

A Study of a Retro-Projected Robotic Face and its Effectiveness for Gaze Reading by Humans

Frédéric Delaunay, Joachim de Greeff and Tony Belpaeme
Centre for Robotics and Neural Systems
School of Computing and Maths
University of Plymouth
B110 Portland Square, PL4 8AA Plymouth, United Kingdom
Email: frederic.delaunay@plymouth.ac.uk

Abstract—Reading gaze direction is important in human-robot interactions as it supports, among others, joint attention and non-linguistic interaction. While most previous work focuses on implementing gaze direction reading on the robot, little is known about how the human partner in a human-robot interaction is able to read gaze direction from a robot. The purpose of this paper is twofold: (1) to introduce a new technology to implement robotic face using retro-projected animated faces and (2) to test how well this technology supports gaze reading by humans. We briefly discuss the robot design and discuss parameters influencing the ability to read gaze direction. We present an experiment assessing the user’s ability to read gaze direction for a selection of different robotic face designs, using an actual human face as baseline. Results indicate that it is hard to recreate human-human interaction performance. If the robot face is implemented as a semi sphere, performance is worst. While robot faces having a human-like physiognomy and, perhaps surprisingly, video projected on a flat screen perform equally well and seem to suggest that these are the good candidates to implement joint attention in HRI.

Index Terms—human-robot interaction; eye gaze; joint attention; robotic face

I. GAZE DETECTION IN HUMAN-ROBOT INTERACTION

Humans are exceptionally good at inferring where others are looking. This ability highly facilitates the establishment of joint attention, deemed to be very important for a wide variety of interaction schemes, both between human-human [1], [2], human-robot and robot-robot [3], [4] interaction. In human-robot interaction, the detection of gaze direction can be considered from both perspectives: the robot reading gaze direction of the human partner and the human partner reading the robot’s gaze. The former has received most attention from the HRI field, while the later will be the topic of this paper.

The reason for having a robot read the gaze of a human partner is important for social learning, including the acquisition of language and conceptual structures. This has been acknowledged in the HRI field for quite some time and several studies have proposed algorithms for gaze direction detection, both in humans and other robots [5], [6], [7] (see [8] for a survey of eye and gaze detection). Another reason is that appropriate eye gaze behaviour facilitates interaction: Yoshikawa et al. [9] for instance describe how a responsive robotic gazing system increases the feeling of people of being looked at, thus enhancing the interaction experience. Related to this, in [10] and [11] a bidirectional eye contact method

was described that facilitates the communication between a robot and a human. Picot et al. [12] describe a virtual agent displayed on a flat screen monitor was able to interpret scenes and direct its gaze in a lifelike manner.

However, the ability to detect the direction of an agent’s gaze needs to be present for *both* interacting partners, and hence it is not only important that the robot can interpret human gaze, but also that a human can perceive where a robotic partner is looking. This is of significant interest in developmental robotics where the robot-human dyad supports mental development [13], [14]. In young children, for example, cyclical changes in gaze to and from the adult serves as a signal function of the infant’s affect, which in turn modulates the adult’s behaviour towards the infant [1]. Cognitive psychology shows how gaze direction reading is essential in joint visual attention [2] or how object permanency can be read from the gaze being fixed on the expected location of an occluded object. In adults, gaze is a powerful signal; gaze aversion, for example, is used to signal that one is considering a cogitating [15].

This suggests that it is not only important to have machines that can read gaze direction, but that machines should be able to accurately display gazing behaviour that can be correctly interpreted by human users. To the best of our knowledge, there are no studies that address this issue. There are several open questions: What factors influence the ability of people to infer where another (artificial) agent is looking? What is the influence of the physiognomy of an agent’s face and eyes on the user’s ability to infer where it is looking? And do the dynamics of eye movements have an influence? Answers to these questions would enable us to design robotic faces in a more principled way, rather than the intuitive approach currently employed. In relation to this, we have developed a new robotic face technology based on retro-projected animated faces (see next section) which we would like to evaluate against a selection of alternative robotic faces.

II. A RETRO-PROJECT ROBOTIC FACE

The robotic face we wish to study and contrast against other robotic faces is based on retro-projected animated face (RAF) technology (more technical details in [16]). In this, a live-generated video is retro-projected into a semi-transparent mask (see figure 1). This technology has a number of advantages

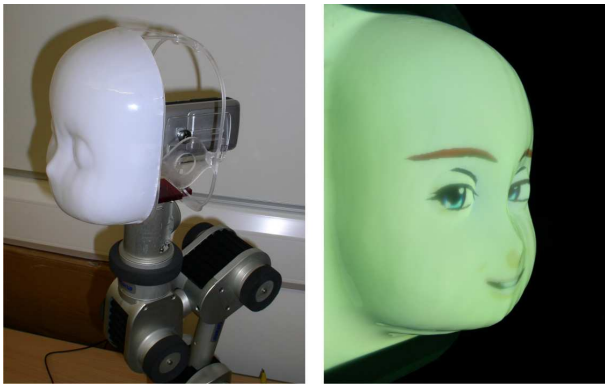


Fig. 1. a prototype of a retro-projected robotic face, (l) the construction of the face showing the semi-transparent mask, the microprojector, support structure and robot arm on which the face is mounted, (r) the face as experienced by users.

over alternative face technologies, specifically animatronic and android technologies. The most important being the flexibility and adaptivity of the face: it can display a wide range of facial expressions, it can change gender or age on the fly and can make use of cues, such as colour or animations, to communicate internal states to the user. The animation speed is unrivalled due to the lack of mechanical parts and this also results in the absence of mechanical motor noise. The RAF technology has very low maintenance requirements due to the lack of mechanical parts and the use of LED projector technology which has a mean time between failure of over 50,000 hours. The weight is relatively low: our lightest prototype weighs about 500gr.

There are drawbacks as well, the most significant ones being that non-linear behaviour of skin needs to be programmed, something which comes “free” with android faces through non-linear interactions and shadows being inherently present in the latex mask. The RAF technology requires a relatively darkened room depending on the power of the projector, however this is expected to improve in the near future with the introduction of laser instead of LED-based microprojectors. Finally, as the projector needs a clear projection volume, there can be no sensors or actuators behind the face. These necessarily need to be placed at the periphery of the face or behind the projector.

III. EXPLORED ROBOTIC FACE TECHNOLOGIES

In the study we performed a series of experiments in which human participants were asked to infer the gaze direction of four different types of facial interface (see figure 2): (1) a real human face, (2) a human face displayed on a flat-screen monitor, (3) an animated face projected on a semi-sphere and (4) and an animated face projected on the 3D mask described in the previous section.

The rationale behind the different face types is as follows: the real human face will serve as the base line, we assume that a real human face will work best for assessing gaze direction. Next to this, we evaluate three artificial faces. The

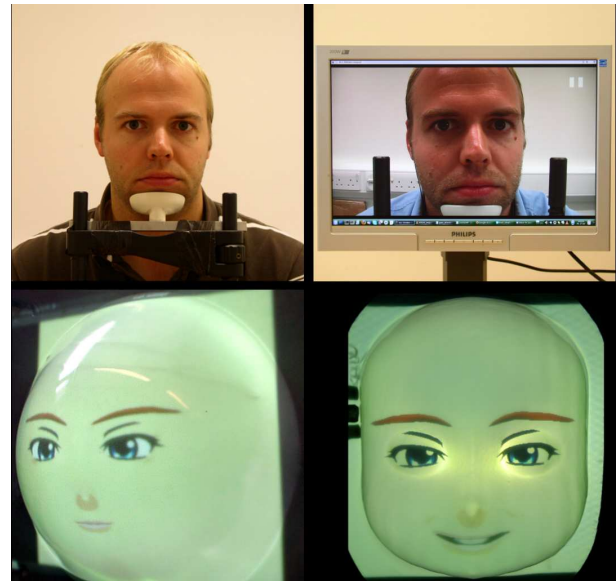


Fig. 2. The four faces used in the experiment.

first being a recording of the human face displayed on a monitor, this condition serves to assess in how far the lack of 3D structure influences gaze reading. The second is a robotic face implemented as a back-projected 3D face, as described in II. The third is the same robotic face, this time projected into a semisphere; this serves to evaluate the technology of [17], who presented a similar robotic setup. However we did not evaluate an android robot face, such as the Albert Hubo or the Ishiguro’s androids due to budgetary constraints.

Human eyes are unique: no other other animal—including primates and apes— have such a large visible sclera to iris ratio [18]. Reading gaze direction is also facilitated by the spherical shape of the eye, which allows one to infer the position of the iris not only when facing a person head on, but also when seeing someone from viewpoints other than frontal. The information we glean from the spherical shape of the eyes is distorted when a face is displayed on a 2D surface (see figure 3). However, under certain the conditions the distortion is minimal: if a video of a person is shown on a screen with the viewer sitting at the relative location where the camera was, the distortion is minimal. The reason for this is that the 3D to 2D transformation is consistent: we have no problems reading gaze direction from the video because we are *aligned* with the camera’s line of sight. However, by moving away from this ideal position, the visual transformation is no longer relevant to our perspective and other cues are needed to read gaze direction. The Mona Lisa effect [19] is related to this, making us believe a person in a portrait is looking at us, even if moving away from the normal of the picture; this however is probably not useful in reading gaze direction from 2D images. In real environments, we are clearly free of this effect as more geometric information is available, and as such multiple observers (at different viewpoints) are able to read a person’s gaze, precision being related to the distance between

the observer and the participant. In scenarios involving more than one person in a robot's field of view, establishing eye-contact should be much faster with gazing than actual head movements.

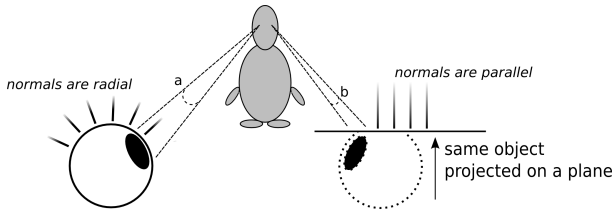


Fig. 3. The iris can be represented as a plane of which the normal is aligned with the eye line of sight. (a) The eye's gaze direction is available to an observer, because there is always a normal vector aligned with the observer's line of sight. (b) By projecting an object onto a plane, distances, angles and normals are lost, and other visual cues are needed to read gaze direction.

The study presented in this paper fits within the larger context of studying human-robot interaction for developmental learning. Specifically, we wish to explore how human-robot interaction can support the acquisition of concepts and associated linguistic utterances. In this, joint attention will be essential in establishing a tutelage relationship between the human caretaker and the robot.

IV. METHODS

In the experiment we wish to study how good people are at reading gaze direction. We have four faces¹ and two viewpoints from which a face can be seen, giving a eight combinations. The viewpoint was either directly facing the face (0 degrees condition) or at an angle of 45 degrees on the right (45 degrees condition). Each participant was asked to evaluate four combinations out the eight possible combinations, so as not to make the experiment too tedious.

See figure 4 for a schematic representation of the setup using the mask as a face being tested and figure 5 for a schematic top view of the positions of the participants with respect to the number grid and face. Such a setup also allowed us to speed up the procedure by testing two participants at the same time. The participant's seat height was adjusted so that the participant's line of sight, the center of grid and the eyes of the face align.

To sum up what is described above, we end up 8 different conditions:

- Human face seen at 0 degrees.
- Human face seen at 45 degrees.
- Human face displayed on a flat screen monitor seen at 0 degrees.
- Human face displayed on a flat-screen monitor seen at 45 degrees
- Animated face projected on a semi-sphere seen at 0 degrees

¹From here on we distinguish between a display (the physical surface) and a face (set of facial features) only when needed. Otherwise, we use the term face to express any combination of both (only a real human face is a display and a face at the same time).

- Animated face projected on a semi-sphere seen at 45 degrees
- Animated face projected on a 3D mask seen at 0 degrees
- Animated face projected on a 3D mask seen at 45 degrees

Twenty four participants participated in a sequence of four sessions (combinations). This yielded 96 records, which gives 12 records for each condition.

To account for any habituation effect, for example, performance increase of participants over the sessions of a sequence, we shuffled the order of sessions for each pair of participants. This way we ensure that the order in which a face is presented to a participant is varried over participants.

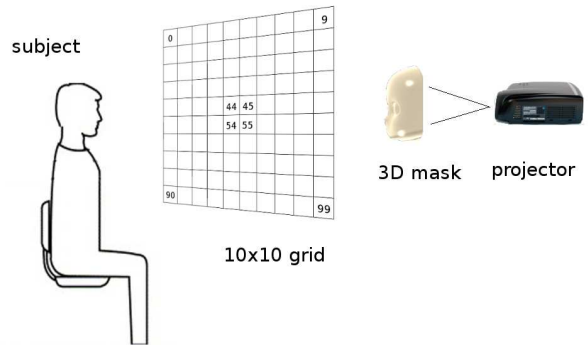


Fig. 4. An instance of an experimental session, here with the retro-projected animated face.

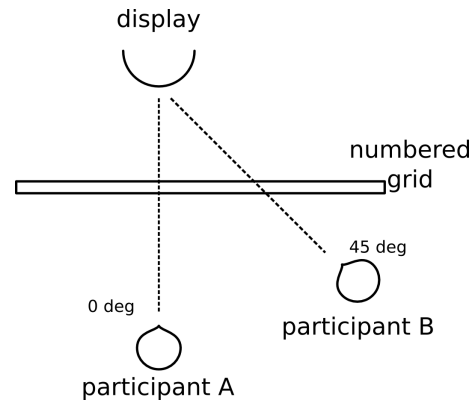


Fig. 5. Top down schematic representation of the position of the two participants with respect to the number grid and face.

Between the participants and the face there is a transparent grid of 50x50 cm, which is divided in 100 squares (10 rows and 10 columns) displaying the numbers 0 to 99 from top left to bottom right, so the center of the grid is between the numbers 44, 45, 54 and 55. We opted for a grid standing upright between the participants and the face so that the distance from eyes of the face to the numbers of the grid would increase evenly from the center of the grid; this would not be the case if the numbered grid was laid flat in front of the face. The position and size of the grid also ensured

downward facing eyelids could not hinder the interpretation of gaze direction when gazing at the bottom of the grid.

A single session consisted of the face looking at a sequence of 50 randomly generated numbers, switching to the next one after a fixed delay of 5 seconds. As numbers are pseudo-randomly generated, we instructed the participants that the same number can appear multiple times in a number sequence.

Once a number was gazed at, an auditory signal was given indicating to the participants that they could perform their observation. A delay of 5 seconds was long enough to give the (human) face enough time to find the proper number and for the participants to write down their observations afterward. When the face was a human (one of the examiners), the number sequence was played over earphones worn by the examiner so it could not be heard by the participants. In the case of the video, the face consisted of a prerecorded sequence of the same examiner looking at a number sequence. In the two cases of the animated faces, the number sequence was generated on the fly and fed into the animated face control module. The same auditory signal was played once the face was looking at the next number to ensure consistency among sessions.

The participants were asked to write down the number they thought the face was gazing at on a paper sheet. Handwriting allows participants to quietly report their results in a very natural way, and they can easily make corrections if needed. Participants were also asked to perform as best as they could with and not to cheat by looking at each others notepads. By observing the participants while they performed the experiment it was acknowledged that they were obeying these rules.

V. RESULTS

The sequence of numbers written by each pair of participants was compared to the actual sequence and the difference was calculated using the euclidean distance between the cell on the grid the participant reported and the cell to robot gazed at. In this way, the difference between a participant's sequence and the real sequence is expressed as a mean error distance.

As can be seen in figure 6, performance for the human face was best (lowest mean error): when having to guess at which number the human face was looking, the participants had an average error of 1.13. As the numbers were 5cm apart, this means that participants were about 5.65cm (5cm x 1.13) off in guessing where the gaze rested. The mask and flat faces are next in accuracy for guessing where the eyes are looking, followed by the dome face.

A 4 x 2 analysis of variance (ANOVA) on gaze interpretation error showed main effects of both display type, $F(3, 88) = 8.121$, $p < .01$, and looking angle, $F(1, 88) = 14.438$, $p < .01$. However, no interaction effects were observed, $F(3, 88) = 0.419$, $p = .740$.

Post-hoc comparison of the ANOVA using Tukey test shows that the participants' performance between the human condition and all other conditions was significant, while this was not the case for any other comparison (see table I).

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT DISPLAY TYPES

condition	versus	<i>p</i>
dome	flat	.176
	human	.000
	mask	.146
flat	dome	0.176
	human	.027
	mask	1.000
human	dome	.000
	flat	.027
	mask	.035
mask	dome	.146
	flat	1.000
	human	.035

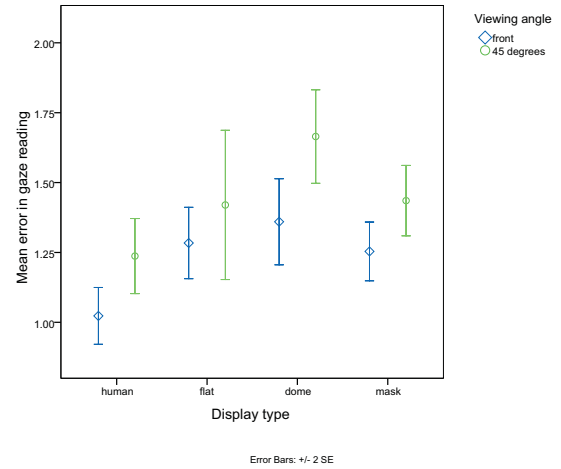


Fig. 6. Display of mean errors per display from 2 angles (front and 45 degree). Error bars indicate Standard Error.

When examining the difference in performance between the two different viewpoints, it is clear that it is much easier for participants to determine the gaze direction when they are facing the face, as opposed to a side view at 45 degrees (figure 6). This difference between viewpoints was significant for the human, mask and dome, but not for the flat screen, due to the large variance in performance (see table II).

Comparing the performance of the participants at gaze reading from a 45° angle using ANOVA gave $F(3, 44) = 3.690$, $p < 0.019$. A post-hoc Tukey test showed there was no performance difference between the human face, flat screen and mask, but there was a significant difference in participant performance between the dome and the human face ($p < 0.01$).

TABLE II
DIFFERENCE TESTS FOR THE FOUR DIFFERENT DISPLAYS BETWEEN THE TWO DIFFERENT ANGLES. THE DIFFERENCE BETWEEN VIEWING ANGLES FOR DOME, MASK AND HUMAN IS SIGNIFICANT, AND FOR FLAT IT IS NOT.

t-test	dome	flat	mask	human
<i>p</i> value	.014	.367	.037	.018

After performing the experiments, participants were asked

to subjectively rate their experience. We asked them to describe how effective they found each of the four different faces at conveying information about gaze direction. This was rated on a seven-point Likert scale, with the range: 1-very ineffective, 2-ineffective, 3-somewhat ineffective, 4-undecided, 5-somewhat effective, 6-effective, 7-very effective. The resulting graph is displayed in figure 8 which shows that participants find the human the most effective in terms of gaze information, followed by mask, flat and dome. The difference between human and all other faces is significant, as is the difference between mask and dome. The difference between flat and mask and flat and dome is not (see table III).

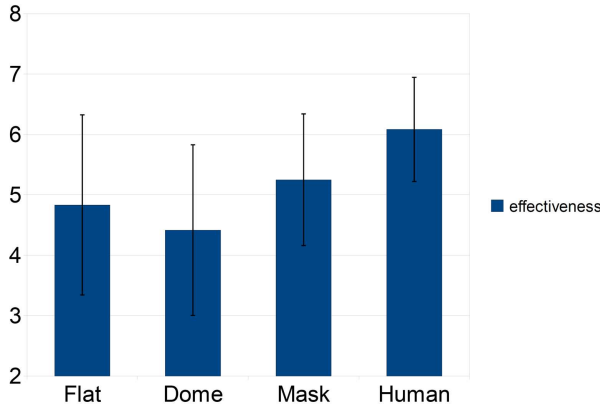


Fig. 7. Participants rating of different faces in terms of effectiveness in conveying gaze information. Error bars indicate Standard Deviation.

TABLE III
SIGNIFICANCE DIFFERENCE TESTS OF THE SUBJECTIVE EFFECTIVENESS SCORE FOR THE 4 DIFFERENT FACES AGAINST EACH OTHER.

	flat	dome	mask	human
t-test/flat	-	$p=.336$	$p=.285$	$p=.001$
t-test/dome		-	$p=.030$	$p=.000$
t-test/mask			-	$p=.006$
t-test/human				-

VI. DISCUSSION

As expected, inferring gaze direction from a real human is easiest and most accurate. Overall though, it can be concluded that a 3D mask with a projected animated face embodies a reasonably setup for which participants are still rather apt at inferring the gaze direction. We hypothesize that although an animated face is missing some human characteristics, and hence this may impair the ability of participants to infer it's gazing direction, the 3D structure of the mask counters this effect. This is reflected in the fact that performance for the dome is significantly lower. A clear advantage of a projected 3D face is its online nature, with eye animations being rendered on the fly which allow for a very direct form of interaction, as opposed to prerecorded sequences. This feature was not some much utilised in the reported experiment, but will be invaluable for planned HRI studies based on this setup.

Comparing the mask and the flat-screen video, participants perform more or less equally well (difference in performance is not significant). A flat-screen video of a human face is also relatively well interpreted, although especially seen from the side the variance in performance is rather large. Based on verbal reports from participants, we hypothesise that these differences are influenced by how much the participants are trying to actually see where the flat face is looking at versus how much they try to reason where it is looking at. From the side, this face is looking at the corners of the screen when it is supposed to look at a number in the grid corner. It seems the gaze is at some vague distant point in space, rather than at the number grid. However, participants may reason that this is the case, because they are viewing a flat-screen, and hence adjust there gaze estimation accordingly. This again shows that human use a number of different cues to read gaze direction [2].

Another factor that most likely influenced the performance of participants when confronted with a human face (both real and on video), was the fact that participants reported to find it helpful to see the eyes of the human face employing recognisable search strategies when looking for the next number. For instance, when switching from number 12 to 86, typical human search behaviour would be to drop the eyes first from the 2nd line (20-29) to the 8th line (80-89), and then move along the horizontal axis from 82 to 86. Occasionally participants reported to recognize this behaviour to be helpful. It may be the case that other participants did not observe this, but were nevertheless employing this information subconsciously. In contrast, the projected animated face (being computer controlled), would drop it's gaze directly from one number to the next. This issue could be addressed in a follow up study by having the animated face mimicking human like search behaviour.

ACKNOWLEDGEMENT

This work is supported by the CONCEPT project (EPSRC EP/G008353/1), the EU FP7 ITALK (no. 214668) and ALIZE-E (no. 248116) projects.

REFERENCES

- [1] M. DeBoer and A. M. Boxer, "Signal functions of infant facial expression and gaze direction during mother-infant face-to-face play," *Child Development*, vol. 50, no. 4, pp. 1215–1218, 1979.
- [2] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? Cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50 – 59, 2000.
- [3] F. Kaplan and V. Hafner, "The challenges of joint attention", *Interaction Studies*, pp. 67–74, 2004. [Online]. Available: <http://cogprints.org/4067/>
- [4] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Advanced Robotics*, vol. 20, pp. 1165–1181(17), October 2006.
- [5] R. Atienza and A. Zelinsky, "Active gaze tracking for human-robot interaction," *Multimodal Interfaces, IEEE International Conference on Multimodal Interfaces*, p. 261, 2002.
- [6] D. H. Yoo and M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 25 – 51, 2005, special Issue on Eye Detection and Tracking. [Online]. Available: <http://www.sciencedirect.com/science/article/B6WCX-4DHXF2H-1/2/1ebc9bb2e46a1aec88d6cd91cc431c6f>

- [7] J. Ruiz-Del-Solar and P. Loncomilla, "Robot head pose detection and gaze direction determination using local invariant features," *Advanced Robotics*, vol. 23, no. 3, pp. 305–328, 2000. [Online]. Available: <http://dx.doi.org/10.1163/156855308X397497>
- [8] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 5555.
- [9] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto, "Responsive robot gaze to interaction partner," in *In Proceedings of robotics: Science and systems*, 2006.
- [10] D. Miyauchi, A. Nakamura, and Y. Kuno, "Bidirectional eye contact for human-robot communication," *IEICE - Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2509–2516, 2005.
- [11] D. Miyauchi, A. Sakurai, A. Nakamura, and Y. Kuno, "Active eye contact for human-robot communication," in *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, 2004, pp. 1099–1102.
- [12] A. Picot, G. Bailly, F. Elisei, and S. Raidt, "Scrutinizing natural scenes: Controlling the gaze of an embodied conversational agent," in *Proceedings of Intelligent Virtual Agents, 7th International Conference, IVA 2007*, Paris, France, 2007, pp. 272–282.
- [13] T. Kishimoto, Y. Shizawaa, J. Yasudaa, T. Hinobayashia, and T. Minamia, "Gaze following among toddlers," *Infant Behavior and Development*, vol. 31, pp. 280–286, 2008.
- [14] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 675–691, 2005.
- [15] A. McCarthy, K. Lee, S. Itakura, and D. W. Muir, "Cultural display rules drive eye gaze during thinking," *Journal of Cross-Cultural Psychology*, vol. 37, no. 6, pp. 717–722, 2006.
- [16] F. Delaunay, J. de Greeff, and T. Belpaeme, "Towards retro-projected robot faces: an alternative to mechatronic and android faces," in *Proceedings of the IEEE Ro-Man 2009 conference, Toyama, Japan*. IEEE, 2009.
- [17] M. Hashimoto and H. Kondo, "Effect of emotional expression to gaze guidance using a face robot," in *Proceedings of the 17th IEEE International Symposium on Robot Human Interactive Communication (RoMan 2008)*, Y. Tamatsu, Ed., 2007, p. 95101.
- [18] H. Kobayashi and S. Kohshima, "Unique morphology of the human eye," *Nature*, vol. 387, no. 6635, pp. 767–768, 1997.
- [19] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.