An Attention-Directed Robot for Social Telepresence

August 9, 2013

Rui Yan, Keng Peng Tee, Yuanwei Chua, Zhiyong Huang, Haizhou Li

Institute for Infocomm Research (I²R)
A*STAR, Singapore

Presented by Zhiyong Huang, 黃智勇
Outline

- Motivation
- Related work
- System prototype
- Technical details
- User study
- Conclusion and future work
Motivation

• Telepresence
  – Addressed network bandwidth, resolution, and the lack of eye-contact
  – The fixed screen and camera lead to disengagement in the telepresence experience
Related Work

- **Galileo**
  - Portable motorized iPad/iPhone holder

- **Kubi**
  - Manually controlled from iPad/iPhone/web browser

- **Helios**
  - Portable motorized iPad/iPhone holder
  - Manually controlled from iPad/iPhone/web browser
Related Work

Polycom EagleEye Director

- Dual cameras
- 7-mic array

- Automatic speech source localization + face zooming
- Must be connected to the Polycom HDX Room Telepresence Systems
- High cost (S$20k EagleEye Director + S$5k HDX system)
- Designed for large meeting rooms & non-portable
System Prototype

• An attention-directed robot
  – Direct attention automatically in a videoconferencing scenario based on *audio-visual stimuli* (speech, face)
Attention-Directed Telepresence Robot

- **user’s computer**
  - audio-visual attention control software
  - client SDK
- **user’s tablet**
  - camera
  - video-conferencing app
- **robot**
  - pan-tilt motors
  - motor control board
  - microphone array
  - audio capture card
- **I/O and communications interface**
- **audio-visual data**
  - attention focus
Technical Details

Visual Module
- Camera
- Face, Eyes & Mouth Detector
- Speaker Detector

Audio Module
- Microphone Array
- Voice Detector
- Sound Localization

Audiovisual Fusion Module

Tracking Control Module

Videoconferencing Communications

Direction of face to track
Direction of speech source

Tablet facing person of interest
Sound Localization

Camera view

External view
Sound Localization

• 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
Visual Tracking

Camera view

External view
Visual Tracking

• The OpenCV implementation of the Multi-Scale Haar Cascade Classifier algorithm
Audiovisual Fusion

Camera view

External view
Audiovisual Fusion
Results: Demo Prototype
User Study

**Subjects**
- Subjects’ age ranged from 20 to 40 years old.
- 8 subjects comprising 6 males and 2 females.
- Subjects are researchers, engineers and graduate students.

**Targets**
- Study what aspects of the proposed robot caused the users’ positive attitude.
Experiment

• The experiment scenario is that of a video-conferencing session between 2 groups of participants in 2 separate rooms.

• 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 have equal chances to speak, and summarize the discussion points.
Experiment

• 2 Conditions
  – The robot is fixed, *fixed*
  – The robot moves automatically to track the faces of users and face the direction of a speaking user, *moving*
Hypothesis

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.
Questionnaire

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.)
Questionaire

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

(rated on a *binary scale*, which asks about the overall user preference)
## Results

Means and standard deviations of users’ scores for the questionnaire. A single asterisk ‘*’ indicates $p < 0.05$ and a double asterisk ‘**’ $p < 0.01$.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Fixed</th>
<th>Moving</th>
</tr>
</thead>
<tbody>
<tr>
<td>The live video and audio were clear enough.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The communication was responsive.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>It was easy to learn the assembly.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The self-introduction round was smooth and easy.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can easily follow the summary points by the facilitator.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt distracted during the meeting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt physical relax and comfortable throughout the meeting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt more fatigue than in a normal face-to-face meeting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt I can show and tell naturally.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt as if I were talking with remote users in the same room.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt as if I were viewing remote users in the same room.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt as if I were being viewed by remote users in the same room.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

• The responses to Q13 reveal that the participants preferred a robot with automatic attention direction to one that is immobilized.
Suggestions from User Study

• 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.
• Face tracking should only be activated when required.
• The robot should allow a user to choose who to view at the remote end.
• 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.
• It is sometimes difficult for a user to view the screen when the tablet is facing another user in the same room.
Conclusion and Future Work

• An Attention-Directed Robot for Social Telepresence
  – Portable and automatic
  – Audio-visual fusion with short term memory
  – Tracking control that ensures face and hands always stay within field of view

• 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
The work was supported by I²R Robotics and A*STAR SERC Human Factor Engineering Programmes

zyhuang@i2r.a-star.edu.sg