2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
September 14-18, 2014. Chicago, IL, USA,

Gesture-Based Attention Direction for a Telepresence Robot: Design and Experimental Study

Keng Peng Tee, Rui Yan, Yuanwei Chua, Zhiyong Huang, Somchaya Liemhetcharat

Abstract— **The application of robotics to telepresence can enhance user interaction experience by providing embodiment, engaging behaviors, automatic control, and human perception. This paper presents a new telepresence robot with gesture-based attention direction to orient the robot towards attention targets according to human deictic gestures. Gesture-based attention direction is realized by combining Localist Attractor Network (LAN) and Short-Term Memory (STM). We also propose audio-visual fusion based on context-dependent prioritization among the 3 types of audio-visual cues (gesture, speech source location, head location). Experiment results are very promising and show that i) the average gesture recognition rate is 92%, i) gesture-based attention direction rate is 90%, and that ii) only by considering the 3 types of audio-visual cues together can the robot perform on par with a human in directing attention to the correct person in a meeting scenario.**

I. Introduction

Telepresence technology, which enables people to communicate face-to-face over remote distances, is gaining importance and popularity as a useful tool at home and in the workplace for increasing productivity and connecting people. The application of robotics-related technologies to telepresence can enhance user interaction experience by providing embodiment, engaging behaviors, automatic control, and human perception.

To this end, research-oriented telepresence robot platforms have been developed to study human-robot interaction. The MeBot [1] is a small mobile articulated telerobot that allows the operator to express non-verbal cues such as hand and head gestures. The Texai robot is a mobile platform with a screen at standing-height, and can be controlled to move around a remote office to attend meetings. Powered by HARK, an advanced audition software with sound source localization, tracking and separation capabilities, a sophisticated dialogue management and auditory awareness has been demonstrated on the Texai robot [2]. To increase robustness of speaker tracking, audiovisual approaches have also been proposed. Omnidirectional vision combined with sound localization has been shown to track speakers as they are move around the robot [3]. Computational AudioVisual Scene Analysis (CAVSA) based on proto-objects and short term memory [4] has been designed to track speakers even if they disappear for a while. Our previous work studied audio-visual attention control based on speech source localization and visual face tracking for a telepresence robot, and presented a user study showing facilitation of show-and-tell and increased presence [5]. Audio-visual integration has been shown to improve robot audition by increasing speech recognition rate and robustness against noisy conditions [6]. Besides individual robot platforms, a multi-robot system has also been proposed as mobile telepresence solution for following multiple users and keeping them in view [7]. The above works focused on audiovisual human tracking, and do not consider the use of gestures to direct attention.

The authors are with the Institute for Infocomm Research, A*STAR, Singapore 138632. E-mail: {kptee, ryan, ychua, zyhuang, liemhet-s}@i2r.a-star.edu.sg.

Fig. 1. Desktop telepresence robot prototype with tablet and Kinect.

The use of gestures to direct attention has been studied in some robot applications, including instructing the robot on which object to grasp [8], and pointing to a location on the ground for a robot to move to [9] . Vision-based gesture recognition using self-organizing feature map has been used to control an entertainment robot AIBO to perform discrete actions [10]. Gesture recognition has been studied using Localist Attractor Network (LAN) to control a simulated mobile robot to perform simple actions like move, stop, and turn left or right [11]. Gesture-based robot control has also been demonstrated using a wearable sleeve interface with EMG and IMU sensors [12] and wirelessly transmitted accelerometer signals from users' hands [13]. Instead of recognizing human gestures, recognition of full-body gestures of a small humanoid robot is studied in [14] to allow playful interactions with humans. These works focus only on the gesture recognition problem and do not consider fusion with other attention direction modalities.

In this paper, we propose gesture-based attention direction as a natural mode of human-robot interaction, and fusing this interaction mode with other active attention-directing behaviors (speech source localization and visual tracking) to enhance the overall telepresence experience. We highlight the contributions of this paper as follows:

1) Design and proof-of-concept of a novel telepresence robot with gesture-based attention direction enabling

978-1-4799-6933-3/14/$31.00 ©2014 IEEE

automatic orienting of the robot towards attention targets according to human deictic gestures.
2) Fusion of gesture-based attention direction with speech source localization and head tracking to allow group videoconferencing scenarios. To the best of the authors' knowledge, this is the first desktop telepresence robot to provide all 3 features together.
3) Quantitative empirical evaluation of the performance of the complete fused system in a structured meeting scenario with human participants.

II. System Design

It is the object of this paper to enhance telepresence robots with natural interaction modes and automatic control of attention focus based on natural conversation cues. Human gestures are useful cues for the robot to switch attention seamlessly to another person. In one use case, a speaker who has finished his speaking turn can pass the turn to another speaker in the same room. In another use case, a speaker who is addressing a group of listeners in a remote room can use gestures to alternate attention amongst the listeners.

We highlight the design requirements of the robotic system as follows:

1) Fluid and quick-stabilizing point-to-point motion without vibrations/oscillations
2) Ability to visually track a moving human in the camera's field of view
3) Ability to localize a speech source and direct attention to the (out-of-view) speaker
4) Ability to recognize deictic gestures and direct attention to the target person

In our previous work [5], we have developed a desktop robot for group telepresence that addressed the first 3 points. Therefore, it is the focus of this work to address the last requirement, namely the provision of gesture based attention direction and fusion with speech source localization and visual head tracking.

A. Telepresence Robot Hardware

The desktop telepresence robot 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, which provides video steaming of users at the far end through built-in video-videoconferencing communications, is mounted on the pan-tilt unit, and a Kinect sensor is, in turn, mounted at the top of the tablet, as shown in Figure 1. Since the tablet moves together with the Kinect 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.

In our prototype, the microphone array comprises 8 Shure omnidirectional microphones evenly distributed on a curved surface. The top of the microphones can be seen in Figure 1 as silver small circles on the surface of the black base. The audio signals are amplified and synchronized by a MOTU 896mk3 audio interface, before being fed to the computer for audio tracking. We use Robotis Dynamixel MX-106 servomotors, which are computer-controlled via the usb-plug-and-play USB2Dynamixel device, and provide more than sufficient torque to move and hold the tablet in all possible orientations. The audio tracking, visual tracking and motion control modules all run on a PC with Intel Core i7 64-bit dual-core CPU and Microsoft Windows 7 operating system.

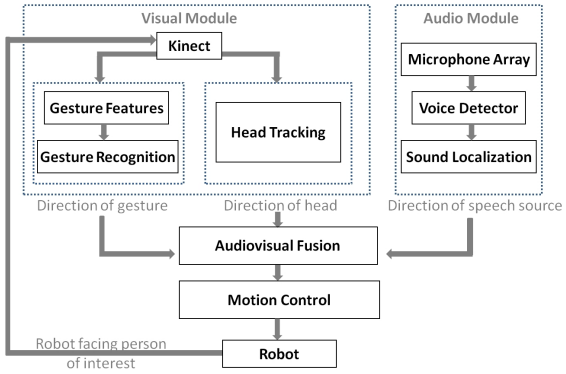

Fig. 2. Schematic overview of audiovisual attention control module

B. Attention Control Software

The attention control architecture is shown in Figure 2. The system consists of the input layer, two hidden layers and the output layer. A Kinect and a 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 head tracking sub-system and a gesture recognition sub-system. In the audio module, a speech source localization sub-system detects and localizes human voice. In the second hidden layer, an audiovisual fusion module is used to integrate multiple attention cues from the first hidden layer to obtain an attention target. Then, the motion control module ensures that this attention target is tracked smoothly by the robot.

The head tracking module tracks the head position of the nearest skeleton detected in the Kinect's field of view using Microsoft Kinect for Windows SDK. Speech source localization is based on the algorithm in [15] , which uses a Voice Activity Detector to discriminate human voice from irrelevant sounds, followed by a combination of Time Delay of Arrival and Steered Beamformer techniques to determine the direction of the speech source. Gesture recognition and full audiovisual fusion with speech source localization and head tracking will be covered in the subsequent sections. To move the robot a smooth and human-like fashion, we use a minimum-jerk trajectory for each joint angle. Proportional-derivative control is used to drives the motors to track this minimum-jerk trajectory.

III. Gesture-Based Attention Direction

Gesture-based attention direction is realized by combining gesture recognition with Short-Term Memory (STM). The rationale for using STM is that building memory helps

the robot to reduce search space and confirm the precise locations of the attention targets.

We use the Localist Attractor Network (LAN) [16] for gesture recognition because it requires only a small amount of training data but can obtain high accuracy [5]. In the LAN model, a recognition target is reached when the state converges to an attractor. Let w_j be the attractor for the jth gesture class, π_j the connection weight, σ_j the width of the attractor basin, ε the initial state, and $y(t)$ the current state. The LAN model is described by [16]:

$$y(t+1) = \alpha(t)\varepsilon + (1-\alpha(t))\sum_j q_j(t)w_j \quad (1)$$

$$q_j(t) = \frac{\pi_j e^{-|y-w_j|^2/2\sigma_j^2}}{\sum_l \pi_l e^{-|y-w_l|^2/2\sigma_l^2}} \quad (2)$$

$$\alpha(t) = \sigma_y^2(t)/(\sigma_y^2(t)+\sigma_z^2) \quad (3)$$

$$\sigma_y^2(t) = \frac{1}{n}\sum_j q_j(t)|y(t)-w_j|^2 \quad (4)$$

where n is the state dimension and σ_z is a non-negative constant that accounts for the unreliability of observation ($\sigma_z = 0$ means wholly unreliable observation).

Let the low pass filtered skeleton data of Kinect be:

$$\mathbf{X} = \{X'_{t,le}, X'_{t,lw}, X'_{t,re}, X'_{t,rw}\}_{t=1}^{N}$$

where $X'_{t,le}$, $X'_{t,lw}$, $X'_{t,re}$ and $X'_{t,rw}$ are, respectively, the position vectors at time t for the left elbow, left wrist, right elbow and right wrist, taken from the ipsilateral shoulder. By using Fast Fourier Transform (FFT), the coefficients of the m lowest frequencies of $\mathbf{X}$ are obtained as the feature vector $F = \left[\begin{array}{cccc} f_1 & f_2 & \cdots & f_{2m} \end{array}\right]$, which consists of $2m$ features for each gesture candidate. Combining k example gestures for the jth class and ith person, we have

$$F_{i,j} = \left[\begin{array}{cccc} f_{1,1}^{i,j} & f_{2,1}^{i,j} & \cdots & f_{2m,1}^{i,j} \\ \vdots & & & \\ f_{1,k}^{i,j} & f_{2,k}^{i,j} & \cdots & f_{2m,k}^{i,j} \end{array}\right]$$

Finally, we obtain the feature matrix of gesture class j for n different persons as $F_h^j = [F_{1,j}^T, ..., F_{n,j}^T]^T$.

We use a self-organizing map to obtain a low dimensional space where it is easier to find the center of F_h^j. After the mapping, F_2^j is denoted as the corresponding position of F_h^j in the 2D space. Thus, F_2^j is taken as the attractor basin for the jth class, and the center of F_2^j is chosen as the jth attractor, denoted by w_j.

Now, we propose Algorithm 1 to direct the robot's attention based on recognized gestures and STM, which contains the memorized locations of the users as they were last observed by the robot. For a new gesture candidate with feature vector F, we initialize $\varepsilon = F$. Then, we iterate y using (1) until y converges to some w_j, which yields the recognized gesture class. After that, the gesture recognition result is associated with STM to find the attention target P_a. Specifically, it is the position in STM that is closest to the pointing direction.

Algorithm 1: Gesture-Based Attention Direction.

Data: $\{P_1, \cdots, P_n\}$ as positions in STM, $\{w_1, ..., w_m\}$ as attractors to m gesture classes, and F as features for new detected gesture.
Parameters: π_j, σ_j and σ_z in LAN model.
Result: Attention target P_a.
begin
1. Initialize $\varepsilon = F$.
2. Iterate y using (1) until y converges to an attractor w_j.
3. Calculate distance $e_k = \|y - w_k\|$, for $k = 1, ..., m$, and find the directions D_1 and D_2 corresponding to the least and second least e_k values respectively.
4. Determine the attention target as follows:
   If there is a position P_j in STM corresponding to the speaker's direction D_1, then $P_a = P_j$.
   Else find the position P_j between the two directions D_1 and D_2 and set $P_a = P_j$.

end

IV. Full Audio-Visual Fusion

Audio-visual fusion resolves conflict among multiple attention cues from head tracking, speech source localization and gesture recognition, and uses short term memory to associate user-specific audiovisual features with the last tracking location, so that more accurate localization of out-of-view attention targets identified from audio or gesture cues can be obtained. We require the robot to turn around and memorize the accurate positions of all persons in the meeting room by combining the results of head tracking and sound localization. These positions will be put into STM. The head tracking module confirms precise location and updates the memory when any user changes position.

Define the following conditions in the architecture:
Condition 1: The speech source has the same azimuth angle region as that of the person detected in the robot's view.
Condition 2: Human gesture is detected by the robot.

We detail the steps for fusing audio-visual cues to determine the attention target P_a and update the STM:

1) Check for attention cues from gesture recognition, head tracking and speech source localization modules. If the received cue comes from a single module only, then the attention target is trivially obtained. If multiple cues are received, check Condition 1.
2) If Condition 1 is satisfied, either head tracking or gesture recognition takes priority over the auditory cue. If Condition 2 is true, then the attention target is determined from gesture as outlined in Algorithm 1. Otherwise, the attention target is the head of the person in the robot's view.
3) If Condition 1 is not satisfied, speech source localization takes priority over visual cues to determine the attention target. Algorithm 2 describes speech-based attention direction and STM updating.
4) The STM is updated with the current focus.

For the updating of STM in Algorithm 2, a simple way is to directly store the attention target P_a into STM. But a more accurate position can be achieved with the help of head

tracking. Thus we update STM by adding this new position and not the attention target P_a.

Algorithm 2: Speech-Based Attention Direction and STM.

Data: $\{P_1, P_2, \cdots, P_n\}$ as positions in STM,
P_s as target from speech source localization.

Parameter: Bound of $O(\Omega) \in R^2$, where $O(\Omega)$ is a neighborhood region of $O = (0, 0)$.

Result: Attention target P_a and updated STM.

begin

1. Compute the distance $e_i = \|P_i - P_s\|$ for $i = 1, \cdots, n$
2. Find minimum e_i and memorize index j.
3. Find position P_j corresponding to index j and determine attention target:
   If $(P_j - P_s) \in O(\Omega)$
   then $P_a = P_j$ and STM remains unchanged;
   Else $P_a = P_s$ and update STM.

end

V. Experiment Results

The experiment consists of two parts: the first part tests the accuracy of gesture-based attention direction, while the second part investigates the performance of the full audio-visual attention-directed robot in a scenario involving real-time interactions with users.

First, to train the LAN model, 5 users demonstrated 20 deictic gestures in each of 5 directions, namely the Front, Right, Right45, Left and Left45. The following parameters are chosen in the LAN model: $\pi_j = 1$, $\sigma_j = 0.03$ and $\sigma_z = 0.1$. For online testing, 3 persons participated, two of whom come from the training set (Users 1 and 2) and one is a new user (User 3). Each person performed 40 instances of each gesture class. The LAN model is able to classify 94.45%, 94% and 88% of the gestures correctly for Users 1, 2 and 3 respectively, even though User 3 is not in the training set.

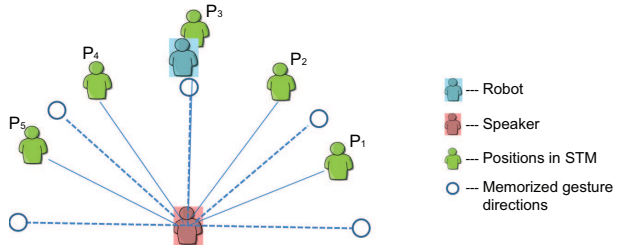

Fig. 3. Meeting room layout for gesture-based attention direction test.

A. Gesture-Based Attention Direction

The objective of this experiment is to test the ability of Algorithm 1 in determining the correct attention target from deictic gestures. We have 5 STM positions, P_1 to P_5, whose directions are 60 and 30 degrees to the right, front, and 30 and 60 degrees to the left, respectively, as depicted in Fig. 3. Assume that the current speaker is always facing the robot camera. According to Algorithm 1, the two most similar gesture classes in the LAN model are found, and the attention target determined by interpolating between the two directions. For example, if the speaker gestures towards P_1, then the two nearest memorized gesture directions, Right and Right45, are determined after the LAN model converges. Thus the attention target is found in the region between these two directions. Table I shows the recognition rates based on 20 gestures towards each STM position. Overall, about 90% of the attention targets can be inferred correctly, even if the gestures point to directions not memorized in the LAN model. This illustrates the robustness of the gesture-based attention direction algorithm.

TABLE I
Gesture-Based Attention Direction

		Attention Target				
		P_1	P_2	P_3	P_4	P_5
Gesture	100	22	18	20	26	14
Right30	20	19	1	-	-	-
Right60	20	3	17	-	-	-
Front	20	-	-	20	-	-
Left30	20	-	-	-	20	0
Left60	20	-	-	-	6	14

TABLE II
Sequence of Scenario Events

Phase	Description	Trigger	Target	Cues
start	Y is seated facing the robot camera. Nobody is talking.	-	Y	Vision
A	X gives an introduction and gestures towards Y to invite Y to the white board to present something.	X talks.	X	Speech Vision Gesture
B	Y stands up and walks to the white board.	X gestures towards Y.	Y	Vision
C	Y talks and points to material on the white board.	Y talks.	Y	Speech Vision
D	Z interrupts with a question when Y pauses, and gestures towards Y to give the floor to Y.	Z talks.	Z	Speech Vision Gesture
E	Y ponders the question and prepares to reply.	Z gestures towards Y.	Y	Vision
F	Y answers the question.	Y talks.	Y	Speech Vision
end	X concludes the session.	X talks.	X	Speech Vision

B. Full Audio-Visual Attention Direction

To test the integration of all modules together, we designed a scenario[1] that is representative of typical meeting situations, and that includes behaviors like taking turns to talk, walking to a white board and presenting with material on the white board, as well as using gestures to reference another person.

The main aim of this study is to capture the view seen by the other end of the video-conferencing under various conditions, namely:

[1] Supplementary video available at:
http://www1.i2r.a-star.edu.sg/~kptee/Video/

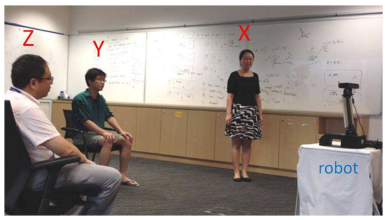

Fig. 4. Experiment scenario.

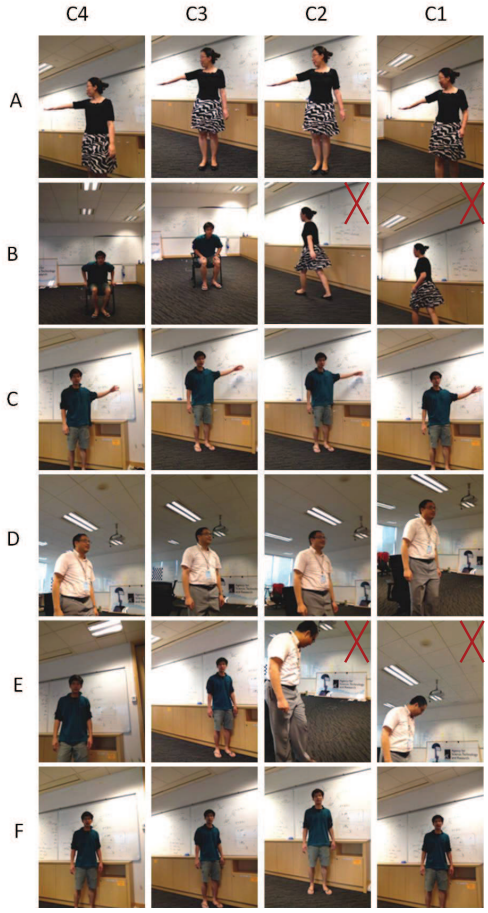

Fig. 5. Snapshots from tablet camera at different phases A-F and different conditions C1-C4. A cross denotes wrong target fixation.

C1) Speech source localization only.
C2) Speech source localization and head tracking.
C3) Speech source localization, head tracking, and gesture recognition.
C4) Manual control by a user (i.e. "ground truth").

We use the tablet camera to capture video footage for analysis. Since the video-conferencing software is launched from the tablet, the video recorded by the tablet camera is exactly what the users at the far end would see. We use OpenCV face detection together with manual annotation to estimate the target subject's head position in the videos.

Three subjects, "X", "Y",and "Z", participated in the experiment. Figure 4 shows the room layout and the positions of the subjects and robot. A round of self-introduction is initiated in order to memorize the positions of the subjects in the STM based on Algorithm 2. After that, the scenario commences based on the sequence detailed in Table II.

The importance of C4 is that it is a "ground truth" for which we can compare the robot's performance in C1-C3. C4 involves manual control where a fourth user, standing behind the robot, uses both hands to freely orientate the tablet about its pan and tilt axes.

Figure 5 provides a qualitative comparison between C1-C4 in terms of the actual view seen from the tablet camera. C3 gives views that are most consistent with C4 across all phases A-F. The main difference between the views of C3 and C4 is that C3 positions the person's head nearer to the top of the image. Unlike C3, C1 and C2 do not capture the views of the correct attention target Y in phases B and E.

For quantitative comparison, we define an error measure:

$$d = \begin{cases} \|p - p^*\|, & \text{if correct target} \\ \bar{p}, & \text{otherwise} \end{cases} \quad (5)$$

where p and p^* are the actual and desired locations of the attention target's head in the image, respectively, and $\bar{p}$ a penalty when the attention target according to Table II is not in the image. We select $p^* = (240, 213)$ pixels, counting from the top left corner of 480×640 resolution images. This desired head position provides ease of capturing body language. Also, we set $\bar{p} = 240$ pixels, which is the minimum estimate of the location of the out-of-view attention target.

Figure 6 shows a comparison of the error measure d for Conditions C1-C4. We observe that:

1) In phase A, the error magnitudes and profiles of C1-C3 are generally similar to that of reference C4. The initial discontinuity is due to fast robot movement to a new off-display target.
2) In phase B, C3 yields low errors since the robot turned to the correct target Y after recognizing the gesture from X. In contrast, the errors for C1 and C2 are high because the correct target Y is not in the image for some time. The fluctuation of error for C3 and C4 after finding Y is due to Y standing and walking to the white board. For C2, the error decreases when Y walks into the robot's view. The error for C1 exhibits a V shape because Y walks across the robot's view without being tracked.
3) Phase C has similar errors for all the conditions since only Y is standing at the same position while talking.
4) In phase D, the errors for all conditions are similar, but C1 is slightly higher since it is unable to track the fine motion of Z.
5) In phase E, the error for C3 is close to that for C4 due to gesture-based attention direction, but those for C1 and C2 peak throughout the entire phase.
6) In phase F, the error for C3 is again similar to that for C4. For C1 and C2, the error decreases from a large value as the robot turns attention from Z to Y upon the speech cue.

Taking the integral of the error over the entire duration of the scenario, we see, in Table III, that C4 has the least value, and C3 is the closest to C4. Without gesture detection,

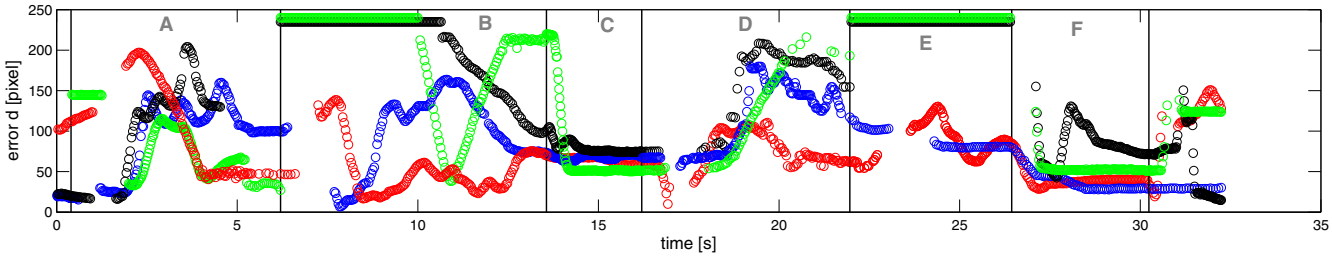

Fig. 6. Error measure for Conditions C1-C4, represented by green (C1), black (C2), blue (C3) and red (C4) markers. C3 has similar error to C4, but C1 and C2 have large errors during phases B and E where attention targets are specified by gestures.

TABLE III
COMPARISON OF ERROR INTEGRAL

Trial	C4	C3	C2	C1
Error Integral [pixel s]	2010.8	2722.7	3026.7	3094.5

C1 and C2 incurred higher error than C3 and C4 in phases B and E. In summary, it is necessary for the telepresence robot to have all 3 features – speech source localization, head tracking *and* gesture recognition – in order to perform on par with a human on directing attention to the correct person in a meeting scenario.

VI. CONCLUSION

We have developed a telepresence robot that automatically directs attention with speech source localization, head tracking, as well as gesture recognition. Audio-visual fusion resolves potential conflict among multiple attention cues. Compared with existing approaches which use the separate visual or auditory module or even simply integrate the two modules together, our proposed system provides the following advantages: (i) gesture recognition sub-system has been included into the whole system to improve the attention detection from speakers; (ii) a short-term memory has been integrated in the audio-visual fusion to enhance robustness and accuracy of attention direction. The outcome is a more natural user interface with automatic control of attention focus based on natural conversation cues. Our experiment results show that the gesture-based attention direction algorithm can achieve more than 88% accuracy. Additionally, the full-feature audio-visual attention-directed robot performs on par with a human in directing attention to the correct person in a meeting scenario. Excluding gesture recognition, or both gesture and head tracking, results in the robot fixating on the wrong person. Future work will include large-group user study under real unstructured group videoconferencing scenarios. It is important to investigate user experience from both sides of the telepresence session, as well as robustness against noise in the environment.

ACKNOWLEDGMENTS

This work was partially supported by the I^{2}R Productivity Programme. The authors are grateful to Mr Jason Teo Kok Yung for design work and Ms Yow Ai Ping for help with the experiments.

REFERENCES

[1] S.O. Adalgeirsson and C. Breazeal, "MeBot-A robotic platform for socially embodied telepresence," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2010, pp. 15–22.
[2] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H.G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using HARK," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2011, pp. 2130–2137.
[3] J. Imai T. Suwannathat and M. Kaneko, "Omni-directional audio-visual speaker detection for mobile robot," in *Proc. IEEE Int. Conf. Robot and Human Interactive Communication*, 2007, pp. 141–144.
[4] R. Yan, T. Rodemann, and B. Wrede, "Simple auditory and visual features for human-robot dialog scene analysis," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 700–706.
[5] R. Yan, K. P. Tee, Y. Chua, Z. Huang, and H. Li, "An attention-directed robot for social telepresence," in *Proc. Int. Conf. Human-Agent Interaction*, 2013, pp. III–1–2.
[6] T. Yoshida, K. Nakadai, and H. G. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2009, pp. 604–609.
[7] N. Karnad and V. Isler, "A multi-robot system for unconfined video-conferencing," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2010, pp. 356–361.
[8] J. J. Steil, G. Heidemann, J. Jockusch, R. Rae, N. Jungclaus, and H. Ritter, "Guiding attention for grasping tasks by gestural instruction: the gravis-robot architecture," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2001, pp. 1570–1577.
[9] S. Abidi, M. Williams, and B. Johnston, "Human pointing as a robot directive," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2013, pp. 67–68.
[10] T. Hashimaya, K. Sada, M. Iwata, and S. Tano, "Controlling an entertainment robot through intuitive gestures," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 2006, pp. 1909–1914.
[11] R. Yan, K.P. Tee, Y. Chua, H. Li, and H. Tang, "Gesture recognition based on localist attractor networks with application to robot control," *IEEE Computational Intelligence Magzine*, vol. 7, no. 1, pp. 64–74, 2012.
[12] M. T. Wolf, C. Assad, M. T. Vernacchia, J. Fromm, and H. L. Jethani, "Gesture-based robot control with variable autonomy from the JPL BioSleeve," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2013, pp. 1160–1165.
[13] X. H. Wu, M. C. Su, and P. C. Wang, "A hand-gesture-based control interface for a car-robot," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2010, pp. 4644–4648.
[14] M. D. Cooney, C. Becker-Asano, T. Kanda, A. Alissandrakis, and H. Ishiguro, "Full-body gesture recognition using inertial sensors for playful interaction with small humanoid robot," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2010, pp. 2276–2282.
[15] E.S. Chng, "A microphone array with a 3-dimensional configuration for the i2r social robot," in *Technical Report, Institute for Infocomm Research, A*STAR.*, 2012.
[16] R. S. Zemel and M. C. Mozer, "Localist attractor networks," *Neural Computation*, vol. 13, pp. 1045–1064, 2001.