An Attention-Directed Robot for Social Telepresence

Rui Yan Keng Peng Tee Yuanwei Chua Zhiyong Huang Haizhou Li

Institute for Infocomm Research
Agency for Science, Technology and Research, Singapore (A*STAR)
{*ryan,kptee,ychua,zyhuang,hli*}*@i2r.a-star.edu.sg*

Abstract: In this paper, we present an attention-directed robot with audiovisual attention control system for telepresence applications. The robot is able to automatically direct attention to the person of interest Attention direction is based on an integration of 3D speech source localization and visual face tracking. To study the effect of automatic attention direction on the telepresence experience, we conducted a user study for a video-conferencing session between two groups of participants in two separate rooms. The users' responses show that, in the presence of automatic attention direction, the feeling of presence (i.e. that a remote person is in the same room) increases, the ease of show and tell increases, and the flow of the video-conferencing communications is smoother. These results suggest that the attention-directed robot can enhance social telepresence.

1 Introduction

Telepresence technologies enable people to feel that they are present in the same locality even though they are communicating from different locations. Some problems in telepresence have been addressed in existing research work including cost, network bandwidth, resolution, and the lack of eye-contact [1]. Besides these problems, the fixed screen and camera of many existing solutions is also a major limitation. When one person moves, the display and camera do not follow the movement and this may lead to disengagement in the telepresence experience, especially when a person unintentionally leaves the camera view in the middle of a conversation. One way to address the field-of-view problem is to move the camera using actuators. Commercially available solutions include Motrr's Galileo, a pan-tilt unit for smart phones, and Double Robotics' mobile platform for smart tablets. However, these motorized platforms lack intelligent or interactive behavior and can only be manually controlled by a remote user.

For a robot to show intelligent and interactive behavior in the presence of humans, it is important that both verbal and non-verbal behaviors of humans, such as facial expressions and body language that accompany speech, be detected. This can be achieved by audiovisual integration based on scene understanding and position information, as shown in several works that aim to make human-robot dialog more natural and flexible [2, 3, 4]. By adding short term memory, people can be tracked even if they disappear from the camera view for a while [5, 6, 7]. Attention control has also been developed for robots to enhance interactions with humans, using synchrony-based neural network architecture [8] and spatiotemporal perceptual learning mechanism [9]. Non-verbal elements such as hand and head gestures, postural mirroring, interpersonal distance and eye contact have been studied on a mobile and articulated robot designed for embodied telepresence [10].

Motivated by the above mentioned works, particularly the embodied telepresence approach, we have developed a telepresence robot comprising a microphone array and a pan-tilt unit on which a camera and a tablet display are mounted. The tablet display moves together with the camera to face a speaker, much like one turns his/her head to face another person during a conversation. Attention direction to the person of interest in the meeting is achieved by a fusion of visual face tracking using the camera, and 3D sound localization using the microphone array.

We conduct a small scale preliminary user study that attempts to gain insight into the possibility that such a telepresence robot can enhance the telepresence effect by automatically directing attention to the person of interest in the meeting. The experiment scenario is that of a video-conferencing session between 2 groups of participants in 2 separate rooms, each with a robot. During the session, the participants performed 3 interactive tasks, including self-introduction, open-ended discussion on a specific topic, as well as show and tell of a self-designed lego assembly. After going through a similar session with an immobilized robot for comparison purpose, they filled a questionnaire that evaluates the experience in terms of an increase/decrease in the feeling of presence, the ease of show and tell, and the smoothness of communication flow.

The remainder of this paper is organized as follows. Section 2 provides a description of the attention-directed telepresence robot, including the methods used for face tracking and speech source localization. Following that, the study task, hypotheses, setup, and procedures of the experiments are described in Section 3. Then, in Section 4, the results of the user study are analyzed with respect to the experimental hypotheses, before conclusions are drawn on the effect of our attention-directed robot on the telepresence experience.

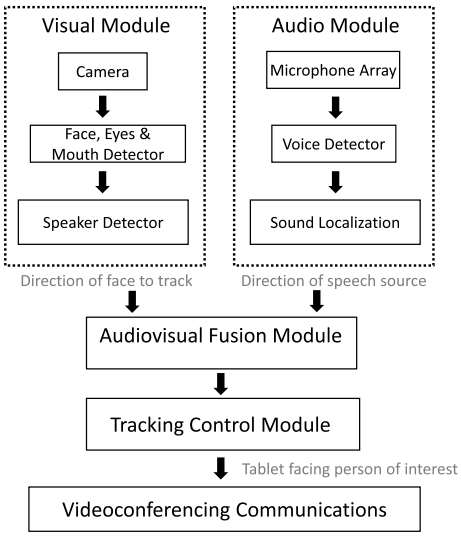

Figure 1: Schematic overview of attention-directed control.

2 Attention-Directed Telepresence Robot

Our attention-directed robot is able to localize a speech source and this is useful for directing attention to a person who is speaking but is out of view. Also, the system is able to continuously track a moving person, thus giving the speaker the freedom to stand up, pace around, or walk to a white board to illustrate a point.

A schematic overview of the attention-directed control is shown in Figure 1. The system consists of the input layer, two hidden layers and the output layer. Camera and microphone array are used in the input layer. The first hidden layer is composed of visual and audio modules. The visual module consists of a visual computing sub-system that detects and tracks human's face, eyes and mouth. In the audio module, a sound source localization sub-system is proposed to detect, localize and track sound sources in the environment. In the second hidden layer, an audiovisual fusion module is used to integrate the output from the first hidden layer.

2.1 Robot Description

The robotic system consists of a pan-tilt unit, computing unit, microphone array, camera, audio interface, motor controller board, and tablet. The pan axis is perpendicular to the table while the tilt axis lies on a plane parallel to the table. The tablet provides video steaming of users at the far end through in-built video-conferencing communications (e.g. Facetime). It is mounted on the pan-tilt unit, and a webcam is, in turn, mounted at the top of the tablet, as shown in Figure 2. Since the tablet moves together with the camera to face a speaker, it provides an embodiment of the head, and emulates how one turns his/her head to face another person during a conversation. The microphone array comprises 8 microphones placed on a curved surface that provides a non-coplanar configuration necessary for 3D sound localization.

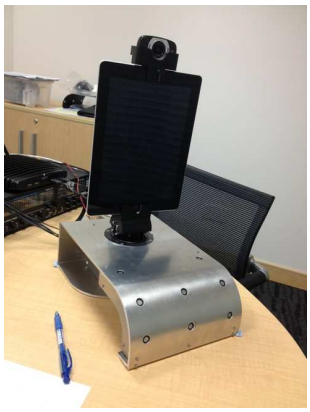

Figure 2: The prototype of the pan-tilt robotic base with a smart tablet.

2.2 Sound Localization

Our speech source localization method is based on the algorithm in [11] , which uses a Voice Activity Detector (VAD) to discriminate human voice from irrelevant sounds, followed by a combination of Time Delay of Arrival (TDOA) and Steered Beamformer methods to determine the direction of the speech source.

A periodic VAD algorithm is used, motivated by the periodicity nature of speech. The ratio of the peak portions and the valley portions of speech spectrum is used as a key feature for voice detection since it is generally robust to noise effects. The voiced sounds often contain pitches of high power components which can be searched out even with noise, and the valleys are located between two adjacent pitches.

The TDOA method is a very popular method for the sound localization. It consists of two key steps: (i) Obtain the TDOA values between each pair of the microphones in the array; (ii) Compute the direction of sound source based on the TDOA values. Besides TDOA, a steered beamformer approach is also used to improve the robustness of sound localization. The direction of maximum output energy corresponds to the sound source direction. The basic idea is that all the microphone signals are used by a beamformer (spatial filter) that is steered in all possible directions in a spherical space to look for the maximal output energy.

2.3 Visual Tracking

In the scenario of telepresence between two parties, each comprising multiple users, the attention arbitration problem needs to be solved. This can be done by the following rules: (i) Visually detect which of the users in view is speaking, and then focus attention on the speaker. (ii) If no speaker is detected, then focus attention on the mean face position of the users in view.

The first rule makes sense because, in a telepresence, it is generally true that only one speaker is speaking at any one time. Hence, there is no compelling need to select attention between multiple speakers. The second rule ensures that, if no one is speaking, all users currently

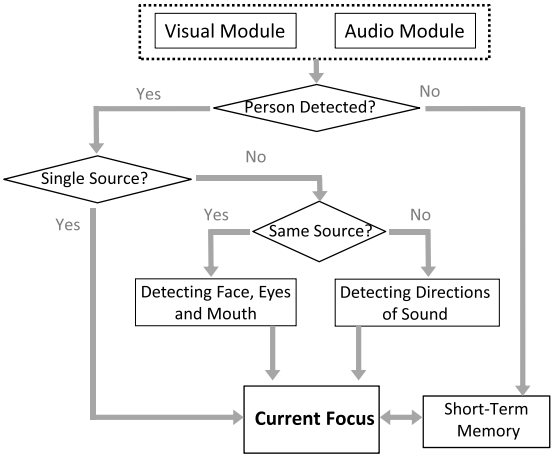

Figure 3: Architecture of multi-modal fusion attention system.

in view are kept in view as much as possible. If one or more of these users walks away from the cameras view, the mean of face position of the remaining is automatically re-calculated and attention re-directed. The visual speaking detection is based on the algorithm of [12], which takes into account visual features from the face, eyes, and mouth. Specifically, we use OpenCV Haar-Based Cascade Classifier to find the face and eyes of each user in view, which allow us to estimate the mouth position. Then, cropped images of the mouth area are obtained and passed into a speaking detection SVM classifier, which determines if the said user is speaking.

2.4 Audiovisual Fusion

In this part, the first important problem is how to decide the attention from the multi-modal results which include the face detection from visual computing and the sound source detection. In order to achieve the accurate and robust performance, the users are required to initialize the system before starting to use the platform for the teleconference. We require the robot to turn around and memorize the accurate positions of all persons in the meeting room by combining the results of face detection and sound localization. These positions will be put into a short-term memory (STM). We will update any position when any user changes the position. The architecture of multi-modal fusion attention system is shown in Fig.3.

Now we will show the detailed steps of how the system finds the attention focus. In the proposed approach, it is required to memorize the attention position P_a of current speaker in the STM. P_a will be updated after the new attention is detected. The main steps are:

1. The first step is to check the attention detection from the visual and audio modules. If there is not a new attention detection, the current attention will be kept the same as the previous one in the STM. With the help of STM, it can reduce search space and confirm the precise locations. Otherwise, if a new attention is detected it will be passed to the next step .

2. If the new source comes from a single module, then the attention can be directly determined. Otherwise, we check if the condition, that sound detection subsystem provides the same azimuth angle region as that of the face detection of visual computing subsystem, is true.

3. Audio module takes priority when the condition is not satisfied. In this case, the new attention will be inferred from both the sound localization and user's positions in the STM.

4. Visual module takes priority when the condition is satisfied.

5. The STM is updated with the current attention focus.

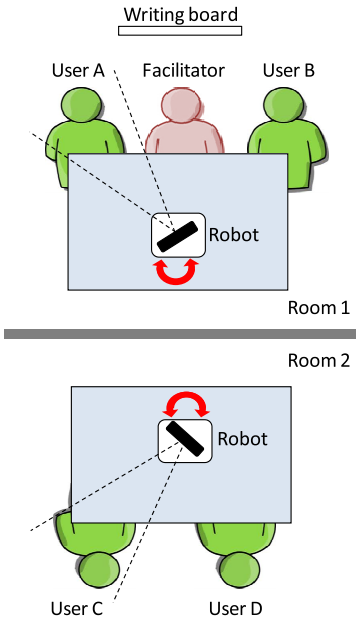

Figure 4: Conceptual design of telepresence experiment.

3 Experiment

3.1 Study Task

The experiment scenario is that of a video-conferencing session between 2 groups of participants in 2 separate rooms. The conceptual design is shown in Figure 4. Each session consisted of 4 users and 1 facilitator. One of the rooms contained the facilitator and 2 users, while the other contained 2 users only. The facilitator is an experimenter whose role was to lead the discussions, ensure that the participants had equal chances to speak, and summarize the discussion points.

During the session, the participants performed 3 tasks:

T1: First, led by a meeting facilitator, the participants took turns to introduce themselves to the others. The self-introducing participant was required to show his/her face to the camera so that those at the

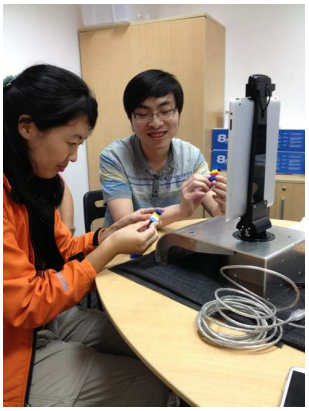

Figure 5: Demonstration the structure of one meeting room.

far end are able to see and hear him/her at the same time.

T2: In the second task, they engaged in an open-ended discussion on a specific topic. The facilitator ensured that everyone had ample opportunities to express his/her opinions on the topic, and stood up to summarize the key points on a writing board.

T3: Finally, they took turns to demonstrate a simple assembly of 6 lego pieces into any structure of their choice. The participants at the far end replicated the lego structure with the demonstrator's help if needed. The demonstrator was required to verify that the structure is correctly assembled, before the facilitator passed the turn to the next user to demonstrate.

These tasks were first carried out with a pair of *immobilized* telepresence robots (experimental condition 1), one in each room, and subsequently *attention-directed* telepresence robots (experimental condition 2). The same group of users participated in both experimental conditions. As such, to avoid repeating the same discussion, which may disengage the users, we used a different topic of discussion for each experimental condition, namely,

- *"What is your prefered way to reach out to young students and arouse their interest in robotics?"*

for experimental condition 1, and

- *"Is it possible that the robots enable old or sick people to live independently for longer?"*

for experimental condition 2.

For experimental condition 1, users are free to manually move the base of the robotic platform to re-orient the camera or display. They are also free to adjust their sitting positions with respect to the robot by shifting the lightweight foldable chair.

3.2 Setup

We invited eight persons, comprising six males and two females, to participate in the experiment. Six of them are researchers and engineers from our robotics group, and the remaining are engineering graduate students. Their age ranged from twenty-two to forty years. The users were divided into two groups, each with three males and one female.

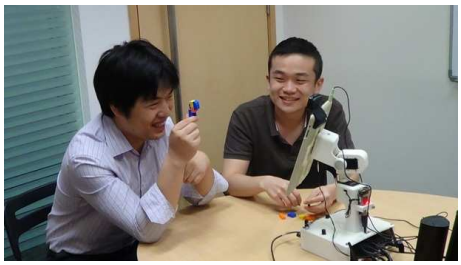

Figure 6: Demonstration the structure of the other meeting room.

Figure 7: The facilitator summarizes key points on a writing board.

Figure 5 and 6 provide a glimpse of the rooms used for the experiment. We set up similar robotic platforms and structures in the two conference rooms.

There is a table in each meeting room with a height of 1 meter from the ground. The chairs for the users to sit on were without wheels, lightweight and foldable, and the users were able to adjust their sitting positions freely. The robot is put on the table and participants sat around this table, with the distance of participants to the robot ranging from 0.6 to 1.7 meters. There is also a writing board placed at about 2 meters to the robot in one of the rooms for the facilitator to record the key points. Figure 7 shows the location of the writing board relative to the robot and the participants.

The video-conferencing software used was Apple's FaceTime, running on iOS 6 operating system and over high speed WiFi connection.

3.3 Hypotheses

In the experiment, we compare two conditions. For condition 1, the robot is fixed, i.e. when one person moves or the new source sound comes, the display and camera do not follow the movement and the new sound. For condition 2, the robot moves automatically to track the faces of users and face the direction of a speaking user, according to the methods described in Section 2.

We propose the following three hypotheses, which are to be tested in the user study:

1. The feeling of facing a remote person, as though he/she were in the same room, increases in the presence of automatic attention direction.

2. The ease of show and tell in video-conferencing increases in the presence of automatic attention direction.

3. The flow of video-conferencing communications is smoother in the presence of automatic attention direction. By this, we mean that there are less extraneous events or stimuli that will disturb the way the meeting proceeds.

4 Results Analysis and Discussion

After all of the tasks had been completed for both experimental conditions, the users were requested to fill out a questionnaire (adapted from [13]) with 12 questions:

Q1. The live video and audio were clear enough.

Q2. The communication was responsive.

Q3. It was easy to learn the assembly.

Q4. The self-introduction round was smooth and easy.

Q5. I can easily follow the summary points by the facilitator.

Q6. I felt distracted during the meeting.

Q7. I felt physical relax and comfortable throughout the meeting.

Q8. I felt more fatigue than in a normal face-to-face meeting.

Q9. I felt I can show and tell naturally.

Q10. I felt as if I were talking with remote users in the same room.

Q11. I felt as if I were viewing remote users in the same room.

Q12. I felt as if I were being viewed by remote users in the same room.

Each question was rated by the users on a 9-point Likert scale, where 1 = strongly disagree, 3 = disagree, 5 = neutral, 7 = agree, and 9 = strongly agree. Furthermore, we have a question, rated on a binary scale, which asks about the overall user preference:

Q13. I prefer a robot with automatic attention direction to one without automatic attention direction.

The results of the questionnaire are shown in Figure 8. Each bar represents the mean value of the responses to each statement in each condition, and each bar represents the standard error of the mean value. To assess if there is a statistically significant difference in the means of the user ratings over the 2 experimental conditions, we perform a paired-sample t-test.

We found a significant difference in the feeling that the self-introduction was smooth and easy (Q4, $p = 0.0122 < 0.05$). The comparison shows that the feeling was significantly stronger in the moving condition than in the fixed platform. There is also a significant difference in the ease of following the summary points by the facilitator (Q5, $p = 0.0302 < 0.05$). The comparison shows that it was significantly easier to follow the summary points in the automatic tracking condition than in the fixed condition. We found a significant difference in the feeling that it is natural to show and tell in the two different conditions (Q9, $p = 0.0039 < 0.01$). The comparison results suggest that users felt more natural to show and tell in the automatic tracking condition. The feeling of talking with remote users in the same room was significantly stronger in the moving condition than in the fixed condition (Q10, $p = 0.0049 < 0.01$). Moreover, we found the feeling of being viewed by remote users in the same room was significantly stronger in the moving condition. $p = 0.018 < 0.05$ for the question Q12.

There was not much difference in the effects of the live video and audio conversation for the fixed and moving cases. Thus the audio-visual quality seemed not to affect the results. Though there was no statistically significant difference in the other questions, from the comparison of mean values in the moving condition and fixed condition, most of subjects had the better feeling for the moving condition.

Now, we analyze the results with respect to Hypothesis 1: that the feeling of facing a remote user in the same room increases in the presence of automatic attention direction. The comparison of the response to the last three questions (Q10, Q11 and Q12) shows that the feeling of talking with remote users, as though they were in the same room, was significantly stronger in the moving condition than in the fixed condition (Q10). The feeling of being viewed by users at the far end, as though they were in the same room, was also increased in the moving condition (Q12). For question Q11, the t-test yields $p = 0.0522$, which is, in fact, very close to the 0.05 significance level. Though not conclusive, it seems to suggest that participants' feeling of facing a remote user in the same room also increases when the robot can automatically track the focus in the video conference. Thus, the results from Q10, Q11 and Q12 provide support for Hypothesis 1.

Next, we consider Hypothesis 2: that the ease of show and tell increases when the robot involves automatic attention direction tracking. Because of field-of-view constraints in the video-conferencing scenario, we need to look at both sides of the process, i.e., if the subjects is easy to show and tell naturally (Q9) and meanwhile if they are easy to learn and follow the motions from the other person (Q3 and Q5). Users' responses to question Q9 indicate that it was significantly easier to show and tell naturally in the automatic attention tracking condition. At the same time, the responses to question Q5 also showed a significantly stronger feeling to follow the summary points by the facilitator. Although the participants did not feel a significant increase in the ease of learning the assembly from a remote user in the

Figure 8: Means and standard deviations of users' scores for the questionnaire. A single asterisk '*' indicates $p < 0.05$ and a double asterisk '**' $p < 0.01$.

presence of automatic attention direction, we found, from the mean values and standard deviations (no overlap between the two conditions), that most of subjects thought it was still easier to learn the assembly when the robot has automatic attention direction. Collectively, these results provide good support for Hypothesis 2.

Hypothesis 3, which concerns the flow of video-conferencing communication, is related to questions Q4, Q6, Q7 and Q8. Firstly, the result of Q4 showed a significant increase in the feeling that the self-introduction was smooth and easy. Besides this, the feeling of being distracted during the meeting is also an important factor to measure the smooth flow of the video-conferencing communications. From the result of Q6 ($p = 0.0703$), though there was no *significant* difference in the feeling of being distracted, the subjects still have a diminished feeling of being distracted in the presence of automatic attention direction. Another concern related to Hypothesis 3 is if the subjects felt physically relaxed and comfortable throughout the meeting, or if they feel more fatigued than in a normal face-to-face meeting (Q7 and Q8). Though there was neither any significant difference in the responses to Q7 nor Q8, the mean values still show an increase in the presence of automatic attention direction, i.e., subjects on average felt more relaxed and less fatigued.

A limitation of the current user study is the small sample size. Although there is no minimum sample size for the t-test to be valid, a small sample size results in a low statistical power. For this study, we chose only 8 samples after considering the trade-off between the cost of having more participants and the benefit of getting more statistical power at this preliminary stage, though clearly the small sample size is insufficient to obtain a high power. In our future work, we plan to increase the number of participants to at least 25, in order to achieve a more reasonable statistical power.

Finally, the responses to Q13 reveal that the participants preferred a robot with automatic attention direction to one that is immobilized with a mean score of 7.13. In addition, they gave a number of constructive suggestions and comments as follows:

1. It should be easy for users to switch between an automatic mode and a fixed one, to allow the users to take over the control of the robot at any time.
2. Face tracking should only be activated when required.
3. The robot should allow a user to choose who to view at the remote end.
4. The robot at the side with listening users should alternate attention among the listeners so that the speaker at the other end is not locked on a single listener.
5. It is sometimes difficult for a user to view the screen when the tablet is facing another user in the same room.

Gathering these useful feedback from the participants, we can take further steps to improve the functions and performance of the social telepresence robot. For example, items 1 and 2 can be easily realized by a remote control. In order to have the functions in item 3 and 4, it is necessary to establish the communication and cooperation between the two robots. We can solve the problem in item 5 by connecting a projector.

5 Conclusion

We have presented a robotic telepresence platform with audiovisual attention control that enables automatic attention direction in response to audio-visual stimuli. Based on the results of a user study designed for a telepresence scenario, in which two conditions (fixed and attention-directed) were compared, we observed that in the presence of attention direction, the feeling of presence increases, the process of show and tell becomes easier and more natural, and the flow of the video-conferencing communications is smoother. As such, our study suggests that it is possible to enhance social telepresence by using an attention-directed robot. Feedback from participants shed light on further improvements and expanded functionalities of the telepresence robot that are required to enhance its usability, the chief among them being the ability for users to override the automatic control of the robot in an ad hoc manner should the need arise.

References

[1] Yang, R., Zhang, Z.: Eye gaze correction with stereovision for videoteleconferencing, *Proc. 7th European Conference on Computer Vision-Part II*, pp. 479-494 (2002)

[2] Kim, H., Komatani, K., Ogata, T., Okuno, G.: Auditory and visual integration based localization and tracking of humans in daily-life environments, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2021-2027 (2007)

[3] Bernardin, K., Stiefelhagen, R.: Audio-visual multi-person tracking and identification for smart environments, *Proc. 15th Int. Conf. on Multimedia*, pp. 661-670 (2007)

[4] Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., Horaud, R.: Audio-visual clustering for multiple speaker localization, *Proc. 5th International Workshop on Machine Learning for Multimodal Interaction*, pp. 86-97 (2008)

[5] Nakadai, K., Hidai, K., Mizoguchi, H., Okuno, H.G., Kitano, H.: Real-time auditory and visual multi-object tracking for humanoids, *Proc. 17th International Joint Conference on Artificial Intelligence*, pp. 1425-1432 (2001)

[6] Rodemann, T., Joublin, F., Goerick, C.: Audio proto objects for improved sound localization, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 187-192 (2009)

[7] Yan, R., Rodemann, T., Wrede, B.: Simple auditory and visual features for human-robot dialog scene analysis, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 700-706 (2012)

[8] Hasnain, S.K., Gaussier, P., Mostafaoui, G.: Synchrony as a tool to establish focus of attention for autonomous robots, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2423-2428 (2012)

[9] Aryananda, L.: Attending to learn and learning to attend for a social robot, *Proc. 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 618-623 (2006)

[10] Adalgeirsson, S.O., Breazeal, C.: MeBot-A robotic platform for socially embodied telepresence, *Proc. 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 15-22 (2010)

[11] Chng, E.S.: A microphone array with a 3-dimensional configuration for the I2R social robot, *Technical Report, Institute for Infocomm Research, A*STAR.* (2012)

[12] Li, L., Xu, Q., Tan, Y.K.: Attention-based Addressee Selection for Service and Social Robots to Interact with Multiple Persons, *Proc. Workshop at SIGGRAPH Asia*, pp. 131-136 (2012)

[13] Nakanishi, H., Murakami, Y., Kato, K.: Movable cameras enhance social telepresence in media spaces, *CHI - Telepresence and Online Media*, pp. 433-442 (2009)