International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2016)
D. Reiners, D. Iwai, and F. Steinicke (Editors)

Natural Interaction in Asymmetric Teleconference using Stuffed-toy Avatar Robot

Samratul Fuady[1], Masato Orishige[1], Li Haoyan[1], Hironori Mitake[1], and Shoichi Hasegawa[1]

[1]Laboratory for Future Interdisciplinary Research of Science and Technology, Tokyo Institute of Technology, Japan

Abstract

In this paper, we propose an asymmetric teleconference system using stuffed-toy robot as the representative of the remote user. Our main goal is to realize the system that can provide natural conversation both on the remote and the local side without immersing the user into a virtual environment. We consider envelope feedback gesture as the significant part to realize natural interaction, specifically turn-taking, eye gazing and beat gesture. We use stuffed-toy robot as the avatar robot because it can move very fast as well as increase familiarity and improve the interaction due to its softness structure. Furthermore, softness structure also brings safety and robustness because it will not be easily broken. In the remote side, we use wearable sensor and eye-tracking sensor to capture remote user's movement, thus enabling him/her to move and interact naturally as well. We conducted the teleconference experiment to evaluate our system. The result suggested that our system can improve the experience of the teleconference, especially in turn-taking aspect and transferring the beat gesture.

Categories and Subject Descriptors (according to ACM CCS): H.4.3 [Information Systems Applications]: Communications Applications—Computer conferencing, teleconferencing, and videoconferencing

1. Introduction

Nowadays, the research on improving the experience and interaction in telecommunication receiving a lot of attention. One of them is telepresence or tele-existence, which is defined as a system that enabling user to feel his/her presence in the different location from their real location. The face-to-face teleconference is one type of telepresence that is widely used today. This system can be grouped into two: symmetric and asymmetric. In the symmetric case, both sides will experience the similar thing. For example, the video conference where the group of people talks to real-time video of the other group, and the same thing happens on the other side. Another example of the symmetric telepresence system is by immersing all the conference members to the virtual environment; thus every member will experience the same thing with the other member [KWK10]. In the case of the asymmetric system, both sides will experience the different thing. The avatar robot is used as the substitute of the remote user in the local environment. In many cases, this system is necessary because only some of conference members are unable to attend the meeting. Furthermore, direct communication is still really needed instead of virtual communication in the near future.

Several telepresence mobile robots have been available commercially. For example Anybots, double robotics, mantarobot, etc [KCL13]. These robots are very helpful to give the remote user more access to the local environment by the ability to move, but the interaction that can be done is very limited [TDYU11]. They lack non-linguistic communication which is believed to be significantly important in communication [CT99].

Another system has been developed by [LTSB08] [JSG*15]. They develop a bear-like semi-autonomous robot avatar for family communication, education, and pediatric care. The good interaction can be achieved in local side, but in the remote user side, the interaction is not very natural, since the user needs to select the response or emotion manually. [CYL16] developed anthropomorphic robot as avatar robot. They use Kinect sensor and wearable hand glove to capture remote user movement. As for the face, they back project user's face into the robot's face so the emotion of the user can be shown. However, the configuration in the remote user side is not very convenient because the remote user needs to set up some equipment which is not easy and require quite big space. Besides, the hard mechanism they use for robot can decrease the interaction to the local user and life-like movements are also still become a tough challenge.

Other related research is Beaming Project [BES15]. They use HMD and eye-tracker to capture remote user movement and transfer those to the robot. They immerse the remote user in the virtual environment, which is the local environment captured by the camera. However, this system only focussed on head and eye movement. These type of feedback in many cases is not enough to convey the message from the speaker.

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

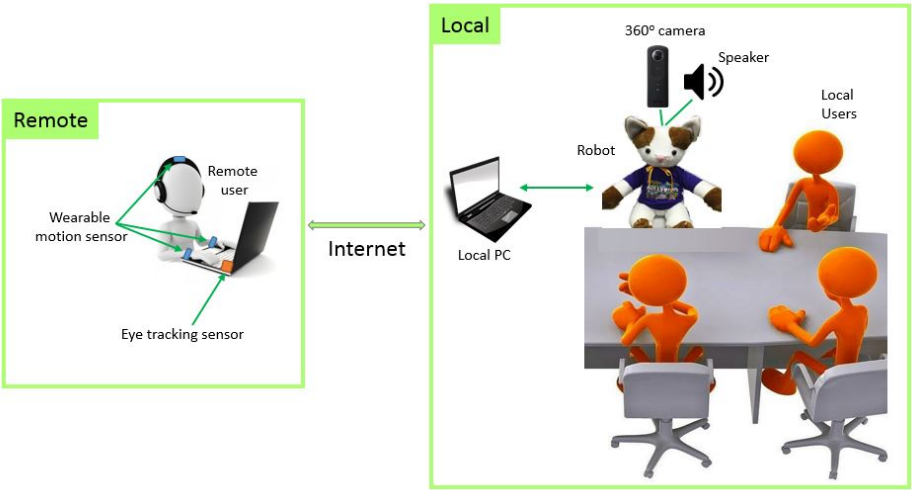

Figure 1: *Illustration of overall systems*

In creating natural interaction, compared to emotional feedback, envelope feedback gesture is more important in supporting the dialog [CT99]. Envelope feedback is the nonverbal behavior that is related to the process of the conversation, for example, nodding, gaze, and beat gesture.

In this paper, we develop telepresence system that enables natural interaction between the remote user and local user. We don't put the remote user into a virtual environment, but still facilitate his/her natural interaction. We use stuffed toy robot as the avatar robot that can increase the interaction in the local environment and able to perform some envelope feedbacks with fast response. As for the remote user, wearable devices is used to capture his/her movement naturally.

This paper is organized as follows. We describe the overall system overview in Section 2. Then we explain our avatar robot in section 3. In Section 4, we describe the system for the remote operator, while for the local user we describe it in Section 5. Then we present results of this system in section 6. We finally conclude our paper in Section 7.

2. System Overview

In order to perform conversation naturally, the turn-taking, eye gaze, and gesture are the minimal element that is needed. Eye gaze and gesture are important and act as nonverbal communication channel when doing the conversation. These also help in expressing attention and interest. The avatar robot should at least be able to express these elements to the local user. On the contrary, physical reproduction of gaze by interpreting and synthesizing it and reproduction of real-time images of the remote user and conference room are not so significantly important. Physical reproduction of gaze requires reproduction of physical positions of the conference side.

Our proposed telepresence system is described in Figure 1. On the local side (conference room), we have the stuffed-toy avatar robot in the form of a cat, local users, and theta camera to capture the whole local environment. The camera is 360° camera so that the remote user can see a wide range of local environment. We do not put the camera in the robot because it will make the remote user need more effort to understand the image due to the movement of the robot head. By installing the camera in the fix position, the local environment image will be easier to interpret by the remote user. And generally, it is also reasonable to set up the conference room before the meeting start.

On the remote side, we have remote user equipped with eye tracking sensor and wearable motion sensor. We use tobi eyeX eye tracker that can be easily placed in the remote user PC. For the motion sensor, we use ROHM motion sensor, which consists of an accelerometer, gyroscope, magnetometer, and pressure sensor. The sensor is placed on the remote user arms in the form of watches, and one in the head along with the headphone. The movement captured by motion sensor will be translated to robot movement, and the eye movement will be translated as robot head movement according to what the remote user focus on.

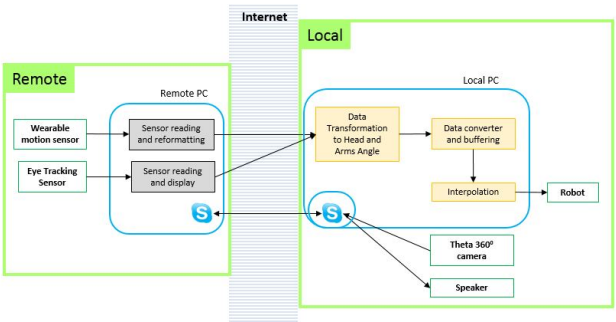

Figure 2: *Block diagram of the system.*

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

The current implementation of the system can be seen in Figure 2. For voice and video data, we use the available video chat application like Skype or Google Hangouts. The remote user will see the video of local side environment from this application. We run another program along with this application to capture the eye gaze of the remote user and read the data from the motion sensor. These data is sent to the local side, where the data will be processed to head and arm angle. This data will then be converted to string length and forwarded to the robot.

3. Stuffed-toy robot

3.1. Structure

We use stuffed toy robot which is modeled as a cat because it is generally liked by everyone and have a cute and nonintimidating look. Normally, the robot is seated, with the dimension of 400 mm x 300 mm x 240mm and weight about 1.5 kg.

It is built using a cotton-based soft material for the moving part. Thus, the moving parts have a softness and good hand feeling like the stuffed-toy animal. This configuration will improve familiarity, safety, and user interaction compared to the robot with hard material and mechanism. The softness also brings robustness because the robot will not be easily broken, unlike another robot with a hard link or shell. We can see the robot structure in Figure 3.

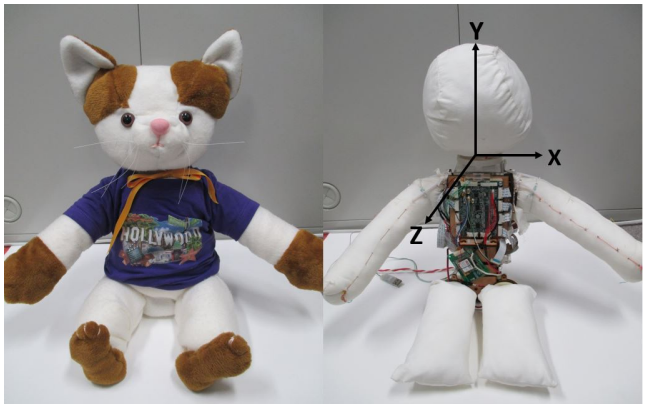

Figure 3: *Outside appearance (left) and inner structure of the stuffed-toy robot with the coordinate system (right).*

There are three moving parts of the robot: two arms and head. Each moving part has three strings attached to it. Each string is connected to the motor. By pulling these strings, the cylindrical cloth will be bent as shown in Figure 4. The moving part can move to any desired direction by pulling strings in combination. The motors and other circuit are put in the body parts to keep the softness of the moving parts of the robot.

3.2. Motion

Formulating equation model for this mechanism is quite difficult because it consists of multiple soft materials. A lot of factors can affect the characteristic of the moving parts, for example, the different in stitching or the amount of cotton stuffing. So we perform trajectory planning using interpolation from collected numerical data

Figure 4: *String mechanism used in stuffed-toy robot.*

[TMYH13]. We collect the data that map the position to the length of strings by pulling each string bit by bit and record the position using NDI Optotrak motion capture. After we collect the map data, the desired pose can be interpolated from the map data to get the length value for each string. Additionally, we interpolate several poses between the current pose and the desired pose, so the robot can move gradually which give the smoother motion.

The robot has the movement speed of $4.04 rad/s$. This speed is faster than commercially available robot like nao ($3.29 rad/s$) [GHB09], or iCub ($1.25 rad/s$) [TMS*07]. This is due to the robot moving part that is very light because it is made from cotton and fabric.

The control for each string is using PID control. Figure 5 shows the position of the robot's nose when shaking and nodding its head with predefined motion and captured by NDI Optotrak motion capture.

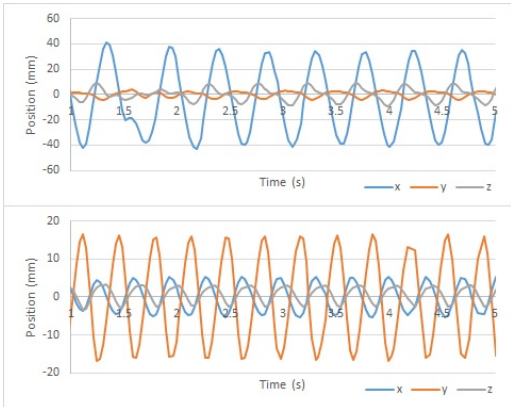

Figure 5: *Robot's nose position when shaking (top) and nodding (bottom).*

3.3. Packet Handling

In the telepresence system, the robot needs to move in real time. To handle network uncertainty, we store the target pose of the robot temporarily in the buffer. Target pose consists of the desired length for all strings which is calculated from the desired angle position. The buffer allows n_{pose} number of poses to be stored with Δ_{pose} range time between poses. The buffer is continuously updated every time the new pose arrived (t_{rcv}), and also send the pose to the robot at t_{send}. An update rule is used to handle unwanted cases caused by network uncertainty, which is described in Algorithm 1.

The more number of n_{pose}, the longer latency will be, but too

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

Algorithm 1 Pose Handling

Everytime receiving P_{rcv} pose at time k
if $0 < t_{rcv} - t_{send} < \Delta_{pose}$ **then**
  $P_k \leftarrow P_k$
  $P_{k+1} \leftarrow P_{k-1}$
  $P_{k+2} \leftarrow P_{rcv}$
else if $\Delta_{pose} \leq t_{rcv} - t_{send} < 2\Delta_{pose}$ **then**
  $P_k \leftarrow P_{k-1}$
  $P_{k+1} \leftarrow P_{rcv}$
else if $-\Delta_{pose} < t_{rcv} - t_{send} \leq 0$ **then**
  $P_k \leftarrow P_k$
  $P_{k-1} \leftarrow P_{rcv}$
else
  $P_k \leftarrow P_{rcv}$

small number of n_{pose} will give a high chance of the lack of data. As mentioned before, between two poses, we perform interpolation to create several additional poses to make the robot move smoothly. Bigger Δ_{pose} will optimize this function and give smoother movement of the robot. However, it will add the latency and risk of missing high-frequency movement.

4. Remote User Interface

On the remote side, we use the wearable devices from ROHM Semiconductor, SensorMedal-EVK-001, to capture user's motion. The sensor consists of an accelerometer, gyroscope, magnetometer, and pressure sensor. We use gyroscope data from the device to get the angular velocity of the remote user's movement because we believe the envelope response describe in [CT99] is best captured in the form of velocity rather than the position like commonly done by other research [SYK*08] [KBB14]. The gyroscope has 16bit resolution with a frequency of 50Hz. There are three devices placed in remote user, one for each arm in the form of watches, and one on the top of the head embedded to the headphone. The sensor in the arm supposedly capture the beat gesture of the remote user when talking, and the sensor in the headphone is to capture the head movement like nodding or shaking. The remote user also uses the headphone for voice communication during the teleconference.

We place Tobi Eye in the remote user's PC to capture his/her eye movement. Remote user will see the real time video of the conference room, and we capture the looking point of the remote user and transform it to the robot's head rotation, so the robot will also look at the point where the remote user looks at the image. Figure 6 shows the remote user when doing the teleconference.

We use the available video chat application to transmit the audio (both side) and video (from local to remote). On the top of that, we create transparent window program made with C# script Unity to capture user gaze using tobi eyeX. The remote user can place and remove marker in this program by clicking the mouse. The marker will mark the focus of participant in the conference room. When the user's gaze is within the marker, then the robot will look at the corresponding direction in the conference room. This method is simple but highly reduce the instability of robot head movement while maintaining good eye contact with the participant in the conference room. Another program is also running in the background

Figure 6: *Remote user.*

to read the data from the motion sensor and send it to local PC. Figure 7 shows the display interface in the remote user side.

Figure 7: *Remote PC software.*

5. Local User Interface

As explained in Section 2, in the conference room we prepare the avatar robot, 360° camera, the microphone and speaker to communicate with the remote user. The configuration can be seen in Figure 8.

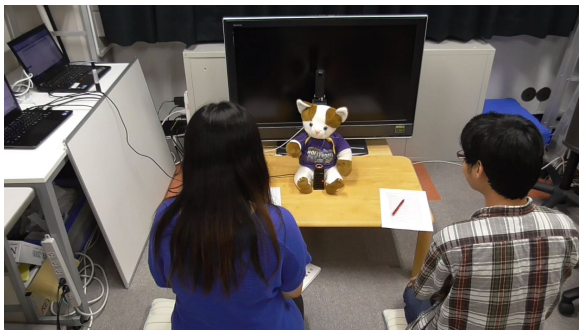

Figure 8: *Configuration in conference room (local).*

The sensor data from the remote side will be processed in local PC. We create a program made with C# script Unity to process the

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

sensor data, which is consist of the angular velocity of the head and arm movement and position of eye gaze, to desired angular position of the robot's arm and head. The angular velocity will be transformed to angular position θ and ϕ which is shown in Figure 9. Since we use the velocity data, the position error will be accumulated. To avoid this kind of problem, we always return to the neutral position when no angular velocity is detected. This approach is also suitable to mimic beat gesture of human when talking.

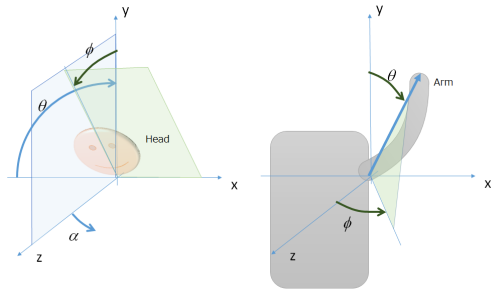

Figure 9: *Robot's angular coordinate for head (left) and arm (right).*

The position of eye gaze will be transformed into the turning angle of the robot's head α. This can be calculated by scaling the image from real-time video to the turning angle of the robot. After getting the angular position, the length of each string is calculated to achieve that desired position. We call this combination of string length as the pose. This pose is then stored in the pose buffer and wait to be executed by the robot.

6. Result and Discussion

We conduct the experiment to evaluate our teleconference system, especially in the local side. The participants in our experiment consist of 10 people (six men and four women) with age range from 23 to 35 years old. For each experiment, two participants will be in the conference room, and the remote user is played by our operator.

The experiment consists of two sessions. For the first session, the teleconference will be conducted using conventional video conference. As for the second session, we use the stuffed-toy robot to represent the remote user. Thus, the participant will have the comparison and easier to tell the difference. For each session, the remote user will ask several questions to the participant in turn. The turn and the question are randomly selected by the remote user.

We observed the participants response when the remote user asks or talk to one of them without mentioning their name throughout the experiment. For each session in an experiment, we perform this scenario six times, so the total is 30 times for each video conference and avatar robot. We found that when using the video conference, this scenario caused confusion on the participant side 15 times (50%), whereas only three times (10%) this confusion happens when we use avatar robot. To break this confusion, remote user and participants need to do repetitive confirmation which reduces communication smoothness.

After the experiment, participants are asked to fill the questionnaire to get their impression of doing the teleconference both with the video conference and the avatar robot. We asked their impression about the different aspect of the experience. They were asked to scale their agreement with the following statements using visual analog scale range from -5 (disagree) to 5 (agree).

S1. I easily understand who the remote user is talking to
S2. The turn taking of speaking during the conversation was easy for me
S3. The gesture of the remote user/robot is easy to understand
S4. I know when the remote user/robot is agree by gesture
S5. I know when the remote user/robot is disagree by gesture
S6. I think the remote user/robot movement is natural
S7. I can feel eye gaze of the remote user/robot
S8. I can feel eye contact with the remote user/robot
S9. I can feel existence of the remote user

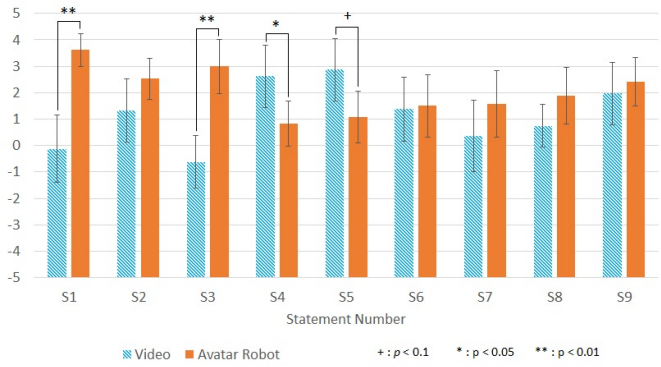

Figure 10: *Questionnaire result from teleconference using video conference versus stuffed-toy robot.*

The result of participant's impression can be seen in Figure 10. We use t-test to analyze the result from questionnaires collected in this experiment.

Statement 1 and 2 which is related to taking the turn in speaking receive very good impression when using stuffed-toy robot. The significant difference is found in the easiness of distinguishing who the remote user is talking to by using the robot compare to video conference (p-value= 0.00044). This result supports the observation we mentioned earlier where avatar robot help reduces the confusion in participant side about who the remote user is asking to. The ability of the robot to turn its head while speaking to the corresponding person make it easier to communicate in a conference situation, where two or more people are involved. This experience is quite difficult to perform in a conventional video conference because participant can't easily tell who the remote user is talking to. The video of the remote user always shows his/her eyes staring at the display monitor, making it tough to distinguish where the remote user are looking at, even though the remote user actually looks to a particular person in his/her display monitor. Repetitive confirmation or mentioning participant name everytime is needed to break this confusion, which is not something people naturally do in direct conversation.

Furthermore, the beat gesture of the robot (S3) also shows positive impression quite significantly (p-value= 0.00217). We believe capturing the remote user gesture in the form of velocity play an

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

important role in this impression. When we have a direct conversation, we tend to produce beat gesture along with other gesture. Beat gesture happens outside the center of the listener's gaze and will be caught by peripheral vision, which is more sensitive to velocity [MN84]. In the case of the video conference, even though the upper part body (head and part of arms) of the remote user is visible to the participants (captured from embedded camera on the PC), the participants tend to really focus on the face. Capturing the whole body of the remote user may transfer better beat gesture through video, but this will need additional camera configuration which is not convenient to be set up in the remote user side.

The head nodding and shaking expression corresponding to agree and disagree (S4 and S5) are better perceived from the video conference, although the impression from the robot is still positive. This is partly due to the limitation of the robot head movement as well as the priority to make the robot movement smoother. Robot movement perceived by the participant as a natural movement (S6), roughly the same with the video conference with the real human.

The smoothness movement of the robot is easily affected by the network. The uncertainty in wireless communication can result in jaggy motion in the robot. One way to handle this behavior is to extend the period between poses of the robot Δ_{pose}, as we mention in Section 3.3, so the robot will have more time to interpolate the pose. However, it can cause more delay in the robot response and possibly missing high-frequency movement like nodding or shaking head. Creating recognizer for high-frequency movement can handle this kind of problem, however, we will loose the variation of the nodding (or another kind of gesture) because it will be interpreted as the same movement.

Eye gazing and eye contact when using the robot show positive impression (S7 and S8). It is better than video conference but not really significant. This behavior is also very important to experience natural conversation. Lastly, about the feeling of the existence of the remote user (S9), the participants have a positive impression both when using video conference and avatar robot.

7. Conclusion and Future Work

We have developed a telepresence system that enables natural interaction between the remote user and local user. We use stuffed-toy robot as the avatar robot that can increase the interaction in the local environment and able to perform some envelope feedbacks which is important in creating natural interaction. The wide angle camera is used in the conference room to capture local environment and send it to the remote user by using available video chat application. As for the remote user, wearable devices is used to capture his/her movement naturally.

The experiment shows that our system can enhance teleconference experience in some aspects. The turn-taking, specifically, improved a lot when using this system compare to conventional video conference. Furthermore, beat gesture also perceived better than using the video conference. As for the other aspect, using avatar robot is still showing positive impression even though not significantly different with the video conference.

In the future, we would like to improve the richness of the robot movement so it can show better life-like movement. A better way to deal with network problem also needs to be considered.

8. Acknowledgement

This research is (partially) supported by the Center of Innovation Program from Japan Science and Technology Agency, JST.

References

[BES15] BAILLY G., ELISEI F., SAUZE M.: Beaming the gaze of a humanoid robot. In *Proc. Int. Conf. on Human-robot Interaction (HRI)* (2015). doi:10.1145/2701973.2701992. 1

[CT99] CASSELL J., THORISSON K. R.: The power of a nod and a glance: Envelope vs emotional feedback in animated conversational agents. *Applied Artificial Intelligence Vol.13*, 4 (1999), 519–538. 1, 2, 4

[CYL16] CHING P. W., YUE W. C., LEE G. S. G.: Design and development of EDGAR – a telepresence humanoid for robot-mediated communication and social applications. In *Proc. Int. Conf. on Control and Robotics Engineering '16* (2016). doi:10.1109/ICCRE.2016.7476147. 1

[GHB09] GOUAILLIER D., HUGEL V., BLAZEVIC P.: Mechatronic design of nao humanoid. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (2009), pp. 769–774. doi:10.1109/ROBOT.2009.5152516. 3

[JSG*15] JEONG S., SANTOS K. D., GRACA S., O'CONNELL B., ANDERSON L., STENQUIST N., FITZPATRICK K., GOODENOUGH H., LOGAN D., WEINSTOCK P., BREAZEAL C.: Designing a socially assistive robot for pediatric care. In *Proc. Int. Conf. on Interaction Design and Children (IDC) '15* (2015), pp. 387–390. doi:10.1145/2771839.2771923. 1

[KBB14] KOENEMANN J., BURGET F., BENNEWITZ M.: Real-time imitation of human whole-body motions by humanoids. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (2014), pp. 2806–2812. doi:10.1109/ICRA.2014.6907261. 4

[KCL13] KRISTOFFERSSON A., CORADESCHI S., , LOUTFI A.: A review of mobile robotic telepresence. In *Advances in Human-Computer Interaction* (2013). doi:10.1155/2013/902316. 1

[KWK10] KANTONEN T., WOODWARD C., KATZ N.: Mixed reality in virtual world teleconferencing. In *Virtual Reality Conference* (2010). doi:10.1109/VR.2010.5444792. 1

[LTSB08] LEE J. K., TOSCANO R. L., STIEHL W. D., BREAZEAL C.: The design of a semi-autonomous robot avatar for family communication and education. In *Proc. Int. Symp. on Robot and Human Interactive Communication '08* (2008), pp. 166–173. doi:10.1109/ROMAN.2008.4600661. 1

[MN84] MCKEE S. P., NAKAYAMA K.: The detection of motion in the peripheral visual field. *Vision Research Vol. 24*, 1 (1984), 25–32. doi:10.1016/0042-6989(84)90140-8. 6

[SYK*08] SULEIMAN W., YOSHIDA E., KANEHIRO F., LAUMOND J.-P., MONIN A.: On human motion imitation by humanoid robot. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (2008), pp. 2697–2704. doi:10.1109/ROBOT.2008.4543619. 4

[TDYU11] TSUI K. M., DESAI M., YANCO H. A., UHLIK C.: Exploring use cases for telepresence robots. In *Proc. Int. Conf. on Human-robot Interaction (HRI)* (2011), pp. 11–18. doi:10.1145/1957656.1957664. 1

[TMS*07] TSAGARAKIS N. G., METTA G., SANDINI G., VERNON D., BEIRA R., BECCHI F., RIGHETTI L., SANTOS-VICTOR J., IJSPEERT A. J., CARROZZA M. C., CALDWELL D. G.: icub: The design and realization of an open humanoid platform for cognitive and neuroscience research. *Advanced Robotics Vol. 21*, 10 (2007), 1151–1175. 3

[TMYH13] TAKASE Y., MITAKE H., YAMASHITA Y., HASEGAWA S.: Motion generation for the stuffed-toy robot. In *Proc. SICE Annual Conference '13* (2013), pp. 213–217. 3

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.