus toward a task is high, humans try with stronger actions (e.g., turning head more than once, waving the hand, coming closer to the person and turning the head, or even using voice) to attract her/his attention. Robots should use the same conventions. In this research, the robot detects the target person's level of sustained VFOA and the region of shifted VFOA to choose the appropriate control signal. We chose the head turning action (to look at the person) as the weak signal when the sustained VFOA attention level is low and the shifted VFOA is in the either CFV/LNPFV/RNPFV area. We use the head shaking action when the sustained VFOA attention level is low and the shifted VFOA is in the LFPFV/RFPFV area. We also use the head shaking action when the sustained VFOA level is high and the robot needs to attract the attention of the target person. We use the head shaking action as a strong AA signal because abrupt object motion draws people's attention [48]. See [49] for a detailed description of the cues. The visual stimuli by the robot's nonverbal behaviors cannot affect a person if she/he is in a position where she/he cannot see the robot action. Thus, we do not consider situations when the shifted VFOA is in the out of the FOV area.

### E. Establishing Communication Channel

To establish a communication channel, the robot needs to make the person notice that it is looking at her/him. The robot should be able to display its awareness through some actions (for example, facial expressions, eye blinking, or nodding). We adopt eye blinking to create such awareness since it is one of the most important cues for forming a person's impressions [30]. These actions are designed to evoke the target person's sense of being looked at by the robot.

*a) Eye blinks:* If the robot successfully attracts the target person's attention, or she/he notices the robot's action, she/he will direct her/his gaze at the robot. The robot recognizes her/his face while she/he is looking at it. After detecting the face stability of the target person, (i.e., $FA_H = 1$), the robot starts blinking its eyes about three times (1 blink/s) to establish a communication channel. Eye blinks are produced by rapid closing and opening of the eyelid of the CG images, and displayed through the LED projector onto the robot's eyes.

## V. Evaluation

We conducted an experiment using our static robot head for four different tasks. We then implemented the proposed system into a commercial robot.

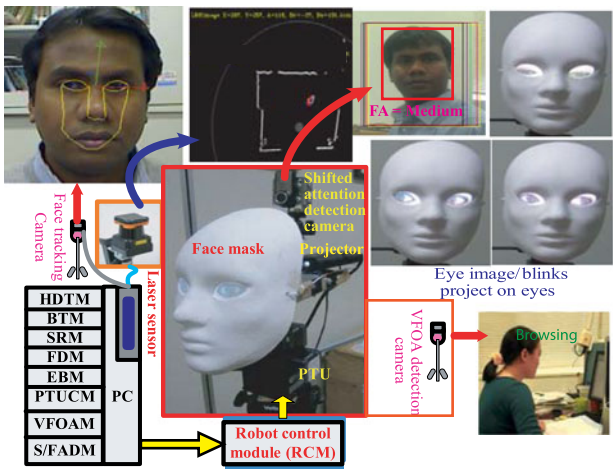



Fig. 5. Experimental robotic platform.

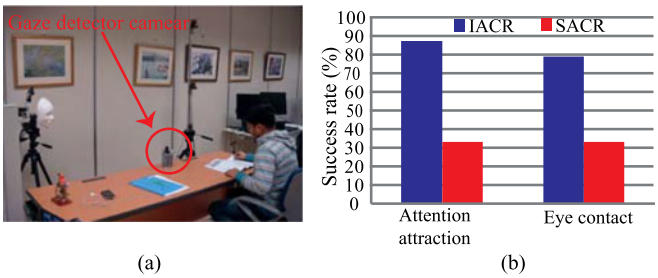



Fig. 6. (a) Experimental scene. (b) Success rate of the system.

### A. Static Robotic Head Interaction

The proposed human–robot interaction scenario based on the level of VFOA of the target human was implemented on a static robot head. We conducted experiments to verify that the proposed system caused less disturbances and was more successful at initiating interactions with the target while she/he was involved in some task.

*1) Methods:*

*a) Participants:* The 24 unpaid participants (19 males) were students at Saitama University (mean age 30.7 years and standard deviation 5.4).

*b) Apparatus:* Fig. 5 shows an overview of our robotic platform [50].

The system includes the head detection and tracking, situation recognition, body tracking, face detection, eye blinking, pan-tilt unit control, VFOA detection, and shifted/focused attention detection modules.

*c) Procedure:* The participants were asked to pay attention to their tasks and wore headphones with music to avoid the sound effect of the pan-tilt movement of the robot. We used two video cameras to capture all interactions. Fig. 6(a) shows the experimental environment. For viewing paintings, the simple head turning action fails if the robot exists in the LFPFV/RFPFV area because the robot tried to attract attention when people were focusing their attention on the paintings. Thus, in this case, we placed the robot in LNPFV area of the visitor.

*d) Independent measures:*

1) *Intelligent attention control robot (IACR).* The robot determines the level of attention to the current task and considers the situation of the target person. For reading, writing, and browsing, the robot attracts attention when the target person loses her/his attention on the task (i.e., sustained VFOA is low). For viewing painting, the robot attracts attention when the target person focuses her/his attention on the task (i.e., sustained VFOA is high). The robot sends the head turning action when the attention is shifted to either CFV/LNPFV/RNPFV region. However, if the attention is shifted to LFPFV/RFPFV region, then the robot uses the head shaking action.
2) *Simple attention control robot (SACR).* This robot does not consider the target person's VFOA. After detecting the target person, the robot tries to attract her/his attention. When the target participant is involved in a task, after 5–30 s, we start the experimental robot. The body-tracking module of the robot immediately detects the target person and tries to attract her/his attention. The robot uses two types of AA signals. If the head turning fails, then the robot uses the head shaking action to attract her/his attention.

*e) Dependent measures:*

1) *Impression of the robots*: Participants filled out a four item questionnaire for each condition (after two interactions). The measurement was a rating on a Likert scale of 1 to 7 (1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, 4: Neither agree nor disagree, 5: Somewhat agree, 6: Agree, 7: Strongly agree). The questions were: Did you feel that the robot attracted your attention? Was the robot's interruption acceptable to you for attracting your attention? Did the robot's interruption to attract your attention disturb you? Did you make eye contact with the robot?
2) *Success rate*: We counted the number of times the target participants looked at the robot after AA actions. We counted the number of times that the robot was successful in detecting attracted attention from the participants.

*f) Data analysis:* Our hypotheses were the following:

1) The proposed method (IACR) outperforms the other (SACR) in attracting the participant's attention toward the robot.
2) The proposed method is more acceptable than SACR in attracting the participant's attention.
3) The proposed method creates less disturbance than SACR in attracting the participant's attention.
4) The proposed method outperforms SACR in establishing a communication channel.

Participants were divided into four groups and asked to do the task assigned to each group: reading (12 participants), writing (four), browsing (four), and viewing paintings (four participants; each participant was asked to randomly fix her/his attention on a painting in the environment). The order of all experimental

TABLE III
QUESTIONNAIRE RESULTS WHERE 7 IS STRONGLY AGREE

| Questionnaire (detailed are in Section V-A1e) | IACR | | SACR | |
|---|---|---|---|---|
| | Median | Mode | Median | Mode |
| Robot attracted your attention? | 6 | 6 | 3 | 3 |
| Robot's interruption acceptable? | 6 | 6 | 3 | 2 |
| Robot's interruption disturbed you? | 3 | 3 | 4.5 | 5 |
| Did you make eye contact with robot? | 6 | 6 | 3.5 | 3 |

trials was counterbalanced. We use the Wilcoxon signed rank test to test the above hypotheses.

*2) Results:*

*a) Impression:* The questionnaire results are shown in Table III.

IACR is more effective in attracting attention than SACR ($Z = -4.136$, $p < 0.001$, neg. rank $= 11.5$, pos. rank $= 0.0$). The result supports hypothesis 1. The robot's interruption time to attract the target participant's attention was more appropriate and acceptable ($Z = -4.208$, $p < 0.001$, neg. rank $= 12.89$, pos. rank $= 3.5$). The result verifies hypothesis 2. The participants felt less disturbed when the robot considered her/his attention for interaction behaviors ($Z = -4.194$, $p < 0.001$, neg. rank $= 5.00$, pos. rank $= 12.83$). The result supports hypothesis 3. The proposed method is more effective for establishing a communication channel ($Z = -4.047$, $p < 0.001$, neg. rank $= 11.0$, pos. rank $= 0.0$). The result supports hypothesis 4.

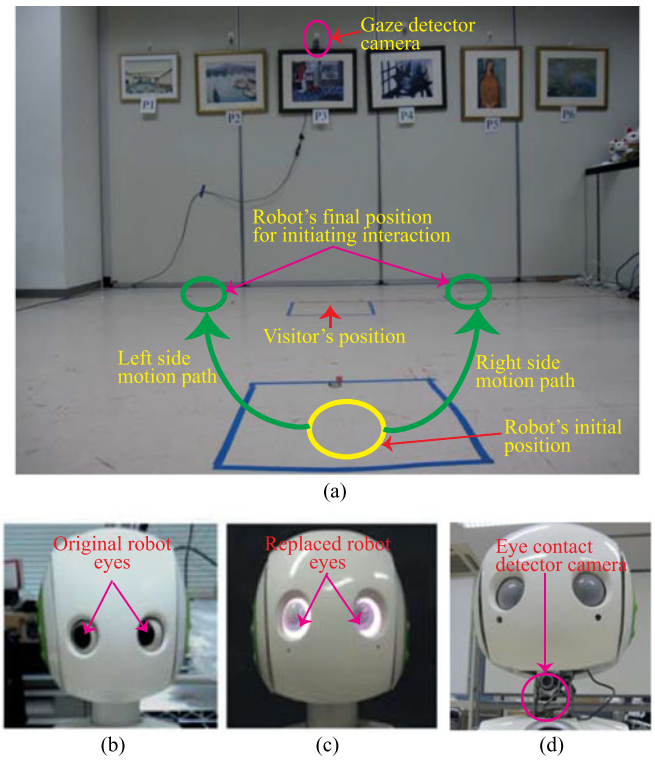
*b) Success rate:* Fig. 6(b) shows the success rate of our system for two types of robotic behaviors. A two-tailed $Z$ test of proportions ($Z = 3.837$, $p < 0.001$) showed that the proposed robot, IACR (87.5%, 21 times attracted the attention of the target participant among 24 trials) is significantly more successful than SACR (33.3%, eight times attracted the attention of target participant among 24 trials) in attracting the target participants' attention. In the eye contact stage, there were no significant differences. However, since the overall success of eye contact depends on the success rate of the AA stage, the results revealed that the proposed method is better in making eye contact with the target participant.

## B. Robovie-R3 in a Museum Scenario

We conducted an experiment to verify that the proposed system is useful to initiate interaction between visitors and the robot in a museum scenario. We assumed that a visitor observes paintings in a museum and fixes her/his attention at a particular painting after a few moments. The robot is situated far from the paintings; therefore, it may not interfere with the visitor's movement and attention. When the robot detects a high attention level of the visitor, it classifies her/his head orientation to select from which side or position the robot should initiate interaction. The robot classifies the visitor's head orientation into five angular regions: LFPFV, LNPFV, CFV, RNPFV, and RFPFV. Then, the robot selects a suitable motion path and position for initiating interaction (see Fig. 7).

(a) (b)

Fig. 7. Robot's position for interaction: (a) Robot selects the left-side interaction path when the visitor's orientation of attention is detected in the LNPFV area. (b) Robot selects the right-side interaction path when the visitor's orientation of attention is detected in the RNPFV area.




(a) (b) (c) (d)

Fig. 8. (a) Experimental environment. (b) Original Robovie-R3's eyes. (c) Replaced eyes for gaze communication. (d) Position of the camera for eye contact detection.

*1) Methods:*

*a) Participants:* The 12 unpaid participants (ten males) were students at Saitama University (mean age 29.7 years and standard deviation 4.8).

*b) Apparatus:* We hung six paintings (P1–P6) on the wall at the same height [see Fig. 8(a)]. These paintings were placed to make participants look in various locations and fix VFOA to a particular painting from a fixed standing position. A USB camera (Logicool) was located on the top of the painting (P3) to detect the visitor's gaze and head orientation. Paintings P2, P3, and P4 were placed in the LNPFV, CFV, and RNPFV areas, respectively. If the robot is present in front of a visitor, she/he may be attracted by the robot even though it does not perform any action due to its human-face like appearance [51].

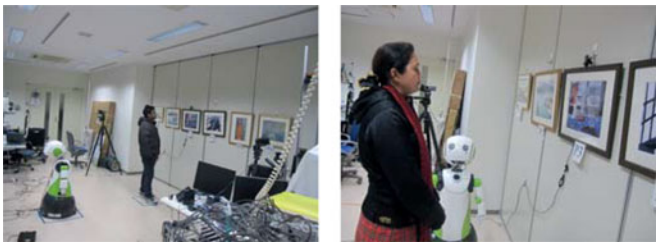


Fig. 9. Snapshots of the experimental scene.

Thus, the robot (Robovie-R3) was situated behind the visitor [see Fig. 8(a)] far from the paintings so that the robot might not interfere with her/his attention. We replaced the Robovie-R3 eyes [see Fig. 8(b)] with the computer graphics generated projected eyes for gaze communication, because it was shown to be effective for gaze communication [52] [see Fig. 8(c)]. To confirm eye contact of the visitor with the robot when it turns its head toward the visitor, we put a USB camera in the lower side of the robot head [see Fig. 8(d)]. We designed two possible motion paths to move the robot autonomously toward the visitor and settle it at a suitable position for initiating interaction with her/him depending on her/his orientation of attention.

*c) Procedure:* In the interaction scenario, each participant was asked to stand in a fixed position, to freely move her/his gaze and head orientation among the paintings, and finally to fix her/his attention to picture P2, P3, or P4 (see Fig. 9). The robot determined the visitor's gaze pattern, gaze point, and head orientation. For simplicity, among five head orientation regions (see Section V-B), we considered only three categories of the visitor's head orientation: RNPFV (when the visitor looks at the picture P4), CFV (when the visitor looks at the picture P3), and LNPFV (when the visitor looks at the picture P2). When the robot detects the visitor's orientation of attention in the LNPFV or RNPFV region, then the robot selects the left-side or right-side interaction path and position. However, when the robot detects visitor's orientation of attention in the CFV zone, then the robot can select either right-side or left-side interaction path and position. We changed the motion path and orientation of attention so that each participant experienced 6 trials. In each trial, the robot tried to select the appropriate motion path and position and to attract the participant's attention by moving its head toward her/him, because head movement is an effective cue for attracting the attention of the target human [35]. Two video cameras were used to capture all interactions.

*d) Independent measures:* Two methods were considered.

1) *Method 1 (M1):* To initiate interaction with the visitor, the robot selects the motion path based on the visitor's orientation of attention so that the robot and the visitor can be face-to-face.
2) *Method 2 (M2):* The robot appears from the opposite direction of the visitor's orientation of attention in the LNPFV or RNPFV area cases.

*e) Dependent measures:*

1) *Visitors' impression:* We asked participants to fill out a questionnaire for each method (after interactions). The measurement was a rating on a Likert scale of 1 (strongly disagree) to 7 (strongly agree). The questionnaire had two subjective questions:
   a) Q1: Did you feel that you made eye contact with the robot during the initiation of interaction?
   b) Q2: Did you think that the robot's approach was effective for initiating an interaction?
2) *Success rate:* From the videos and experimental site, we observed how many times the robot detected the visitor's gaze point and established a successful interaction. The success rate was measured by the ratio of the number of successful interactions to the total number of attempts that the robot made.

TABLE IV
VISITOR'S IMPRESSION FOR QUESTIONNAIRE 1 (Q1)

| | Picture-2 (P2) | | Picture-3 (P3) | | Picture-4 (P4) | |
|---|---|---|---|---|---|---|
| | M1 | M2 | M1 | M2 | M1 | M2 |
| Median | 6.5 | 5 | 6 | 5.5 | 6 | 4.5 |
| Mode | 7 | 5 | 6 | 6 | 6 | 4 |
| $Z$ | $-2.831$ | | $-1.930$ | | $-2.609$ | |
| $P-$value | 0.005 | | 0.054 | | 0.008 | |
| neg. rank | 5.50 | | 4.17 | | 5.67 | |
| pos. rank | 0.0 | | 3.0 | | 3.5 | |

TABLE V
VISITOR'S IMPRESSION FOR QUESTIONNAIRE 2 (Q2)

| | Picture-2 (P2) | | Picture-3 (P3) | | Picture-4 (P4) | |
|---|---|---|---|---|---|---|
| | M1 | M2 | M1 | M2 | M1 | M2 |
| Median | 6 | 5 | 6 | 6 | 6 | 5 |
| Mode | 6 | 5 | 6 | 6 | 6 | 4 |
| $Z$ | $-2.836$ | | $-1.897$ | | $-2.687$ | |
| $P-$value | 0.005 | | 0.058 | | 0.007 | |
| neg. rank | 5.50 | | 4.08 | | 5.0 | |
| pos. rank | 0.0 | | 3.5 | | 0.0 | |

*f) Data analysis:* The experiment was performed in a within-participant design, and the order of all experimental trials was counterbalanced. We compared the Likert scale data of questionnaire measures using the Wilcoxon signed rank test.

*2) Results:*

*a) Impression:* Subjective measures for both Method 1 and Method 2 are shown in Tables IV and V, respectively. We considered three different gaze points when the visitor was looking at pictures P2, P3, and P4 and compared the interaction impression with the robot.

For questionnaire Q1 (see Table IV), the differences between the two methods (M1 and M2) were statistically significant when the visitors looked at picture P2 ($Z = -2.831$, and $p < 0.01$). For picture P4, the differences between two methods were also statistically significant ($Z = -2.609$, and $p < 0.01$). For picture P3, we did not find any significant differences. Thus, for making a successful eye contact during interaction initiation, the robot should select the left-side and right-side motion path when the visitor looks at picture P2 and P4, respectively. However, when the visitor looks at picture P3, the robot may select either the left-side or right-side motion path for making eye contact.

For questionnaire Q2 (see Table V), the Wilcoxon signed rank test shows significant differences when the visitors looked at picture P2 ($Z = -2.836$ and $p < 0.01$). For picture P4, the differences were also statistically significant ($Z = -2.687$, and $p < 0.01$). For picture P3, no significant differences were found. Thus, for initiating an interaction scenario, the robot should select the left-side and right-side motion path when the visitor looks at picture P2 and P4, respectively. When the visitor looks at picture P3, the robot may select either the left-side or right-side motion path.

*b) Success rate:* Each visitor experienced three trials as the target for each method. We observed a total of 72 interactions ($12 \times 3 \times 2$). Among 72 interactions, our system was able to detect 66 times the visitor's gaze point and make a successful interaction at a rate of 91.7%. The proposed system is effective for initiating interaction with visitors.

## VI. Discussion

We developed a robot that can attract the attention of a particular person and establish a communication channel with her/him depending on her/his LVFOA. The proposed method is effective in initiating an interaction process to a target person in terms of initially attracting her/his attention, and establishing a communication channel with her/him. Our results with a commercial robot (Robovie-R3) in a museum scenario confirms the proposed method is making polite and successful interaction.

A robot naturally initiating interaction to control someone's attention is one of the major capabilities to be implemented in social robots. In the real world robots may wait for people to approach them, which is one strategy for robots to initiate interaction. Alternatively, robots can proactively approach people to initiate interaction. In this research we have implemented the proposed system in a museum scenario.

The current system has the following limitations. First of all, it needs cameras in the environment to observe people's gaze patterns. This may be acceptable in the museum scenario. However, this should be modified so that it can be used in various situations. Present robots make loud noises when they move. If they move, they attract people's attention and interrupt their work. Thus, in our experiments, we asked the participants to wear headphones with music to mitigate these noise effects. However, if such robots were developed to move as quietly as humans, they could move to positions where it could more easily observe the target person with onboard cameras. The robot can first estimate attention level by human head movements and other actions of the body from a fixed position. If the robot is not sure about the attention level of the human subjects, it can move to obtain more precise information.

We use a constant value for the maximum span of sustained VFOA, $T_s$. This is the maximum time span that the robot will wait if people do not show their low attention level or the robot cannot detect their low attention level. People often exhibit a low attention level before this time. Thus, if we set it to be large enough as we did in the experiments, there would be no serious problems. However, it would be better if $T_s$ could be adjusted depending on the situation. Humans determine the maximum waiting time by taking account of various factors. If the person seems to be really concentrating on her/his current work, we may wait longer. However, if we cannot wait too long, we may interrupt.

## Acknowledgment

The authors would like to thank the participants of Saitama University, Japan. This research was supported by JST, CREST.

## References

[1] E. Andre, T. Rist, S. V. Mulken, M. Klesen, and S. Baldes, "The automated design of believable dialogues for animated presentation teams," in *Embodied Conversational Agents*. Cambridge, MA, USA: MIT Press, 2000, pp. 220–255.

[2] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 928–938, Jul. 2002.

[3] D. Das, M. M. Hoque, T. Onuki, Y. Kobayashi, and Y. Kuno, "Vision-based attention control system for socially interactive robots," in *Proc. IEEE Int. Symp. Robot Human Interactive Commun.*, Paris, France, Sep. 9–13 2012, pp. 496–502.

[4] Z. Yücel, A. A. Salah, Ç. Meriçli, T. Meriçli, R. Valenti, and T. Gevers, "Joint attention by gaze interpolation and saliency," *IEEE T. Cybernetics*, vol. 43, no. 3, pp. 829–842, Jun. 2013.

[5] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: an ethnographic study," in *Proc. 6th Int. Conf. Human-Robot Interaction*, 2011, pp. 37–44.

[6] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends Cognitive Sci.*, vol. 4, no. 2, pp. 50–58, 2000.

[7] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, Jan. 2011.

[8] D. Das, Y. Kobayashi, and Y. Kuno, "Attracting attention and establishing a communication channel based on the level of visual focus of attention," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 2194–2201.

[9] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 4, pp. 205–218, Dec. 2003.

[10] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head-pose and gaze direction measurement," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2000, pp. 2127–2132.

[11] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," in *Proc. 3rd Int. Conf. Visual Inf., Inf. Syst.*, 1999, pp. 761–768.

[12] S. Asteriadis, K. Karpouzis, and S. D. Kollias, "Robust validation of visual focus of attention using adaptive fusion of head and eye gaze patterns," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2011, pp. 414–421.

[13] S. Asteriadis, K. Karpouzis, and S. D. Kollias, "Visual focus of attention in non-calibrated environments using gaze estimation," *Int. J. Comput. Vision*, vol. 107, no. 3, pp. 293–316, 2014.

[14] J. S. Babcock and J. B. Pelz, "Building a lightweight eyetracking headgear," in *Proc. Symp. Eye Tracking Res. Appl.*, 2004, pp. 109–114.

[15] R. Vertegaal, R. Slagter, G. C. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more the conversational agents than meets the eyes," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2001, pp. 301–308.

[16] S. O. Ba, H. Hung, and J.-M. Odobez, "Visual activity context for focus of attention estimation in dynamic meetings," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 28–Jul. 2, 2009, pp. 1424–1427.

[17] M. Voit and R. Stiefelhagen, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in *Proc. 10th Int. Conf. Multimodal Interfaces*, Oct. 20–22 2008, pp. 173–180.

[18] R. Vertegaal, J. Shell, and S. Lahlou, "Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects," *Soc. Sci. Inf.*, vol. 47, no. 2, pp. 275–298, 2008.

[19] C. Yu, P. W. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *TiiS*, vol. 1, no. 2, p. 13, 2012.

[20] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles

using gaze cues," in *Proc. Int. Conf. Human-Robot Interaction*, 2009, pp. 61–68.
[21] M. Staudte and M. W. Crocker, "Visual attention in spoken human-robot interaction," in *Proc. Int. Conf. Human-Robot Interaction*, 2009, pp. 77–84.
[22] Z. Kang and S. J. Landry, "Eye movement analysis of a multielement target tracking task: Maximum transition-based agglomerative hierarchical clustering algorithm," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 1, pp. 13–24, Feb. 2015.
[23] R. J. K. Jacob, "The use of eye movements in human-computer interaction techniques: What you look at is what you get," *ACM Trans. Inf. Syst.*, vol. 9, no. 2, pp. 152–169, 1991.
[24] L. Fletcher and A. Zelinsky, "Driver inattention detection based on eye gaze - road event correlation," *Int. J. Robot. Res.*, vol. 28, no. 6, pp. 774–801, 2009.
[25] M. Johansson, G. Skantze, and J. Gustafson, "Head pose patterns in multiparty human-robot team-building interactions," in *Proc. 5th Int. Conf. Social Robotics*, 2013, pp. 351–360.
[26] U. Weidenbacher, G. Layher, P. Bayerl, and H. Neumann, "Detection of head pose and gaze direction for human-computer interaction," in *Proc. Int. Tut. Res. Conf. Perception Interactive Technol.*, 2006, pp. 9–19.
[27] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu, "Development and evaluation of an interactive humanoid robot 'robovie'," in *Proc. IEEE Int. Conf. Robot. Autom.*, Washington, DC, USA, 2004, pp. 1848–1855.
[28] B. Mutlu, J. K. Hodgins, J. Forlizzi, and T. Shiwa, "A stroytelling robot: Modeling and evaluation of human-like gaze behavior," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2006, pp. 518–523.
[29] M. V. Cranch, *The Role of Orienting Behavior in Human Interaction*, A. H. Esser, Ed. New York, NY, USA: Plenum Press, 1971.
[30] K. Takashima, Y. Omori, Y. Yoshimoto, Y. Itoh, Y. Kitamura, and F. Kishino, "Effects of avatar's blinking animation on person impressions," in *Proc. Graphics Interface*, May 28–30, 2008, pp. 169–176.
[31] Y. Yoshikawa, K. Shinozawa, and H. Ishiguro, "Social reflex hypothesis on blinking interaction," in *Proc. 29th Annu. Conf. Cognitive Sci. Soc.*, Nashville, TN, USA, Aug. 1–4, 2007, pp. 725–730.
[32] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends Cognitive Sci.*, vol. 9, no. 4, pp. 188–194, 2005.
[33] *Facetrackingapi version 3.2.6*, Seeing Machines Limited, Tucson, AZ, USA, Aug. 2010.
[34] M. M. Hoque, D. Das, T. Onuki, Y. Kobayashi, and Y. Kuno, "An integrated approach of attention control of target human by nonverbal behaviors of robots in different viewing situations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 1399–1406.
[35] M. M. Hoque, T. Onuki, Y. Kobayashi, and Y. Kuno, "Effect of robot's gaze behaviors for attracting and controlling human attention," *Adv. Robot.*, vol. 27, no. 11, pp. 813–829, 2013.
[36] C. Ware, *Information Visulization: Perception for Design*. San Francisco, CA, USA: Morgan Kaufmann, 2004.
[37] F. Tarrés, (2013, Mar.). "GTAV face database." [Online]. Available: http://gps-tsc.upc.es/ GTAV/ ResearchAreas /UPCFaceDatabase /GTAVFaceDatabase.htm
[38] D. Cornish and D. Dukette, *The Essential 20: Twenty Components of an Excellent Health Care Team*. Pittsburgh, PA, USA: Dorrance Publishing, 2010.
[39] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–467, 1995.
[40] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[41] A. J. Glenstrup and T. Engell-Nielsen, "Eye controlled media: Present and future state," Ph.D. dissertation, Dept. Inf. Psychol., Univ. Copenhagen, København, Denmark, 1995.
[42] D. Kahneman, *Attention and Effort*, A. H. Esser, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[43] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods*. Cambridge, MA, USA: MIT Press, 1999.
[44] Y. Kobayashi, D. Sugimura, Y. Sato, K. Hirasawa, N. Suzuki, H. Kage, and A. Sugimoto, "3d head tracking using the particle filter with cascaded classifiers," in *Proc. Brit. Mach. Vision Conf.*, 2006, pp. 37–46.
[45] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vision Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2005, pp. 886–893.
[47] Y. Kobayashi and Y. Kuno, "People tracking using integrated sensors for human robot interaction," in *Proc. IEEE Int. Conf. Ind. Technol.*, Ann Arbor, MI, USA, Mar. 14–17, 2010, pp. 1597–1602.
[48] W. James, *The Principles of Psychology*. New York, NY, USA: Dover, 1950.
[49] M. M. Hoque, T. Onuki, D. Das, Y. Kobayashi, and Y. Kuno, "Attracting and controlling human attention through robot's behaviors suited to the situation," in *Proc. Int. Conf. Human-Robot Interaction*, Mar. 5–8, 2012, pp. 149–150.
[50] M. M. Hoque, D. Das, T. Onuki, Y. Kobayashi, and Y. Kuno, "Model for controlling a target human's attention in multi-party settings," in *Proc. IEEE RO-MAN*, 2012, pp. 476–483.
[51] P. Downing, C. Dodds, and D. Bray, "Why does the gaze of others direct visual attention," *Vis. Cog.*, vol. 11, no. 1, pp. 71–79, 2004.
[52] T. Onuki, T. Ishinoda, E. Tsuburaya, Y. Miyata, Y. Kobayashi, and Y. Kuno, "Designing robot eyes for communicating gaze," *Interaction Studies*, vol. 14, no. 3, pp. 451–479, 2014.

**Dipankar Das** received the B.Sc. and M.Sc. degrees in computer science and technology from the University of Rajshahi, Rajshahi, Bangladesh, in 1996 and 1997, respectively, and the Ph.D. degree in science and engineering from Saitama University, Saitama, Japan, in 2010.

He is currently a Professor with the Department of Information and Communication Engineering, University of Rajshahi. His research interests include object recognition and human–computer interaction.

**Md. Golam Rashed** is currently working toward the Ph.D. degree in engineering from the Computer Vision Laboratory, Saitama University, Saitama, Japan.

He is a Faculty Member (on study leave) with the Department of Information and Communication Engineering, University of Rajshahi, Rajshahi, Bangladesh. His research interests include robotics and human–robot interaction.

**Yoshinori Kobayashi** received the Ph.D. degree from the Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan, in 2007.

He is currently an Associate Professor with the Graduate School of Science and Engineering, Saitama University, Saitama, Japan. His research interests include computer vision for human sensing and its application to human–robot interaction.

**Yoshinori Kuno** (M'80) received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from the University of Tokyo, Tokyo, Japan, in 1977, 1979, and 1982, respectively.

After working with Toshiba Corporation and Osaka University, since 2000, he has been a Professor in the Department of Information and Computer Sciences, Saitama University, Saitama, Japan. His research interests include computer vision and human–robot interaction.