2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Attracting Attention and Establishing a Communication Channel Based on the Level of Visual Focus of Attention

Dipankar Das[1], Yoshinori Kobayashi[1,2], and Yoshinori Kuno[1]

Abstract— **Recent research in HRI has emphasized the need to design affective interaction systems equipped with social intelligence. A robot's awareness of its social role encompasses the ability to behave in a socially acceptable manner, the ability to communicate appropriately according to the situation, and the ability to detect the feelings of interactive partners, as humans do with one another. In this paper, we propose an intelligent robotic method of attracting a target person's attention in a way congruent to satisfying these social requirements. If the robot needs to initiate communication urgently, such as in the case of reporting an emergency, it does not need to consider the current situation of the person it is addressing. Otherwise, the robot should observe the person to ascertain who or what s/he is looking at (VFOA), and how attentively s/he is doing so (VFOA level). Moreover, the robot must identify an appropriate time at which to attract the target person's attention so as to not interfere with his/her work. We have realized just such a robotic system by developing computer vision methods to detect a target person's VFOA and its level, and testing the system's effectiveness in a series of experiments.**

I. INTRODUCTION

Attracting someone's attention and establishing a communication channel with him/her is a fundamental skill in human social interaction and cognition [1]. This ability plays a critical role in a wide range of social behaviors: it sets the stage for learning [2], develops the capacity for mutual understanding [3], and facilitates communication [4].

In natural human-robot interaction (HRI), robots should behave as humans do with one another, and initiate interaction in the same way. Since robots are perceived as social actors [5], it is essential for them to exhibit social intelligence and awareness, rather than performing merely as computational machines that perform tasks assigned to them by the user [6]. A robot's awareness of its social role encompasses the ability to behave in a socially acceptable manner, the ability to communicate appropriately according to the situation, and the ability to detect the feelings of interactive partners, as humans do with one another. In this paper, we propose an intelligent robotic method of attracting a target person's attention and establishing a communication channel with him/her based on his/her level of visual focus of attention (VFOA). If the robot encounters a situation wherein it must initiate communication urgently, such as in the case of reporting an emergency, it does not need to consider the current situation of the person it is addressing. Otherwise, the robot should observe the person to ascertain who or what s/he is looking at (VFOA), and how attentively s/he is doing so (VFOA level). The robot must then ascertain an appropriate time at which to attract the target person's attention so as not to interfere with their current task. We therefore propose a system in which the robot interacts with the target person intelligently and in a socially acceptable manner, enabling it to interact without disturbing their current VFOA as well as other persons in the environment. In HRI, if an agent (A1) wants to communicate with another agent (A2) without disturbing his/her (A2) current focus or task, then A1 will wait for A2 to either complete their task or lose their current VFOA [7]. Recently, researchers in HRI have been interested in developing models inspired by human cognitive processes, because such models can result in more natural interaction behaviors [8]. Providing robots with skills that foster the impression of a more intelligent and intuitive interaction ensures a high level of satisfaction for interacting humans [9].

The VFOA is an important cue for attracting attention and initiating interaction because, (i) it is an effective cue to understand what the person is doing, and (ii) it is also a good indicator of addressee-hood (i.e., who is looking at whom). For instance, if the target person's VFOA is directed toward the robot, then 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 an opportune moment at which to attract his/her attention and establish a communication channel. In our proposed system, proper timing is determined by detecting a low level of attention (or the loss of attention) of the target person to his/her current task. We use visual cues (such as head pose, head movement, face stability, etc.) and the task context of the target person to recognize the VFOA and its level of intensity. These visual cues, such as head pose, can be used as an approximation for VFOA, as supported by psycho-visual evidence [10] and empirically demonstrated by Stiefelhagen et al. [11]. On the other hand, the task context (what the person is engaged in doing) plays an important role in relating the set of circumstances in which the task takes place [12]. The context is also relevant to derive a precise understanding of the task behavior. Knowledge of the context can then be utilized to make the decision as to when to interrupt the target person. For instance, if the target person is involved in "reading", contextual cues such as "turning a page" or "change in the tilt angle of the head" can be used to determine the momentary loss of the person's current VFOA. Attracting the target person's attention and establishing mutual gaze plays an important role in initiating

[1] The authors are with Graduate School of Science and Engineering, Saitama University, 255 Shimo-Okubo, Sakura, Saitama 338-8570, Japan. {dipankar, yosinori, kuno}@cv.ics.saitama-u.ac.jp

[2] Y. Kobayashi is also with JST, PRESTO, 4-1-8 Honcho Kawaguchi, Saitama 332-0012, Japan.

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

an interaction [13]. If the target person has no VFOA on a particular task, then our robot immediately acquires his/her attention using the approaches we proposed in [14], [33], [15]. However, with the proposed approaches in [14], [33], the target person's field of view is divided into different zones and the robot attracts his/her attention depending on the viewing situation. On the other hand, in [15] the robot waits for the target person to detect eye-contact and then shift his/her attention in the intended goal direction. However, in this current project, our primary accomplishment was to determine the current VFOA, and its level and spatial regions, by using the task and task-related contextual cues of the target person. By ascertaining the degree of attention and spatial regions of the target person's VFOA our robot is able to identify an appropriate time to attract his/her attention and establish a communication channel with him/her.

II. Related Work

A. Initiating Interaction

When initiating communication with one another, humans stop at a certain distance [16] and begin an interaction with greetings [17]. Several studies have addressed greeting behaviors used to initiate human-robot conversation [18], [19]. Some robots have been equipped with the capability to encourage people to initiate interaction by displaying cues related to, for instance, approach direction [5] and path [20], standing position [21], and following behavior [22]. These studies all assumed that the target person was facing the robot and intended to talk with it; however, in actual practice these assumptions may not always hold. In real-life environments, the interactive partner(s) may already be involved in some task. Thus, the robot may need to wait for the person's VFOA to be directed toward it before it can initiate an interaction [15] or pro-actively attract his/her attention at a suitable time. Although a passive attitude as in the former study [15] can work in some situations, many situations require a robot to employ a more active approach [23]. In this paper, we consider a situation in which a human and a robot may not be facing each other, and prepare the robot to initiate an interaction in a socially acceptable way depending on the target person's level of VFOA.

On the topic of recognizing a human's level of VFOA, two main streams of research exist. Some researchers use techniques based on active sensing using infrared light [24]. Although these methods are accurate, they are a little invasive and restrictive. On the other hand, techniques based on computer vision compute gaze, head and body posture from camera images to recognize VFOA [25]. However, the gaze-based VFOA level recognition requires high-resolution images of eyes to be effective. Moreover, the method restricts the mobility of the subject. Recently, some researchers have used head pose as a cue for VFOA estimation [8], [12]. This notion is supported by the fact that turns of the head represent an informative cue in recognizing where the subject is looking. In this paper, we propose a computer vision technique to recognize the VFOA and its level based on head pose, head movement, face stability, and task context.

B. Establishing a Communication Channel

Attracting a person's attraction can produce observable behavioral responses such as eye movements, head movements, or changes in body orientation. If the target person is attracted by the robot's behavior, they will turn toward the robot, making eye contact easy to establish. It may appear that eye contact can be made via establishing gaze crossing (i.e., looking at each other). Several robotic systems are able to establish eye-contact by gaze crossing [26], [27]. Psychological studies show, however, that this gaze crossing action alone may not be sufficient to establish eye contact. Gaze awareness is also necessary for humans to feel that they have made eye contact with others [28]. Therefore, robots need to be able not only to detect human gaze but also to accurately display their gaze awareness in a way that can be correctly interpreted by humans. To solve this problem the robot should be able to display its awareness explicitly through some sort of action (e.g., facial expression, eye blinking, waving) [29]. Yoshikawa et al. [30] demonstrated that eye-blinking by an on-screen agent gave participants a stronger feeling of being looked upon. In this paper, we demonstrate the effectiveness of generating such awareness by the robot raising its head towards the target person in addition to blinking its eyes when making eye contact.

III. Human VFOA Analysis

Humans cannot keep attending to a given task while looking at the same target object for a long time [31]. Consequently, they may sometimes divert their attention. In addition, even while attending to the current task there may be occasions when they avert their eyes from the target. For example, when reading, they may not concentrate on looking at the pages when turning them. Such occasions would be opportune times for the robot to attempt to attract their attention. The span or duration for which humans can maintain their VFOA may depend on the task, while the level of the current VFOA certainly does. To investigate this phenomenon further, we performed observation experiments.

A. Data Collection

We videotaped 18 participants carrying out four different tasks: reading, writing, web browsing, and other (randomly fixing his/her attention on a painting in the room). The participants were all students at Saitama University, consisting of 14 males and 4 females with an average age of 28. Participants were asked to concentrate on their given task and did not receive any further information. The average recorded lengths for each person carrying out the reading, writing, browsing, and other tasks were 9, 9, 8, and 7 minutes, respectively. Figure 1(a) is an example scene showing a participant involved in a web browsing task.

B. Observation

Our observations focused primarily on measuring the span of VFOA on a task and acquiring task-related contextual information. To measure the span of VFOA, we observed the time period that a participant was able to concentrate on a

(a)

Task	Read	Write	Brows.	Other.
Avg.	2.50	3.25	5.25	1.52
Max.	3.25	4.50	6.00	2.50
Min.	1.50	2.00	3.50	1.00

(b)

Fig. 1. (a) A participant carries out a browsing task, and (b) the span of VFOA from the experiment (in minutes).

task without loss of attention. In all cases in this experiment, a loss of attention was detected when the participant changed his/her current VFOA to another direction. Additionally, in the cases of reading and writing we found occasions of loss of attention occurred when participants turned pages and stopped writing, respectively. For the reading, writing, web browsing, and other tasks, we detected respectively 14, 10, 9, and 19 instances of loss of attention. From the duration of these occasions, we then estimated the span of VFOA for each task as shown in Fig. 1(b). For time-on-task measurements, the type of activity used in the test affects the results, as people are generally capable of a longer attention span when they are doing something that they find enjoyable or intrinsically motivating [31].

From videos of the experiment, we also observed how head pose changed at the time of loss of attention for different tasks. The minimum deviation of head pose can thus be used as a cue to detect the loss of attention. Head pose was detected using the Seeing Machine faceAPI because it produces robust and real-time 3D head poses from video images [32]. In most cases, when participants lost their attention during reading or writing, they changed the tilt angle of their head first and then changed the pan angle. The minimum deviation of tilt angle of the head for the reading and writing were 14° and 18°, respectively. However, in the case of browsing, participants normally changed the pan angle of their head to shift their attention to another direction. In this case, the minimum deviation was 17°. When participants were asked to fix their attention on a painting, the loss of attention was detected using either the pan or tilt angle of the head. In such cases, the minimum deviation of pan and tilt angles were 12° and 10°, respectively.

IV. Outline of the Proposed Approach

The basic steps of the proposed approach are illustrated in Fig. 2. The entire system is divided into two modules. In *the initiating interaction module*, the robot recognizes and tracks the target person's VFOA. If they are initially in a face-to-face situation, the robot generates an awareness signal and establishes eye contact with the target person. Otherwise, the robot tries to attract the target person's attention by recognizing his/her VFOA and its level and spatial region. The robot attempts to detect the loss of the target person's current VFOA until T_s (where T_s is the maximum span of sustained VFOA). When it detects the loss of current VFOA (low VFOA) at time t, it generates an attention attraction (AA) signal (weak or strong) depending on the viewing situation of the target person's shifted VFOA (region of VFOA). We represent the viewing situation (the relation between the target person's gaze (face) direction and the robot's position) by where the robot is seen in the target person's field of view, and classify it into four regions as shown in Fig. 3: the robot can be situated in the central field of view (CFOV), in the near peripheral field of view (NPFOV), in the far peripheral field of view (FPFOV), and outside of the field of view (OFOV). A detailed description of these viewing situations and how they are determined is given in a previous study [33]. If the shifted VFOA is detected in CFOV/NPFOV, then the robot generates a head turning action (weak signal). However, if the detected VFOA is in FPFOV, then the robot generates a head shaking action (strong signal). The robot expects that the target person will lose attention on their current focus within T_s. However, if within this period the robot does not detect the loss of current focus, it then resorts to generating a weak signal. If the weak signal fails, the robot uses a strong signal to attract the target person's attention.

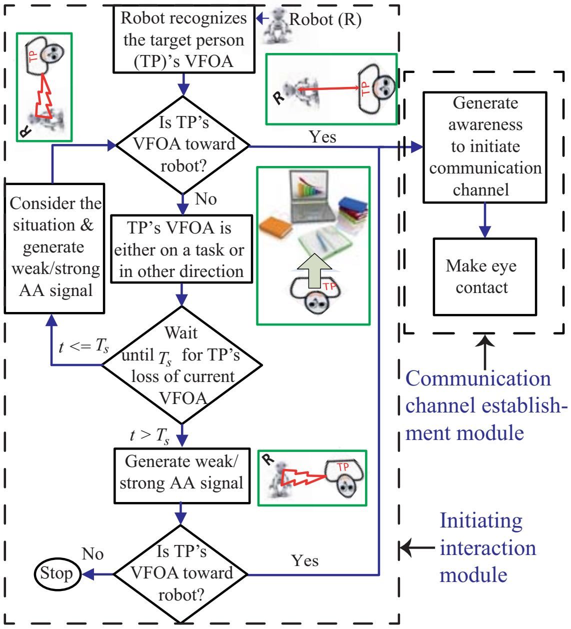

Robot recognizes the target person (TP)'s VFOA
Robot (R)
TP
R
Is TP's VFOA toward robot?
Yes
No
Generate awareness to initiate communication channel
Make eye contact
Consider the situation & generate weak/strong AA signal
TP's VFOA is either on a task or in other direction
Wait until T_s for TP's loss of current VFOA
$t <= T_s$
$t > T_s$
Generate weak/ strong AA signal
Is TP's VFOA toward robot?
Stop
Communication channel establishment module
Initiating interaction module

Fig. 2. Basic steps of the proposed approach.

When the robot succeeds in attracting the target person's attention, *the communication channel establishment module* tries to establish a communication channel with him/her. For this purpose, the robot determines the level of shifted attention. Based on the level of shifted attention, the robot generates an awareness signal toward the target person to indicate that it wants to communicate with him/her. Finally, the robot makes eye contact through eye blinking to establish a communication channel with him/her.

V. Recognition of VFOA and Its Level

We are interested in detecting two kinds of attention: *sustained attention* and *focused or shifted attention*. Focused

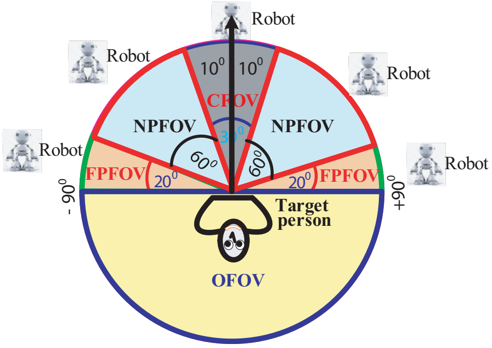


Fig. 3. Representation of the viewing situation.

or shifted attention is a short-term response to a stimulus or any other unexpected occurrence. The span of this attention is very brief [31] 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, on the other hand, is a level of attention that produces consistent results at a task over time. It has already been shown that the duration of sustained attention depends on the task. We use the following cues to recognize VFOA and estimate its level.

A. Visual Cues

Head Pose. We use the Seeing Machines's faceAPI to detect and track the head pose, h_p of the target person. We classify detected head poses into three angular regions: h_p^{cfov}, h_p^{npfov}, and h_p^{fpfov} if they are detected in CFOV, NPFOV, and FPFOV areas, respectively. The pan and tilt angles of the poses are denoted by h_p^p, and h_p^t, respectively.

Head Movement. To detect the head movement, h_m we use the optical flow feature [34]. We generate a rectangular window circumscribing pixels with large flow values. If the total flow values in the window exceeds a threshold, we consider a head movement ($h_m = 1$) cue to have been detected. Thus, h_m indicates a patch of the image where a moving head is detected.

Overlapping Face Window. If a face is detected and overlaps with the most recent head movement widow, h_m, by more than 50%, we consider an overlapping face window, o_f, to have been detected ($o_f = 1$). This means that the target person is turning his/her face toward the robot. Faces are detected using the Viola-Jones AddaBoost face detector [35].

Face Stability. In measuring the visual focus of attention, the stability of face direction, N_{f_s}, is an important cue. N_{f_s} is represented by the number of frames with the same face direction. If the target person shifts his/her attention towards the robot, the system detects the focused attention by detecting the overlapping face window, o_f. After detecting o_f, the robot determines the stability of the face direction (given that a face is detected and its pose is within the CFOV area in several subsequent frames).

B. 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 contextual cues such as "downward-turned head" indicate that his/her attention is toward the book. However, cues such as "turning a page", or "lifting the head upward" indicate that the person has momentarily lost his/her attention.

Task Recognition. From a given video sequence we extract the histogram of orientation gradient (HOG) feature [36] from the whole image for each frame in the sequence. The HOG features are combined for 10 consecutive frames to build a HOG feature pattern, HOG_P. Thus,

$$HOG_P = F_0 + \sum_{i=1}^{9} |F_{i-1} - F_i| \tag{1}$$

where F_0 and F_i are the HOG features of the first and ith frames, respectively. The first frame captures the human appearance features involved in a task, and the rest of the HOG feature frames indicate any changes in the behavioral patterns while engaged in the task. Here, each bin in the histogram represents the number of edges that have the orientation within a given angular range. The angular range is set to 20 degrees and we use an unsigned gradient. Therefore, the bin size $= 180/20 = 9$. With this bin size, we then create a HOG_P feature vector of size 90. This feature vector is used to learn a multi-class support vector machine (SVM) [37]. For detection, we use the SVM classifier in the recognition mode. To evaluate the performance of this system, we used the dataset as described in section III. The dataset was divided into training and test videos. The classifier was learned using 8270 samples (each sample consisting of 10 successive video frames) from training videos for the four different tasks. We used 10,269 test samples from test videos for the four different tasks. Among these, 9612 samples were correctly detected with 93.6% accuracy.

Contextual Cues. After recognizing the task or VFOA of the target person, the system uses the contextual cues related to the task in question to recognize the level of attention. The task-related VFOA span, (T_s), is used to determine how long the robot should wait before disturbing the target person. The robot starts measuring T_s after detecting the task of the target person. We also defined some task-specific cues to assist the robot with determining the level of attention. In a reading situation, we use the page turning, P_t, deviation in tilt angle, $d_{h_p^t}$, and h_m cues to measure the level of attention. In the case of writing, attention level is estimated using halts in writing, W_s, $d_{h_p^t}$, and h_m cues. For web browsing, we use deviation in pan angle, $d_{h_p^p}$, and h_m to determine the level of attention. Finally, for other tasks, $d_{h_p^t}$, $d_{h_p^p}$, and h_m cues are used for estimating attention level. Head movement, page turns, and writing stop 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. A detailed description of the body tracking system (BTS) employed here is given in [38]. Threshold values of 100 and 75 are set to detect the page turns and head movement behavior, respectively, while if the resultant magnitude is approximately zero (in the experiment we set it at 2) for 10 consecutive frames, then a stop in writing is detected. We consider 10 consecutive frames an appropriate number because a person may stop his/her writing motion for a brief moment without actually shifting his/her attention. The threshold values are set manually on a trial-and-error basis. We evaluated the system by using the dataset described in section III to detect a loss of attention (low VFOA). Among the 52 cases of loss of attention in the four different tasks in our experiment, our system detected 44 times correctly, resulting in a detection rate of 84.6%. During these 52 sample observations, false positives occurred 32 times.

VI. Level of Sustained VFOA

The level of VFOA is classified into two categories (low or high) based on the contextual cues of the VFOA in question. When the level of attention drops, the system assumes that a loss of VFOA has been detected. For different tasks, the attention level is detected using the following equations.

$$SA_{L,read} \leftarrow h_m \vee P_t \vee (d_{h_p^t} \geq 14^\circ) \tag{2}$$

$$SA_{L,write} \leftarrow h_m \vee W_s \vee (d_{h_p^t} \geq 18^\circ) \tag{3}$$

$$SA_{L,browse} \leftarrow h_m \vee (d_{h_p^p} \geq 17^\circ) \tag{4}$$

$$SA_{L,other} \leftarrow h_m \vee (d_{h_p^p} \geq 12^\circ) \vee (d_{h_p^t} \geq 10^\circ) \tag{5}$$

If any of Eqs. (2)−(5) are true and their stability is greater than or equal to 3 frames, then the level of attention is deemed low for the corresponding task. Otherwise, the attention level is deemed high, meaning that currently attention remains focused on the task.

VII. Detection of Focused/Shifted Attention

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

Shifted Attention from Sustained VFOA. To initiate a polite social interaction, the robot should attract the target person's attention when s/he loses his/her current sustained VFOA as shown in Fig. 4(b). Thus, the robot first detects the loss of sustained VFOA of the target person using one of Eqs. (2)−(5) when its value is low. After detecting the loss of attention, the robot detects the shifted VFOA (Fig. 4(b)) of the target person. Depending on the environmental factors (e.g. visual stimuli, sound etc.) and the target person's mental focus, the sustained VFOA can be shifted into one of the four regions outlined above (Fig. 3): CFOV, NPFOV, FPFOV, OFOV. The shifted VFOA region is detected using the pan angle of head pose, h_p^p.

Focused/Shifted Attention toward the Robot. The detection of focused/shifted attention directed 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 performs an attention attraction signal directed toward the target person, and awaits his/her attention to be directed toward it in return. When the target person shifts or turns his/her attention toward the robot, the robot must detect this change. To make successful eye contact, the robot classifies the level of focused/shifted attention into three categories: Low, Medium, and High. The robot then performs an attention attraction signal directed toward the target person, and analyzes the input video images on a frame-by-frame basis to detect whether the target person is responding by moving. It is assumed that if the target person is turning to look at the robot from his/her current focus of attention, then some contiguous h_m windows will be detected surrounding their head. Depending on the detected visual cues described in section V-A, the level of focused/shifted VFOA is classified according to the following equations.

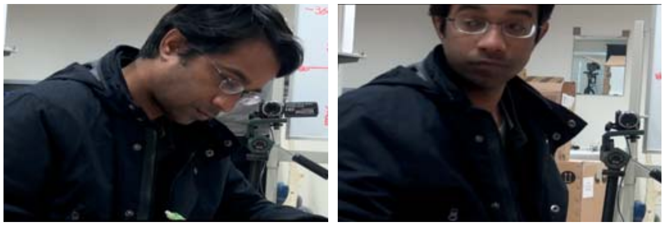


Fig. 4. Target person's attention: (a) target person is focused on writing, (b) target person loses his attention and shifts his focus into FPFOV (seen from the robot's vantage point).

When none of the visual cues are detected except for head movement as in Eq. (6), the robot assumes that the focused/shifted attention level is low, FA_L:

$$FA_L \leftarrow \big((N_{h_m} \geq 1) \wedge (o_f = 0) \wedge (N_{f_s} \leq 1) \wedge (h_p^p \neq CFOV)\big) \tag{6}$$

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 the 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 a head movement is detected and the system identifies an overlapping window of the face within the contiguous head movement area, then it considers the level of attention to be medium, FA_M:

$$FA_M \leftarrow \big((N_{h_m} \geq 5) \wedge (o_f = 1) \wedge (N_{f_s} \leq 1) \wedge (h_p^p = CFOV/NPFOV)\big) \tag{7}$$

Finally, when all of the visual cues are successfully detected and stable enough, then the robot considers there to be a high level of attention, FA_H:

$$FA_H \leftarrow \big((N_{h_m} \geq 5) \wedge (o_f = 1) \wedge (N_{f_s} \geq 5) \wedge (h_p^p = CFOV)\big) \tag{8}$$

When all the conditions on the right-hand side of Eqs.(6)−(8) are satisfied, the corresponding level of attention is detected. The detected level of attention is used in the subsequent awareness generation, and in establishing successful eye contact.

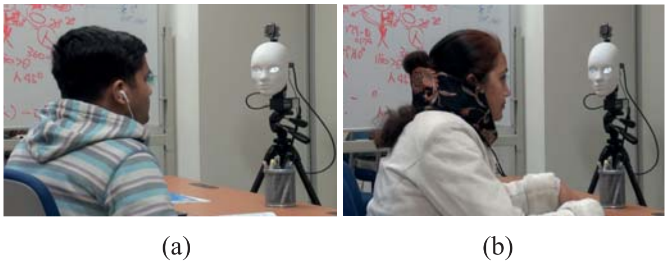

(a) (b)

Fig. 5. (a) a TP's VFOA is shifted from "reading" to the CFOV area, and (b) a TP's VFOA is shifted from "reading" to the NPFOV area.

VIII. Initiating Interaction Based on VFOA

In polite social interaction, humans usually first raise or turn their head toward the person with whom they would like to communicate when s/he loses her/his current VFOA. However, if the target person's degree of focus on a task is high, humans attempt stronger actions (e.g. waving a hand, coming closer to the person and turning the head, or even using their voice) to attract his/her attention. In our system, the robot detects the target person's level of sustained VFOA and the region of shifted VFOA in order to choose an appropriate control signal. From a survey of psychology, HRI literature, and our preliminary experiment, we chose a head-turning action (turning to look at the person) as the weak signal when the sustained VFOA attention level is low and the shifted VFOA is in either the CFOV or NPFOV areas (Fig. 5). We decided to employ a head-shaking action when the sustained VFOA attention level is low and the shifted VFOA is in the FPFOV area as shown in Fig. 4(b). 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 employ a head-shaking action as a strong attention attraction signal because abrupt object motion has been demonstrated to draw people's attention [39]. A detailed description of the cues is given in a previous study [14]. The visual stimuli created by the robot's non-verbal behaviors cannot affect a person if s/he is in a position where s/he cannot see the robot's action, and thus we do not consider situations where the shifted VFOA is in the OFOV area.

IX. Establishing a Communication Channel

To establish a communication channel, the robot needs to make the person clearly notice that it is looking at her/him alone. To solve this problem, the robot should be able to display its awareness explicitly through behavioral cues (for example, facial expressions, eye blinking, or nodding). In our current system, we enabled the robot to display this awareness through a "head-focusing action" [15] and eye blinking [14], since these are recognized as important cues for forming a person's impressions [29].

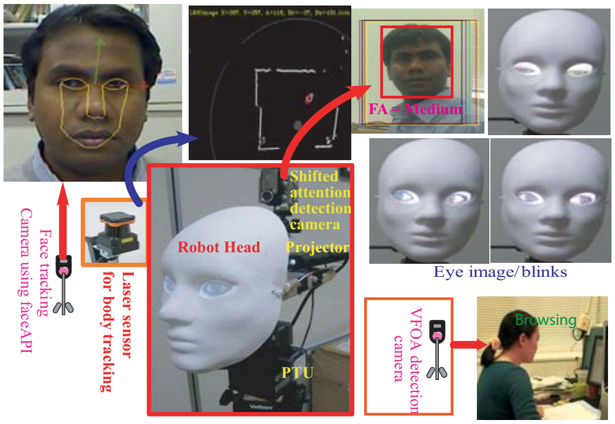


Fig. 6. Experimental robotic platform.

Head-Focusing Action. After generating a head-turning or head-shaking action, the robot observes the target person as long as the attention level toward it continues to be low (i.e., $FA_L = 1$). When the attention level becomes medium (i.e., $FA_M = 1$), meaning that the robot detects any head movement and a face turned toward it, the robot generates an awareness signal by raising its face toward the target person. In the current implementation, the height of the robot is lower than that of humans. Thus, raising its head in this case means that it turns its head to precisely face the person and show that its attention is focused on that person.

Eye Blinks. If the robot successfully attracts the target person's attention, or s/he notices the robot's action, s/he will direct her/his gaze at the robot. The robot recognizes her/his face while s/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 3 times (1 blink/sec) to establish a communication channel.

X. Experiment

We conducted an experiment to verify that the method is useful in attracting the target human's attention from his/her currently direction of attention and in establishing a communication channel through eye contact with him/her.

A. Participants

In this experiment, we used 24-non-paid participants (19 males and 5 females). The participants were all students at Saitama University with an average age of 30.7. They were divided into 4 groups and asked to do the task assigned to each group: reading (12 participants), writing (4 participants), web browsing (4 participants), and other (4 participants; participant was asked to randomly fix his/her attention on any location in the environment except OFOV).

B. Platform

Figure 6 shows an overview of our robotic platform. We have developed a robotic head for HRI experiments mounted on a pan-tilt unit (Directed Perception Inc., PTU-D46). We use two USB cameras (Logicool Inc.): one for face tracking using faceAPI and the other for face detection. A laser range sensor (URG-04LX) is used for body tracking. An LED projector (3M pocket projector, MPro150) is utilized to project CG-generated eyes on the mask (robot head).

C. Experimental Methodology

The participants were to evaluate various behaviors employed by the robot to attract their attention when they were attending to their tasks. They were asked to concentrate on their tasks, and each participant only experienced one task. In order to evaluate the effectiveness of the proposed method, we compared it with another one, so that each participant experienced two types of robot behavior. We asked each participant to put on headphones with music to avoid being distracted by the sound produced by the pan-tilt movement of the robot. We placed two video cameras in appropriate positions to capture all interactions. Figure 7(a) shows the experimental environment.

Intelligent Attention Control Robot (IACR). This system incorporates our proposed method, as described in section IV. The robot determines the level of attention the target person is giving to their current task and considers their situation. The robot attracts attention when the target person loses his/her attention to the task. The robot performs a head-turning action when the target person's attention is shifted to either the CFOV or NPFOV regions. However, if the attention is shifted to the FPFOV region, then the robot uses a head-shaking action.

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 his/her attention immediately. To do so, the robot uses two types of attention attraction signals, beginning with head-turning and then following up with head-shaking if the first action fails to attract the target person's attention.

D. Hypothesis

We expected that the following hypotheses (H1-H4) would be verified by the experiment. **H1**: The proposed method (IACR) outperforms the other method (SACR) in attracting the participants' attention. **H2**: The proposed method is deemed more socially acceptable than the other method in attracting the participant's attention. **H3**: The proposed method disturbs the participants less than the other method in attracting attention. **H4**: The proposed method outperforms the other method in establishing a communication channel.

E. Measures

We measured the following two items in the experiment:

Impression of the robots. After they had experienced two interactions, we asked participants to fill out a questionnaire. The measurement was a simple rating on a Likert scale of 1 to 7. The questionnaire consisted of the following four items: **(Q.1)** Did you feel that the robot attracted your attention? **(Q.2)** Did you feel that the robot's interruption was an acceptable way to attract your attention? **(Q.3)** Did you feel you were disturbed by the robot's interruption when it was attracting your attention?, and **(Q.4)** Did you make eye contact with the robot?

Success ratio. From the videos, we counted the number of times that the target participant looked at the robot after it performed its attention attraction actions. We also counted the number of times that the robot was successful in detecting the participants responding by looking. These numbers can be used to evaluate the success ratio of the proposed method.

TABLE I
QUESTIONNAIRE RESULTS WHERE 7 IS "STRONGLY AGREE"

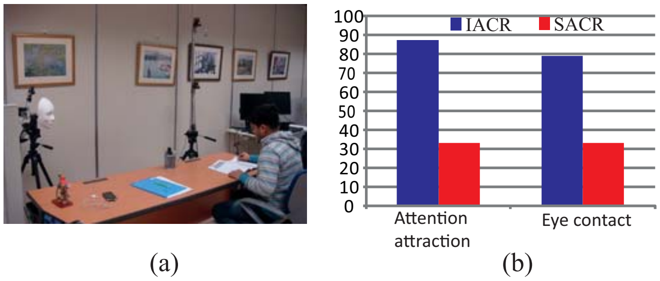
	IACR		SACR	
	Mean	SD	Mean	SD
Q1	5.92	0.43	3.29	1.25
Q2	6.00	0.52	3.21	1.74
Q3	2.38	0.59	4.42	1.73
Q4	6.21	0.52	3.50	2.00

F. Results

The experiment was performed in a within-participant design, and the order of all experimental trials was counterbalanced.

Impression. The experimental results of the questionnaire, measured by the mean (M), and standard deviation (SD) for each type of robotic behavior, are shown in Table I. For Q1, the differences between the conditions were statistically significant $(F(1,23) = 83.37, p < 0.001, \eta^2 = 0.68)$. The result reveals that IACR is more effective at attracting attention than is SACR, clearly verifying the first hypothesis. Concerning Q2, significant differences were found $(F(1,23) = 69.43, p < 0.001, \eta^2 = 0.64)$ between the two behaviors. Consequently, the robot's choice of interruption time to attract the target participant's attention was proven to be appropriate and acceptable, verifying the second hypothesis. For Q3, significant differences were again measured between the two robotic behaviors $(F(1,23) = 59.06, p < 0.001, \eta^2 = 0.48)$. Thus, the participants felt considerably less disturbed when the robot took their attention into account in determining when to initiate interaction. This supports the third hypothesis. Finally, for Q4 as well the differences between the two conditions were statistically significant $(F(1,23) = 55.50, p < 0.001, \eta^2 = 0.60)$, indicating that the proposed method is effective for establishing a communication channel and thus supporting the fourth hypothesis.

Success Ratio. From the videos recorded during the experiment, we measured the performance of our system attracting the attention of the target participants and making eye contact. Each participant experienced two methods, meaning there were a total of $2 \times 24 = 48$ sample observations of interactions with the robot. Figure 7(b) shows the success rate of our system for the two types of robotic behaviors. The result indicates that the proposed robot, IACR, attracted the attention of the target participant 21 times out of 24 trials (a success rate of 87.5%), making it substantially more successful than SACR, which only attracted the participant's attention 8 times out of 24 trials (a rate of 33.33%). We also performed a t-test to ascertain the difference in performance between IACR and SACR. A significant difference between the performances of the two methods were found $(t(1) = 12.06, p < 0.05)$. The experimental results thus clearly reveal that our proposed method (IACR) is far more effective at attracting a target person's attention.



(a) (b)

Fig. 7. (a) experimental scene, and (b) success rate of the system.

XI. Conclusion

In this study, our main focus was to develop a robot system able to interact with a target person in a socially acceptable manner. We considered a general situation where the target person and the robot may not initially be facing each other. The target human may be engaged in different tasks (such as, reading, writing, etc.). Our robot observes the target person's behavior and finds a suitable time at which to attract their attention, depending on the current visual focus of attention and its level. We proposed a method of detecting the current VFOA by using HOG pattern features and an SVM classifier. Different visual cues and task-related contextual cues are used by the system to determine the level of attention. We then assessed our method by comparing it to another type of robot behavior that also sought to attract a target person's attention but did not take into account their VOFA in doing so. The results indicate that our proposed robot is highly effective at capturing and holding a target person's attention.

Acknowledgments

This work was supported by JST, CREST.

References

[1] A. Pentland, *Honest Signals: How They Shape Our World*. MIT Press, 2010.

[2] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: a field trial," *Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 61–84, 2004.

[3] S. E. Brennan, *Seeking and Providing Evidence for Mutual Understanding*. Stanford University Press, 1990.

[4] W. Dong, B. Lepri, T. Kim, F. Pianesi, and A. Pentland, "Modeling conversational dynamics and performance in a social dilemma task," in *ISCCSP*, May 2-4 2012, pp. 1–4.

[5] K. Dautenhahn et al., "How may i serve you?: a robot companion approaching a seated person in a helping context," in *HRI*, M. A. Goodrich et al., Eds. ACM, 2006, pp. 172–179.

[6] E. Andre et al., "The automated design of believable dialogues for animated presentation teams," in *Embodied Conversational Agents*. MIT Press, 2000, pp. 220–255.

[7] E. Arroyo and T. Selker, "Attention and intention goals can mediate disruption in human-computer interaction," in *IFIP TC 13 Int. Con. on Human-Computer Interaction - Volume Part II*, ser. INTERACT'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 454–470.

[8] Z. Yucel, A. A. Salah, C. Mericli et al., "Joint attention by gaze interpolation and saliency," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. In press, 2013.

[9] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: an ethnographic study," in *HRI*, A. Billard, P. H. K. Jr., J. A. Adams, and J. G. Trafton, Eds. ACM, 2011, pp. 37–44.

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

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

[12] 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, 2011.

[13] E. Goffman, *Behavior in Public Places*. Free Press, 1966.

[14] M. M. Hoque, T. Onuki, Y. Kobayashi, and Y. Kuno, "Effect of robot's gaze behaviors for attracting and controlling human attention," *Advanced Robotics*, vol. 27, no. 11, pp. 813–829, 2013.

[15] D. Das, M. M. Hoque, T. Onuki, Y. Kobayashi, and Y. Kuno, "Vision-based attention control system for socially interactive robots," in *RO-MAN*, Paris, France, September 9-13 2012, pp. 496–502.

[16] E. T. Hall, *The Hidden Dimension: Man's Use of Space in Public and Private*. Bodley Head, 1966.

[17] A. Kendon, *Features of the Structural Analysis of Human Communicational Behavior, in Aspects of Nonverbal Communication*, W. V. R. Engel, Ed. Swets and Zeitlinger, 1980.

[18] K. Hayashi, D. Sakamoto, T. Kanda et al., "Humanoid robots as a passive-social medium: a field experiment at a train station," in *HRI*, C. Breazeal et al., Eds. ACM, 2007, pp. 137–144.

[19] R. Gockley, A. Bruce et al., "Designing robots for long-term social interaction," in *IROS*. IEEE, 2005, pp. 1338–1343.

[20] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro et al., "How to approach humans?: strategies for social robots to initiate interaction," in *HRI*, M. Scheutz et al., Eds. ACM, 2009, pp. 109–116.

[21] F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "A model of proximity control for information-presenting robots," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 187–195, 2010.

[22] R. Gockley et al., "Natural person-following behavior for social robots," in *HRI*, C. Breazeal et al., Eds. ACM, 2007, pp. 17–24.

[23] A. Cesta, G. Cortellessa, V. Giuliani, F. Pecora, R. Rasconi, M. Scopelliti, and L. Tiberio, "Psychological implications of domestic assistive technology for the elderly," *Psy. J.*, vol. 5, no. 3, pp. 229–252, 2007.

[24] J. S. Babcock and J. B. Pelz, "Building a lightweight eyetracking headgear," in *Sym. on Eye tracking research & applications*, ser. ETRA '04. New York, NY, USA: ACM, 2004, pp. 109–114.

[25] R. Vertegaal, R. Slagter et al., "Eye gaze patterns in conversations: there is more the conversational agents than meets the eyes," in *CHI*, J. A. Jacko and A. Sears, Eds. ACM, 2001, pp. 301–308.

[26] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu, "Development and evaluation of an interactive humanoid robot 'robovie'," in *ICRA*, Washington DC, USA, 2004, pp. 1848 – 1855.

[27] B. Mutlu, J. K. Hodgins, J. Forlizzi, and T. Shiwa, "A stroytelling robot: Modeling and evaluation of human-like gaze behavior," in *IEEE-RAS HUMANOIDS*, Genova, USA, 2006, pp. 518–523.

[28] M. V. Cranch, *The Role of Orienting Behavior in Human Interaction*, A. H. Esser, Ed. Plenum Press, 1971.

[29] K. Takashima et al., "Effects of avatar's blinking animation on person impressions," in *Graphics Interface*, ser. ACM Int. Con. Proc., C. Shaw et al., Eds. Ontario, Canada: ACM Press, 2008, pp. 169–176.

[30] Y. Yoshikawa, K. Shinozawa, and H. Ishiguro, "Social reflex hypothesis on blinking interaction," in *Annual Con. of the Cognitive Science Society*, Nashville, Tennessee, USA, Aug. 1-4 2007, pp. 725–730.

[31] H. A. Ruff and K. R. Lawson, "Development of sustained, focused attention in young children during free play," *Developmental Psychology*, vol. 26, no. 1, pp. 85–93, January 1990.

[32] "Seeing machines faceapi," March 2013. [Online]. Available: http://www.seeingmachines.com/product/faceapi/

[33] M. M. Hoque, D. Das et al., "An integrated approach of attention control of target human by nonverbal behaviors of robots in different viewing situations," in *IROS*. IEEE, 2012, pp. 1399–1406.

[34] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–467, 1995.

[35] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR (1)*. IEEE Computer Society, 2005, pp. 886–893.

[37] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[38] Y. Kobayashi and Y. Kuno, "People tracking using integrated sensors for human robot interaction," in *IEEE Int. Con. on Industrial Technology*, Michigan, 2010, pp. 1597–1602.

[39] W. James, *The Principles of Psychology*. Dover, New York, 1950.