Role of Embodiment and Presence in Human Perception of Robots' Facial Cues

Ali Mollahosseini[a], Hojjat Abdollahi[a], Timothy D. Sweeny[c], Ron Cole[b], Mohammad H. Mahoor[a,*]

[a]*Daniel Felix Ritchie School of Engineering & Computer Science, University of Denver, Denver, CO, 80208*
[b]*Boulder Learning Inc., Boulder, CO 80301*
[c]*Department of Psychology, University of Denver, Denver, CO, 80208*

Abstract

Both robotic and virtual agents could one day be equipped with social abilities necessary for effective and natural interaction with human beings. Although virtual agents are relatively inexpensive and flexible, they lack the physical embodiment present in robotic agents. Surprisingly, the role of embodiment and physical presence for enriching human-robot-interaction is still unclear. This paper explores how these unique features of robotic agents influence three major elements of human-robot face-to-face communication, namely the perception of visual speech, facial expression, and eye-gaze. We used a quantitative approach to disentangle the role of embodiment from the physical presence of a social robot, called Ryan, with three different agents (robot, telepresent robot, and virtual agent), as well as with an actual human. We used a robot with a retro-projected face for this study, since the same animation from a virtual agent could be projected to this robotic face, thus allowing comparison of the virtual agent's animation behaviors with both telepresent and the physically present robotic agents. The results of our studies indicate that the eye gaze and certain facial expressions are perceived more accurately when the embodied agent is physically present than when it is displayed on a 2D screen either as a telepresent or a virtual agent. Conversely, we find no evidence that either the embodiment or the presence of the robot improves the perception of visual speech, regardless of syntactic or semantic cues. Comparison of our findings with previous studies also indicates that the role of embodiment and presence should not be generalized without considering the limitations of the embodied agents.

Keywords: Social Robot, Embodiment, Physical presence, Retro-Projected Robots

1. Introduction

Social robotics is a rapidly emerging field, which aims to develop robots capable of communicating and interacting with human users in a socio-emotional way (Dautenhahn, 2007; Breazeal, 2005). This is owing to advancements in computer technology, artificial intelligence, and recent innovations in virtual reality and computer graphics. The population of robotic agents including social and humanoid robots made in 2008 was about 8.6 million units (Guizzo, 2010) with a projected annual growth rate of 17% (IDC, 2016). Virtual agents, on the other hand, have received considerable attention in recent years as social agents (e.g. for museum guidance (Kopp et al., 2005), education (Vala et al., 2007), entertainment (Hartholt et al., 2009), and training for job interviews (Hoque et al., 2013)) due to the flexibility of computer rendered faces and the ubiquity of computer screens on mobile devices. Virtual agents are often used when a physical task or interaction such as moving objects is unnecessary. As robotic technologies are focusing more on improving social interaction with users, determining which kinds of robots or virtual agents are best suited for social interaction becomes increasingly important. One fundamental research question is what would be the difference between virtual agents and robots in terms of human interaction, particularly in perceiving major elements of face-to-face communication (both verbal and non-verbal facial cues and skills).

The most salient difference between a robot and a virtual agent on a computer screen is physical embodiment. Several investigations have compared various elements of social interaction among robots and virtual agents (Kidd and Breazeal, 2004; Ju and Sirkin, 2010; Fujimura et al., 2010; Delaunay et al., 2010; Al Moubayed et al., 2013; Mollahosseini et al., 2014), and the majority of these investigations suggested that the physicality of the robot benefits user interaction. However, in the majority of these experiments, a robot with physical embodiment was physically present in front of the subjects. This is potentially problematic since the subject's percepts and evaluations may be affected not only by the robot's embodiment but also by its presence.

Some researchers evaluated the role of presence by comparing a robotic agent with its telepresence or an animated/computer-rendered version of the robot (Kidd and Breazeal, 2004; Lee et al., 2006; Kose-Bagci et al., 2009; Bainbridge et al., 2011). The majority of these investigations suggested that the presence of the robot improves user interaction and social aspects of the robot. However, as shown in Figure 1, few have compared all three conditions in the same experiment/platform. Also, the

*Corresponding author
Email addresses: ali.mollahosseini@du.edu (Ali Mollahosseini), habdolla@du.edu (Hojjat Abdollahi), Timothy.Sweeny@du.edu (Timothy D. Sweeny), rcole@boulderlearning.com (Ron Cole), mmahoor@du.edu (Mohammad H. Mahoor)

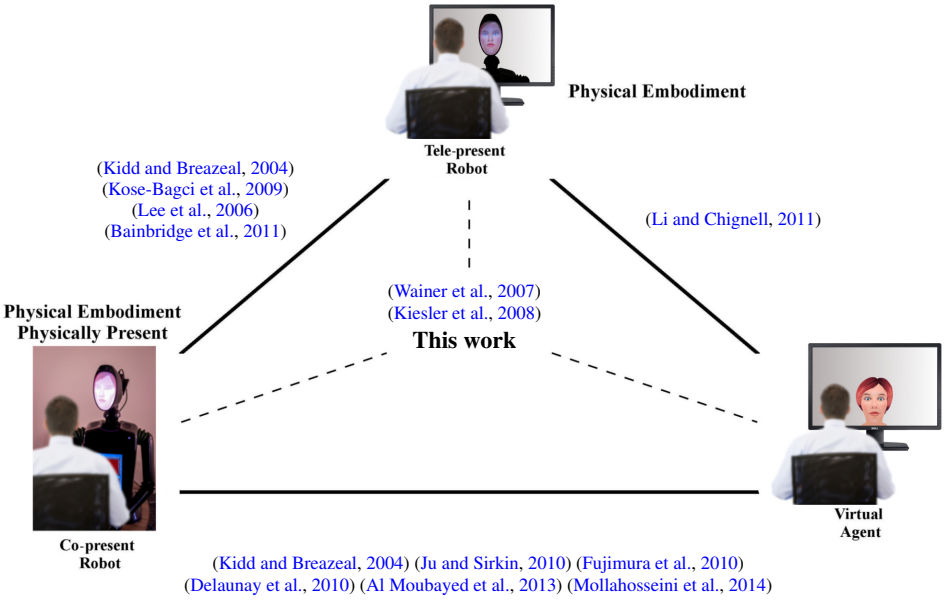


Figure 1: Comparison of presence and embodiment dimensions across three categories of experimental stimuli in the literature (inspired from Li (2015)). The majority of studies do not distinguish the telepresence of a robot (physical embodiment) from the copresence of a robot (physical presence).

majority of these studies compared the influence of these agents on social elements such as likability (Kiesler et al., 2008), enjoyment (Wainer et al., 2007), etc. by requiring subjects to complete a questionnaire after interaction in the lab. Although the reliability of questionnaires can be validated by measurements such as Cronbach's Alpha (Cronbach, 1951), self-report may be an inaccurate quantitative measure, especially with small sample sizes. Hence, better quantitative measures are necessary to determine whether a physically present robotic agent can produce different, and perhaps superior experiences compared to a screen-based version of the same robot.

Recently, retro-projected robotic heads have received much attention (Al Moubayed et al., 2013; Mollahosseini et al., 2014). Retro-projected robotic heads harness character-animation technologies to create an animated human face (aka avatar) and then project this avatar onto a face-shaped translucent mask. The mask and the projector can then be rigged onto a neck mechanism that can move like a human head. By virtue of the computer graphics used to generate the avatar, highly realistic, accurate, and dynamic animations can be generated. These avatars can range from cartoon-like to photo-realistic faces and are usually able to show natural visual speech and facial expressions.

This paper studies the role of embodiment and presence in human perception of a retro-projected robot's facial cues. We used a retro-projected robotic head for this study, since the same animation from a virtual agent could be projected to this robotic face, thus allowing comparison of the virtual agent's animation behaviors with both telepresent and physically present robotic agents. Because face-to-face communication is an important method of social interaction which plays a major role in individuals' socialization and experience (Kendon et al., 1975), we focus on three major elements of face-to-face communication—visual speech, facial expression, and eye-gaze. We leverage three different agency conditions (copresent robot, telepresent robot, and virtual agent) to evaluate whether the embodiment and presence of a social robot provides any extra value for discriminating these social cues compared with an on-screen animation. Similar to other robotic platforms, retro-projected robots have some limitations (e.g., the mask is static and the jaw and lip movements are only optical). We consider these limitations in this study.

The remainder of this paper is organized as follows. Section 2 reviews the definition of physical embodiment and presence and then defines research questions of this study. Section 3 introduces the robotic platform used in this study. Sections 4, 5, and 6 study the role of embodiment and presence in perception of a robot's visual speech, facial expression, and eye gaze, respectively. In each of these sections, a brief review of prior work, the algorithm used to generate the facial cues, the experiments and settings, and the results, as well as a discussion of the results and comparison with previous studies are presented. Finally, Section 7 concludes the paper.

2. Embodiment and Presence

Socially Intelligent Agents (SIAs) are systems that are able to connect and interface to humans via the ability to show aspects of human-style social intelligence (Dautenhahn, 1998). These agents can have a wide range of forms, some of which have physical bodies (e.g. a robot) or virtual observable bodies/faces (e.g. an intelligent avatar), and some of which interact with others using only voice or text without having any appearance (e.g. Siri). Since body gesture and expressions play a crucial role in social interactions and communication (e.g., body language, head gesture, facial expressions, speech, etc.),

researchers try to build SIAs that closely mimic the appearance, behavior, and social skills of human beings (Dautenhahn, 2001). The field of "embodied conversational agents" is an excellent example of this approach (Cassell, 2000).

Mimicking the appearance of humans in SIAs or "*tighter coupling of the* [human] *body to the interface*" (Biocca, 1997) is viewed as central for providing the embodiment to the agents. This embodiment can be both virtual (e.g., embodied conversational virtual agents) and physical (e.g., robot). Pfeifer and Scheier (1999) defined the *physical embodiment* in intelligent robots as "a term used to refer to the fact that intelligence cannot merely exist in the form of an abstract algorithm but requires a physical instantiation, a body."

In-line with this definition, much work has examined the role of embodiment with regard to a variety of social interaction elements such as persuasion (Ju and Sirkin, 2010), likeability (Kidd and Breazeal, 2004; Kiesler et al., 2008), enjoyment (Wainer et al., 2007), trustworthiness (Kidd and Breazeal, 2004), helpfulness (Wainer et al., 2007), direct gaze recognition (Ju and Sirkin, 2010), and ease of interaction (Fujimura et al., 2010). The majority of these reports claimed that the physicality of the robot benefited user interaction. However, many of these studies did not distinguish physical embodiment from the copresence of the robot.

Copresence is a sociological concept describing the condition in which human individuals interact with each other (Goffman, 1963; Zhao, 2003). In our case, copresence refers to how the agent is presented to the user. Zhao (2003) defined copresence in two dimensions: 1) the mode of being with others (i.e., physical conditions that structure human interaction), and 2) the sense of being with others (i.e., subjective experience of being with others). The mode of copresence is related to the concept of "distance" in the taxonomy of copresence, which can be physical proximity (within range of the naked senses) or electronic proximity (outside the range of the naked senses but within the range of senses extended through electronic media) (Li, 2015). In real-world environments, physical and digital presence correspond to "copresence" and "telepresence," respectively (Zhao, 2003). The mode of copresence is also similar to the concept of "directness" in the literature (Milgram et al., 1995; Li, 2015). Physical and digital presence can be simply defined as a situation in which the embodied agent can be touched (or can touch the person). In other words, as Milgram et al. (1995) stated: *"[Physical or digital presence:] [the condition] whether primary world objects are viewed directly or by means of some electronic synthesis process."*

The mode of copresence (e.g., physical or digital) can affect a person's sense of copresence or "social presence" (Zhao, 2003). Some researchers evaluated the role of presence by comparing a robotic agent with its telepresence or a video of the robot (Kidd and Breazeal, 2004; Lee et al., 2006; Kose-Bagci et al., 2009; Bainbridge et al., 2011). For example, Bainbridge et al. (2011) studied the role of physical presence in a simple collaborative task with a humanoid robot that was either physically present or displayed via a live video or an augmented video feed. Multiple social interaction aspects such as greetings, cooperation, trust, and personal space were examined in different parts of the task. Participants in the experiment filled out a questionnaire aimed to evaluate different interactive experiences such as general impressions, characteristics of the interactions, etc. The questionnaire data suggested that overall, participants had a more positive interaction with the physically present robot.

In a recent survey (Li, 2015), the effects of physical embodiment and physical presence were explored through a study of 33 experimental works to compare how people interact with 1) physically present robots, 2) telepresent robots, and 3) virtual agents. The study showed that physical presence plays a greater role in determining a person's response to an agent than physical embodiment. The methods used in these studies include post-treatment questionnaires or measuring subjects' behaviors during laboratory experiments. Among these 33 studies, however, few compared all three conditions in the same experiment/platform (See Fig. 1).

2.1. Research Questions

Based on the above and since face-to-face interaction is one of the essential elements of a social system, we have designed three research questions to be addressed in this paper:

- Q1: What is the effect of physical embodiment on perception of agents' facial cues (telepresent robot vs. virtual agent)?
- Q2: What is the effect of physical presence on perception of agents' facial cues (copresent robot vs. telepresent robot)?
- Q3: What is the joint effect of physical embodiment and presence on perception of agents' facial cues (copresent robot vs. virtual agent)?

In order to answer these research questions, we studied three major facial cues (i.e., visual speech, facial expressions and eye gaze) in this investigation. Each experiment included four conditions:

1. **Virtual Agent (VA):** An animated face was presented on a 2D screen.
2. **Copresent Robot (CR):** The robot was physically present in front of each subject.
3. **Telepresent Robot (TR):** A video or still image of the robotic head was presented to each subject. The videos/images were captured in a frontal angle of the physical agent, and the face in the video was scaled to match the size of the copresent robot.
4. **Human Ground-Truth (GT):** A human performed the task instead of the agent in front of each subject, or the subject was presented with a video recording of the human. If a video was presented, the size of the face in the video was scaled to match the size of the virtual agent's face. The purpose of performing the experiments with GT (human) is to evaluate what we expected to be optimal perception of social cues in our research setting.

In all four conditions, subjects were seated in front of the agent, with the same viewing angle and distance between the subjects and the agent. We used a retro-projected robotic head for this study since computer graphic generated avatars can show natural visual speech and facial expressions, and the same virtual agent animation behaviors can be compared with telepresent and physically present robotic agents. Similar to other robotic platforms, retro-projected robots have some limitations. For instance, Android robotic heads are limited by the number of actuators used in their face, or non-humanoid robots may not be able to show facial expressions. Similarly, since the mask is static in retro-project robotic heads, the jaw and lip movements are only optical and some facial movements (such as nose wrinkling during the expression of disgust) cannot be shown. Therefore, the findings of this investigation cannot be generalized to all other embodiments without considering the relevant differences between the embodied agents.

3. Robotic Platform

Major obstacles for developing realistic robotic faces lie in limitations of the actuators and the skin. The Facial Action Coding System (FACS) (Ekman and Friesen, 1978) codes for approximately 40 primary facial muscles movements (AUs) that are involved in producing facial expressions and mouth movements during speech. Because these actions can be very subtle and quick, mechanical actuators often fail to mimic them. Also, due to cost and space constraints, Android robotic heads have few actuators, and their faces are relatively larger than an average head. For example, the Geminoid H1 robot (Nishio et al., 2007) is approximately five percent larger than its human counterpart (Bartneck and Lyons, 2007). Additionally, the skin of Android robots, which is often made of latex, can produce unnatural wrinkles and folds on the robot's face. For these reasons, and because we aimed to study the effect of embodiment and presence of a robot compared with a virtual agent, we chose a retro-projected robotic platform that can portray natural and realistic facial animation.

In this study, we used Ryan (DreamFace-Tech., 2015), a social robot with the capability of showing facial expressions, eye gaze, visual speech, facial emotion recognition, and subject movement tracking. Ryan uses state-of-the-art character animation technology that is able to show natural visual speech and facial expressions. This platform is designed for face-to-face communication with individuals in different social, learning, and therapeutic contexts.

Ryan (shown in Fig. 2) has a torso equipped with a 10" LCD touch screen which can be used to gather sensory input, display videos, and play games with users. Ryan is equipped with a Microsoft Kinect to track users' movements and two stationary arms for an increased sense of realism. The neck has two degrees of freedom (DoF) providing a total of 180° of yaw, and 45° of pitch. The neck system controls the projector and mask position allowing it to be rotated by the robot application to track faces and head gestures. We developed a face animation system in C# .Net using Microsoft XNA game engine. A

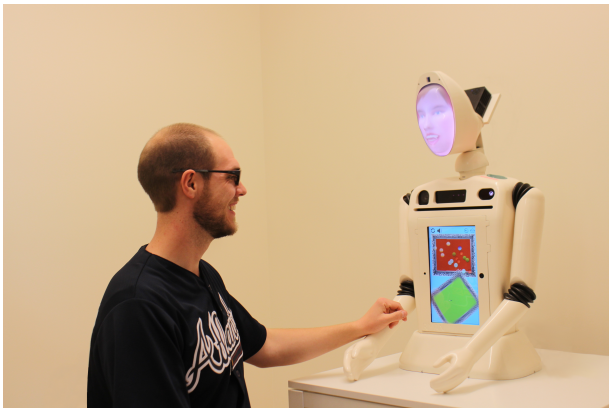


Figure 2: Ryan, the social robot.

graphic artist designed 3D models of different facial expressions and lip movements. The animation software blended the face models to produce accurate natural visual speech and facial expressions based on a multi-target morphing method described in (Ma and Cole, 2004; Mollahosseini et al., 2014). The animation system was used for the virtual agent condition in the rest of the experiments. The same animation was calibrated using the algorithm presented in (Mollahosseini et al., 2014) and then was projected on Ryan's mask.

Animations are based on a multi-target morphing method (Ma and Cole, 2004; Mollahosseini et al., 2014) and .

4. Visual Speech

Visual speech includes the visible oral cues (e.g., movement of the lips, tongue, and jaw) during speech production. These visual cues are not simply a by-product of speech production; they influence auditory perception of speech and vice versa. For example, McGurk and MacDonald (1976) showed that perception of mouth movements can affect the auditory perception of speech and Sweeny et al. (2012a) showed that hearing speech sounds influences the perception of simple visual shapes.

Considering the importance of speech and dialogue in social interaction, it is not surprising that many social robotic platforms have the capability of showing lip synchronization with auditory speech. Mechanical and Android robotic platforms such as Kismet (Breazeal, 2000), HRP-4C (Kajita et al., 2011), FR-i (Oh et al., 2010), Luo Head (Luo et al., 2011) and Alex (Lin et al., 2013) have relatively basic visual speech due to limited actuators and mechanical components that are necessary to control the jaw movements. Virtual agents, on the other hand, have a greater capability for depicting natural visual speech, since advanced computer graphics can be used to generate highly realistic, accurate, and dynamic animations. Nevertheless, lack of physical embodiment and physical presence may constrain the perception of speech in virtual agents. Rear-projected robotic platforms also use computer animation and can thus have faster and smoother lip movement compared with mechanical and Android robots as actuators are not used to control the visual speech. However, since the mask is static (and

therefore the jaw and lips), they might introduce inconsistency between the animation and the final projected face, and possibly even hinder the perception of visual speech . Therefore, it is necessary to study the role of embodiment and presence, especially in this type of robotic heads.

4.1. Related Work

Studies show that virtual embodied talking agents enhance the level of engagement, increase speech comprehension in noisy environments, make agents appear more realistic, and users tend to spend more time with these systems compared to the agents equipped with only voice (Walker et al., 1994; Lester et al., 1999). Siciliano et al. (2003) compared SynFace Virtual Agent (VA) with audio (without visual speech) and video of a human, and concluded that visual-based speech perceptibility of this virtual agent is better than audio only, whereas it is significantly lower than audio-visual perceptibility of human visual speech. Ouni et al. (2003) performed a similar experiment on Baldi virtual agent. They eliminated syntactic and semantic cues by evaluating the perception of visual speech on a non-meaningful series of three Arabic words, and they concluded that speech is better perceived on VA with visual speech than with auditory information only, but still nevertheless significantly lower than audio-visual perceptibility of human visual speech.

Only a few studies have compared the role of embodiment and presence of robotic agents in audio-visual speech perception. Al Moubayed et al. (2013) investigated the role of embodiment of a copresent robotic agent for improving the perception of visual speech. A facial animation on a 2D screen was compared with a retro-projection of the same animation using Furhat (Al Moubayed et al., 2012) and a video of humans from different viewing angles. A collection of short and everyday Swedish sentences with a length of three to six words in each sentence was created. The audio signal quality was reduced using band-pass filtering in specified frequencies and replaced with white noise. Six conditions were studied: audio only, virtual agent viewed at frontal and 45° angle, copresence of a robot viewed at frontal and 45° angle, and the original video recordings of the sentences viewed at the frontal angle. Fifteen sentences were examined in each condition. Auditory-visual perceptual sensitivity was measured as the number of correctly recognized words divided by the number of words in each sentence. This study, conducted on ten subjects with normal hearing, showed that audio-visual speech perceptibility was better perceived with the copresent robot (even though the jaw did not move in the mask), compared with the virtual agent on a flat screen. However, there was no significant difference in the audio-visual perceptibility of the face when it was looked at either from a front-view or a 45° angle on both the virtual agent and robot.

Mollahosseini et al. (2014) studied individuals' experiences and impressions of a proposed visual speech algorithm. In particular, they compared judgments of speech production quality of a virtual agent with retro-projection of the same animation using ExpressionBot. Two short segments of speech were presented with two different lip synchronization approaches (i.e., a proposed approach with kernel smoothing and lip closure in labial phonemes and a basic approach without any further smoothing and processing). The participants (23 typical adults) rated how realistic the visual speech looked on a scale from 0 to 5. Results showed a significant preference for the proposed lip synchronization approach over the basic approach. However, there was no difference in preference for visual speech from the virtual agent compared with the copresent robot.

Table 1 summarizes the results of studies on audio-visual speech perceptibility. As shown, none of these studies compared all three conditions of CR, TR, and VA to distinguish the role of embodiment from the presence of an intelligent agent in perception of visual speech. In this paper, we studied the perception of visual speech from three different types of emotional agents (i.e. VA, TR, CR) as well as from a human (as the optimal case) and based on auditory information alone (as the baseline) using the same experimental setup. Since the methodology and evaluation metrics of evaluating visual speech perception are not standard across the literature, we introduced a new test of visual speech perception along with standard criteria to evaluate the visual speech perception.

4.2. Methodology

We used the same visual speech algorithm presented in (Mollahosseini et al., 2014), which is based on a multi-target morphing method (Ma and Cole, 2004). In particular, the recorded utterances are processed by the Bavieca speech recognizer (Bolanos, 2012), which receives the sequence of words and the speech waveform as input and provides a time-aligned phonetic transcription of the spoken utterance. The aligned phonemes are represented using the International Phonetic Alphabet (IPA), a standard that is used to provide a unique symbolic notational for the realization of phonemes in all of the world's languages (IPA-Handbook, 1999). Having IPA in our system will allow us to add other languages easily as long as the speech recognizer is trained for that language.

For a given language, visually similar phonemes are grouped into units called visemes. For example, the consonants /b/, /p/ and /m/ in the words "buy," "pie," and "my" form a single viseme class. English phonemes are categorized into 20 viseme classes. These classes represent the articulation targets that the lips and tongue move toward during speech production. A graphic artist designed 3D models of these viseme classes in Maya. Finally, natural visual speech was obtained by blending the proper models corresponding to each part of speech with different weights.

The avatar system converts phonetic symbols into the corresponding visemes, and synchronizes them with the audio signal. To achieve a smooth and realistic appearance, the algorithm models coarticulation by smoothing across adjacent visemes using a kernel technique, while ensuring lip closure for labial phonemes (e.g., /b/, /m/, /p/).

4.3. Visual Speech Experiment

Seventeen native English speakers, eight female and nine males, with age range of 19-39 years (Mean= 27.7, SD=6.8) and normal hearing evaluated the audio-visual speech in five

Table 1: Summary and overview of literature comparing audio-visual speech in different conditions

Work	Agent	Condition*				Description	Results**
		CR	TR	VA	GT		
Siciliano et al. (2003)	SynFace			✓	✓	• 12 normal hearing (NH) and 13 hearing-impaired (HI) listeners • Audio signal was degraded for NH group • Video of the original talker was used for GT	• Average perceptibility of VA increased by 22% compared to audio only • Perceptibility of VA was significantly lower than GT
Ouni et al. (2003)	Baldi			✓	✓	• Non-meaningful series of three Arabic words presented to 19 participants • Total of 300 words and 100 trials • Audio signal was degraded	• Average perceptibility of VA increased by 24% compared to audio only • Perceptibility of VA was 15% lower than GT
Al Moubayed et al. (2013)	Furhat	✓		✓	✓	• Audio-visual perception viewed at frontal and 45° angle. • A collection of short Swedish sentences • Reduced audio signal quality	• Audio-visual speech was better perceived on CR compared with VA. • No significant difference between frontal and 45° view angle
Mollahosseini et al. (2014)	Expressionbot	✓		✓		• Two short segments of speech • Examined two different lip synchronization approaches. • Participant rated how realistic the visual speech looked on a scale from 0 to 5	• Significant preference for the proposed visual speech approach over basic method • No significant preferences between CR and VA
This Work	Ryan	✓	✓	✓	✓	• Section 4.2	• Section 4.4

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.
** Only the relevant finding from the original papers are reported in this summary.

conditions (VA, CR, TR, GT, and audio only). Unlike the auditory speech (e.g., an evaluation of hearing ability), there is not a standard methodology to evaluate the perception of visual speech. Several researchers have thus developed their own approaches and evaluation criteria. The sets of sentences in the majority of these studies (See Table 1) are not comprehensive and do not consider syntactic and semantic cues. Measures of performance such as the number of correctly recognized words divided by the number of words in each sentence (Al Moubayed et al., 2013), or subjective evaluation of how realistic the visual speech appeared (Mollahosseini et al., 2014) are not standard, either. To address this issue, we developed an Audio-Visual Speech Perception In Noise (AV-SPIN) test to evaluate the perception of visual speech using a systematic and standardized approach. The AV-SPIN material, including videos, sentences, and IPA aligned auditory information, will be publicly available to the research community.[1]

The Speech Perception In Noise (SPIN) test was developed to address sensory and linguistic cognitive processes of everyday speech (Elliott, 1995; Kalikow et al., 1977). SPIN consists of 250 meaningful sentences categorized as High-Predictability (HP) sentences and 250 non-meaningful sentences categorized as Low-Predictability (LP) sentences. The listener's task is to recognize the last word in each sentence (referred to as the keyword). HP sentences contain syntactic and semantic cues helpful for predicting the keyword (e.g., *The sleepy child took a nap*), while LP sentences do not provide any cues predictive of the keyword (e.g., *Betty knew about the nap*). The sentences were divided into ten sets each containing 50 sentences (25 HP and 25 LP sentences), where odd-numbered sets were complementary of even-numbered sets (i.e., same keywords were in the opposite type of sentence).

Bilger et al. (1984) studied the SPIN test on 128 listeners (aged 19 to 69) with sensorineural hearing loss and proposed a revision (R-SPIN) such that different sets produce equivalent results. Particularly, 31 sentences and their complements were eliminated, 19 sentence pairs were arbitrarily removed, and the remaining sentences were redistributed to create 200 HP sentences and their complementary 200 LP sentences. These 400 sentences were divided into eight sets each containing 50 sentences (25 HP and 25 LP sentences), where odd-numbered sets were complementary of even-numbered sets. Traditionally the R-SPIN is presented with ambient noise at a Signal-to-Noise Ratio (SNR) of 8 dB.

Since R-SPIN is strictly auditory, audio-visual perceptibility cannot be examined with the original R-SPIN materials. Therefore, we created an AV-SPIN corpus by capturing a native English speaker's face as she produced R-SPIN sentences. Similar to the R-SPIN test, the quality of the audio signal was degraded by babble noise. Since the subjects were not hearing-impaired, the audio signal was presented at a high signal-to-noise ratio of -9 dB (i.e., the power of the noise was significantly higher than the auditory speech signal).

In all conditions (VA, CR, TR, GT, and audio only), subjects were seated in front of the agent at a distance of 60 cm. To maintain voice consistency between the conditions, the audio signals were extracted from the videos and force-aligned using Bavieca speech recognizer (Bolanos, 2012). Twenty sentences (10 HP and 10 LP sentences) were randomly assigned to each condition for each subject. A different set of sentences was used to train the subjects at the beginning of the experiment.

Each subject participated in all five conditions (audio-only, VA, CR, TR, and GT) in a random order. The LP and HP sentences in each condition were shuffled and were selected such that each condition did not share any sentences. The sentences were played only once, and at the end of each sentence, the subject had 30 seconds to write down the keyword (last word in each sentence). The subjects could adjust the sound volume at their convenience during the training period, but the same audio volume was used in all conditions of the remaining experiments. A set of headphones with the same audio volume was used in all the conditions. Since headphones were used, the direction of the voice did not play a role in the perception of speech. In addition, this allowed us to eliminate other roles, and

[1] A copy of AV-SPIN is available in: http://www.mohammadmahoor.com/databases-codes/

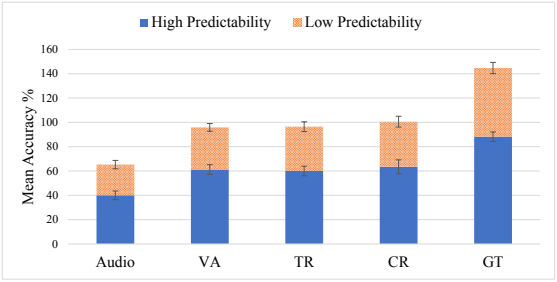


Figure 3: The average accuracy of audio-visual speech perception in different conditions.

only study the psychological effect of presence/embodiment of the robot. Each subject performed the experiment only once, since hearing a keyword in HP could have helped the subject to identify it in an LP sentence.

4.4. Visual Speech Results

We performed a 2 (Predictability; HP, LP) × 5 (Condition; VA, CR, TR, GT, and Audio-Only conditions) ANOVA with both predictability and agent as the within-subject factors. The test showed a significant main effect of agent [$F(4, 64) = 30.48$, $p < .0001$] and a significant main effect of predictability [$F(1, 16) = 134.55$, $p < .0001$]. The interaction between agent condition and sentence predictability, however, was not significant [$F(4, 64) = 1.44$, *n.s.*].

Figure 3 shows the mean accuracy for each condition. To measure whether the differences between different agent conditions was significant, we performed a post-hoc Least Significant Difference (LSD) analysis. Table 2 shows the significance of different comparisons using a post-hoc LSD analysis. All other combinations not included in the table were significantly different from each other. As Table 2 shows, VA (and all other conditions) produced significantly better audio-visual perceptibility than the audio-only condition ($p < .001$). This confirms that visual information can affect speech perception, and shows the efficacy of the visual speech algorithm. The ground-truth (video of the human) had significantly higher audio-visual perceptibility than the other conditions ($p < .001$), which indicates that the proposed visual speech algorithm has room for improvement.

In order to measure the effect of predictability of the sentence in only VA, TR and CR conditions, we performed a separate 2 (Predictability; HP, LP) × 3 (Condition; VA, CR, TR) ANOVA with both predictability and agent condition as the within-subject factors. The analysis showed that the main effect of predictability was still significant [$F(1, 16) = 82.03$, $p < .0001$], however the main effect of agent was not significant [$F(2, 32) = .381$, *n.s.*] nor was the interaction between agent condition and predictability [$F(2, 32) = 1.44$, *n.s.*]. In other words, embodiment and presence did not improve the perception of visual speech regardless of syntactic and semantic cues in the sentences.

Our results indicated that physical embodiment (Research Question 1), physical presence (Research Question 2), and the joint effect of physical embodiment and presence (Research Question 3) did not differ in the extent to which they improved

Table 2: Post-hoc LSD statistical significance of the different conditions in audio-visual speech perception.

Condition1	**Condition2**	***p*-value**
Audio	Other Conditions	<.0001
Virtual Agent	Tele-present Robot	0.949
Co-present Robot	Virtual Agent	0.606
Co-present Robot	Tele-present Robot	0.652
Ground-Truth	Other Conditions	<.0001

the perception of visual speech regardless of syntactic or semantic cues in the sentences. This could be because the mask was static and the jaw and lip movement were only optical in the retro-projected robotic platform. Other types of embodiment, such as Android robots, may express different behaviors. However, since controlling natural lip movement on Android robots necessitates several actuators and a very elastic skin, existing Android robotic faces may even perform worse than computer graphics animations.

This finding is consistent with our earlier study (Mollahosseini et al., 2014), but inconsistent with the study by Al Moubayed et al. (2013), though similar retro-projection technology with a static mask was used in both studies. It is unlikely that the results were influenced by different visual speech algorithms. It is more likely that the difference between Al Moubayed et al. (2013) and our finding is due to different audio-visual corpus and the perception criteria. The audio-visual corpus used in the present study was a standard set considering the syntactic and semantic cues in the sentences, while Al Moubayed et al. (2013) used a collection of short, everyday sentences with the number of correctly recognized words divided by the number of words in each sentence as the criterion of perception. Additionally, the sample size may also have affected the results, as the study performed by Al Moubayed et al. (2013) was evaluated with ten subjects, compared to this study that 17 subjects participated in.

5. Facial Expressions

Facial expression is one of the most critical nonverbal channels used by human beings to convey emotion. Emotion is not only critical in creating more sensitive and effective intelligent agents but also impacts how people respond to the agent (Beer et al., 2011). Hence, facial expression is a vital component in natural social interaction and Human-Robot Interaction (HRI) systems, and has been employed in a variety of robots such as Kismet (Breazeal, 2003), the Philips iCat (Van-Breemen, 2004), Geminoid F (Becker-Asano and Ishiguro, 2011), and on-screen agents (Cassell, 2000; Bruce et al., 2002).

Mechanical and Android robotic platforms control face movement using actuators in their faces. However, due to cost and space constraints, the number of actuators in robotic faces are often limited. Moreover, because facial actions involved in facial expression can be very subtle and quick, mechanical actuators often fail to mimic them. Computer-graphic animations, on the other hand, have a greater capability for controlling facial

movement, but their lack of physical embodiment and physical presence may constrain the perception of facial expression in virtual agents. Retro-projected robotic heads add physical embodiment to computer animation agents, but since the mask is static, some of the facial movements such as nose wrinkling in the expression of disgust cannot be portrayed on a robotic face. Therefore, it is important to investigate the role of embodiment and presence to find out whether physical embodiment and presence can improve the perception of an agent's facial expressions. A few studies have compared the role of embodiment and presence in the perception of robotic agents' facial expressions, and to the best of our knowledge perception of facial expression on retro-projected robotic heads has not yet been investigated.

5.1. Related Work

A few studies have compared the role of embodiment and presence in the perception of robotic agents' facial expressions. Bartneck et al. (2004) studied the role of presence in perception of intensity and recognition accuracy of facial expression using the robotic character iCat (Van-Breemen, 2004) and its telepresence condition (movie on a screen). Subjects were asked to categorize each emotion and rate its intensity. The study found a non-linear relationship between the geometrical intensity (robot's expression intensity) and the intensity of emotions perceived by the user. The results also indicated that emotions depicted by the robot were judged as having greater intensity, but there was no significant difference in the perceived intensity and recognition accuracy between the presence of the robotic character and its telepresence.

Kätsyri and Sams (2008) investigated the effect of dynamics on identifying basic emotions between a virtual agent (Talking Head) and a video of a human. Dynamic and static depictions of six basic emotions from a human face and a virtual agent were shown to 54 subjects. Subjects identified expressions on the human face much better than on the virtual agent. There was no significant difference in the identification of static and dynamic expressions of the human face. Identification of some expressions such as anger and disgust on the virtual agent failed to exceed chance level in the static condition, while dynamics improved it notably in lower intensities.

Mollahosseini et al. (2014) studied the extent to which embodiment and physical presence improved the perception facial expression. The study evaluated how accurately subjects were able to interpret the facial expressions of a virtual 2D agent and its projection on a retro-projected robotic platform. Six basic emotions at their maximum intensity level were displayed in random order, and subjects were then asked to associate each with one of these six categories or to indicate that none were appropriate. They found similar recognition rates for happiness, sadness, surprise, disgust and fear in both a virtual agent and a copresent robot, and superior performance for anger when portrayed by the robot.

Lazzeri et al. (2015) studied the role of embodiment in conjunction with presence on a humanoid Android robot (Robot FACE). Fifteen subjects identified six basic emotions displayed on the robot, in 2D photos of the robot, 3D virtual animation models, as well as a set of 2D photos and 3D models of a human taken from Bosphorus Database (Savran et al., 2008). Preliminary results showed that facial expressions were better identified on the robot than its virtual animation, and the recognition rates of facial expressions performed by the robot were similar to those achieved with human stimuli.

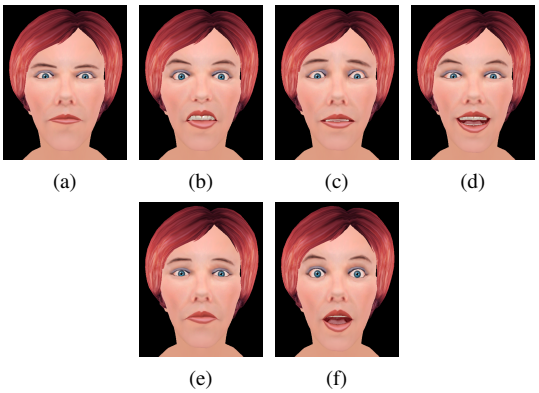


Figure 4: Six basic facial expressions at their maximum intensity: a) Anger, b) Disgust, c) Fear, d) Happiness, e) Sadness, and f) Surprise.

Table 3 summarizes studies of facial expression perception with robots and their relevant findings. As shown, none of these studies compared all three conditions of CR, TR, and VA to distinguish the role of the embodiment from the presence of the robot. In this paper, we studied all three different conditions of emotional agents (i.e. VA, TR, CR) as well as human facial expressions (as the optimal case) in the same experimental setup. We also investigated emotion perception at different intensity levels to study the effect of intensity level on perception of different agents' facial expression.

5.2. Methodology

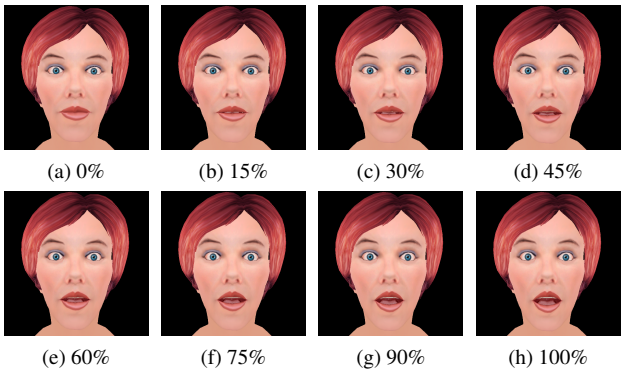
In order to design realistic and standard facial expressions in our animation system, we used the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). The FACS model is a well-known approach for quantifying affective facial behaviors, and describes all possible facial actions in terms of Action Units (AUs). The FACS explains facial movements and does not describe affective state directly. Friesen and Ekman (1983) proposed EMFACS to convert AUs to affect space. For example, EMFACS states that happiness involves raising of the cheek (AU 6) and pulling of the corner of the lip (AU 12), whereas sadness involves raising of the inner brow (AU 1), lowering of the outer brow (AU 4) and depression of the corner of the lip (AU 15). For the current experiment, a graphic artist designed 3D models of six basic expressions (i.e., anger, disgust, fear, happiness, sadness and surprise) in Maya based on EMFACS. Figure 4, demonstrates six basic facial expressions at their maximum intensity used in our animation system.

In order to show facial expressions at different intensities and blend them with visual speech, we used the same algorithm presented in (Mollahosseini et al., 2014). In particular, our animation used the following formula to generate the morph target

Table 3: Summary and overview of literature comparing perception of emotion in different conditions

Work	Agent	Condition*				Emotion†		Description	Results**
		CR	TR	VA	GT	No.	In		
Bartneck et al. (2004)	iCat	✓	✓			5	✓	• Ten geometrical intensities were displayed • Participants recognized the emotion and its intensity	• The relationship between the geometrical and perceived intensity was not linear • No significant difference between CR and TR in the intensity and recognition accuracy
Kätsyri and Sams (2008)	Talking Head			✓	✓	6		• Dynamic and static facial expressions were studied	• GT perceived better than VA • Dynamics did not improve GT • Dynamics improved recognition of subtle emotions, notably anger and disgust of VA.
Mollahosseini et al. (2014)	Expressionbot	✓		✓		6		• Participants selected six categories as well as "none"	• Superior recognition performance for anger in CR • Similar recognition rates for other emotions in both CR and VA
Lazzeri et al. (2015)	The Robot FACE	✓		✓	✓	6		• The robot, its 2D&3D models, and 2D&3D models of human were shown • Physiological signals of subjects were recorded	• CR was better perceived than 2D photos or 3D models (VA and GT) • No significant differences in the subjects' psychophysiological states
This Work	Ryan	✓	✓	✓	✓	6	✓	• Section 5.2	• Section 5.4

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.
†*No.* is the number of studied emotions and *In* stands for whether different Intensity levels are studied.
** Only the relevant finding from the original papers are reported in this summary.



(a) 0% (b) 15% (c) 30% (d) 45%

(e) 60% (f) 75% (g) 90% (h) 100%

Figure 5: Different intensity level of surprised emotion on the virtual agent

based on the current viseme and emotion morph targets:

$$F_j = F_c + \lambda_j(F_j^{max} - F_0) \quad (1)$$

where F_c represents the current viseme, F_j^{max} is the desired expression model at the maximum intensity, F_0 is the Neutral model. The parameter $\lambda_j \in [0, 1]$ is the intensity of the j^{th} expression model F_j.

5.3. Facial Expressions Experiment

We evaluated the perception of facial expressions of emotion performed by different agents with 48 subjects, 23 female and 25 males, with age range of 18-35 years (Mean= 24.6, SD=5.2). Six basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) were displayed in four conditions corresponding to the types of agents (VA, CR, TR, and GT) at seven intensity levels (15%, 30%, 45%, 60%, 75%, 90% and 100%). Each emotion was displayed with an animation/movie starting from neutral until the face's expression reached one of seven intensities. The animations took one second from neutral to the desired intensity and then remained static until the subject responded. Subjects were asked to categorize the emotion of the face as belonging to one of the six basic emotional categories (listed above) or to report "none" if they were unable to assign the facial expression to any of the six categories.

To evaluate the GT condition, subjects were presented with the video recordings of an actress portraying the facial expressions randomly selected from the extended CK+ dataset (Lucey et al., 2010). In order to pair the intensity of GT with the animation, two experts annotated the intensity of emotions between 0 to 100%, frame by frame. The intensity of each frame was considered as the average intensity of the two annotators. Each video in the GT condition took one second, started with a neutral expression, and ended at the desired emotional intensity level. Since the animation uses a weighted blend shape technique defined in (1), the intensity of emotion on the animation was easily defined by changing the parameter λ_j from zero to the desired intensity level over one second. Figure 5 shows different intensity levels of a sample emotion (surprise) on the virtual agent. Clearly, more subtle emotion intensities are more difficult to discriminate and could easily be confused with a different emotion.

Subjects were seated in front of the agent at a distance of 60 cm. Each combination of emotion and intensity was displayed twice in each block of trials, one with each intensity level, where the lowest intensity faces were shown first, then the second lowest, etc. In other words, subjects categorized 84 emotions (2 trials × 6 emotions × 7 intensities) where the first 12 videos/animations portrayed six emotions at intensity level 15% each played twice randomly, the second 12 videos/animations portrayed six emotions at intensity level 30%, and so on. The reason for sorting the trials by intensity level was that the subjects could have recognized the facial movement of an emotion at higher intensity levels and generalized the facial movements for recognition at lower intensity levels. In addition, each subject participated in only one agent condition, since VA, TR, and CR share the same animation and seeing an emotion at a higher intensity level of one condition could have helped the subject to recognize that same emotion at a lower intensity in another condition, on a different agent.

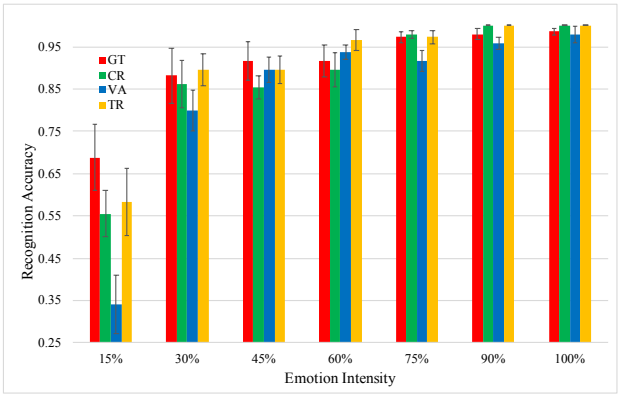


Figure 6: The average accuracy of emotion perception in different conditions.

Each subject participated only in one agent condition (i.e., 12 subjects rated the facial expressions displayed by one particular agent). In each condition, subjects saw facial expressions on an agent, and they were asked to select one of the six basic facial expressions (Anger, Disgust, Fear, Happiness, Sad and Surprise) or None.

5.4. Facial Expression Results

A mixed 6 (emotions) × 7 (intensities) × 4 (agent conditions: CR, TR, VA and GT) ANOVA with emotion and intensity as the within-subject factors and embodiment as the between-subjects factor was conducted. The dependent variable was recognition accuracy. The recognition accuracy differed significantly between emotions [$F(5, 220) = 10.86$, $p < .0001$] and between intensity levels [$F(6, 264) = 129.27$, $p < .001$]. Not surprisingly, faces with higher intensity received higher recognition accuracy. This analysis also revealed a significant interaction between the emotion and agent condition [$F(15, 220) = 1.95$, $p < .020$]. The interaction between agent and intensity was also significant [$F(18, 264) = 3.97$, $p < .001$]. This suggests that there is a difference among agents at low intensities, but not high intensities. In other words, the type of agent is particularly important when recognizing subtle expressions. The three-way interaction was also significant [$F(90, 1320) = 1.55$, $p < .001$]. This suggests that the dependency on intensity is only important for certain emotions (the intensity × agency interaction was significant for anger, fear, sad, and surprise, all p's $< .05$). Figure 6 shows the mean accuracy for each agent condition at different intensity levels, collapsed across the different expressions. As shown, the subjects discriminated emotion better on CR than on VA or TR.

There was also a significant main effect of agency on recognition accuracy [$F(3, 44) = 3.06$, $p = .038$]. Post-hoc LSD analysis on different agent conditions indicated that expression recognition for TR was significantly worse than for human ground-truth and CR (p-values of 0.010 and 0.014, respectively). All other agent conditions were not significantly different from each other or ground-truth. In other words, both embodiment and presence were important factors in improving the perception of emotional expressions. Expression discrimination was better for the ground-truth (video of the human) condition than the other conditions, which indicates that the facial expression of the animation has room for improvement.

Table 4 shows the confusion matrices of the emotion recognition rates for the different agent conditions of CR, TR, and VA. The highest values are shown in bold. As shown, anger, happiness, and sadness were perceived better on CR, while disgust, fear, and surprise were recognized better on the virtual agent. To address whether this difference was significant between different emotions, separate post-hoc LSD analyses were conducted for each emotion. Table 5 shows the result of pairwise comparisons post-hoc LSD analyses and effect sizes of different agent conditions for different emotions. Cohen's d is an effect size used to indicate the standardized difference between two groups defined as:

$$d = \frac{M_1 - M_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \quad (2)$$

where M_i is the mean and σ_i is the standard deviation of group i. Generally, the effect size is considered small if $d > 0.2$, medium if $d > 0.5$ and large if $d > 0.8$ (Cohen, 1977). As indicated in Tables 5:

- Anger was recognized better on both CR and TR compared to VA, with a medium effect size (effect of embodiment, Research Question 1).
- Recognition of disgust, fear, and happiness was equivalent across all the agent conditions.
- Sadness was recognized better on CR compared to TR and VA, with a medium/large effect size (the join effect of embodiment and presence, Research Question 3).
- Surprise was recognized worse on TR comparing with VA, with a medium effect size (effect of embodiment, Research Question 1). However, Surprise was recognized better on CR compared with TR, with a medium effect size (the effect of presence, Research Question 2).

We believe that the negative effect of physical embodiment on the perception of an agent's surprised expression could have occurred because the jaw did not move in the static mask, making subtly surprised faces difficult to perceive. This phenomenon (i.e., the effect of seeing a moving expression on a static mask) was presumably less noticeable when the robot was present in front of users (CR condition), as the difference between CR and VA was not significant for the expression of surprise. Since the only varying factor between TR and CR was the "presence" of the robot, we believe that presence could potentially compensate for the negative effect of seeing facial movements on a static mask.

These results are consistent with our previous study (Mollahosseini et al., 2014), indicating that subjects perceived the facial expression of anger (and sadness in the present study) with greater accuracy in the robotic face than that of the virtual agent. Our finding is also consistent with (Bartneck et al., 2004). We also found a significant difference between the robot

Table 4: Confusion matrix of the emotion recognition rates (in percentage) of CR, TR and VA with presented facial expression (columns) against subjects' judgments (rows).

	Copresent Robot (CR)						**Telepresent Robot (TR)**						**Virtual Agent (VA)**					
	AN*	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
Anger	**95.2**	2.4	0.6	0.0	0.6	0.0	91.1	5.4	0.0	0.0	1.8	0.0	81.5	4.8	0.0	0.0	0.6	0.6
Disgust	0.0	87.5	1.2	1.2	0.0	0.6	3.0	77.4	1.2	0.0	0.0	0.0	3.0	**90.5**	3.0	0.6	0.6	0.0
Fear	0.0	0.6	78.6	0.0	0.0	3.0	1.2	1.2	76.8	0.0	3.6	5.4	1.8	3.6	**81.5**	0.0	3.6	3.0
Happiness	0.0	0.0	0.0	**89.9**	0.0	1.2	0.0	0.6	0.6	83.9	0.0	0.6	0.0	0.0	0.0	89.3	0.0	0.6
Sadness	1.2	3.0	14.3	0.0	**98.2**	1.8	2.4	1.8	14.9	0.0	88.7	3.0	6.5	0.6	10.7	0.0	89.3	1.2
Surprise	0.0	1.2	2.4	7.7	0.0	91.1	1.2	4.2	3.6	10.7	0.0	81.5	0.6	0.6	4.2	10.1	0.0	**94.6**
None	3.6	5.4	3.0	1.2	1.2	2.4	1.2	9.5	3.0	5.4	6.0	9.5	6.5	0.0	0.6	0.0	6.0	0.0
Total Accuracy			**90.08**						83.23						87.80			

*AN, DI, FE, HA, SA, and SU stand for Anger, Disgust, Fear, Happiness, Sadness, and Surprise, respectively.

Table 5: Pairwise comparison (LSD p-value) and Cohen's d effect size of users' perception of facial expressions on different agent conditions.

	TR vs CR		VA vs CR		VA vs TR	
	p	d	p	d	p	d
Anger	.405	.271	**.004**	**.652**	**.031**	**.419**
Disgust	.121	.347	.642	.155	.050	.493
Fear	.447	.139	.913	.002	.386	.135
Happiness	.340	.023	.784	.230	.222	.218
Sadness	**.034**	**.789**	**.046**	**.727**	.891	.008
Surprise	**.010**	**.421**	.426	.237	**.001**	**.658**

and telepresence of the robot for perception of the facial expressions of sadness, similar to Bartneck *et al.*, who found a significant difference between CR and TR for recognizing sadness at intensities lower than 30%.

This finding is inconsistent with a study by Lazzeri et al. (2015) in which all emotions were better perceived on a robotic agent than on a virtual agent. Perhaps, the difference between (Lazzeri et al., 2015) and our finding is mainly due to the difference between the embodiments (i.e., Android vs retro-projected robotic heads). The masks in retro-projected robotic heads are static, thus jaw and the lip movements are only optical and some facial movements such as nose wrinkling in the expression of disgust cannot be shown, whereas Android robotic heads can be more flexible in controlling the skin if enough actuators are provided. In addition, Lazzeri et al. (2015) created a synthesized virtual agent from a set of pictures of a physical robot acquired from various angles and used Unity 3D software to animate the 3D models. Our virtual agent featured an accurate 3D model which was projected on the robotic face. Hence, the same animation and expression dynamics were used in both our robot and virtual agent conditions.

6. Eye Gaze

Eye gaze is one of the most basic and important features of the human face for nonverbal communication. Humans incorporate gaze both consciously and unconsciously into various human-human interaction schemes (Chen and Yeh, 2012). For example, neurons in the primate visual cortex can respond selectively to eye gaze, head orientation, or even the combination of both (Perrett et al., 1985). Eye gaze serves several different functions such as capturing attention, maintaining engagement (Cassell, 2000), conveying information about emotional and mental state (Ruhland et al., 2014), augmenting verbal communication (Emery, 2000), orchestrating turn-taking, and deictic reference (Kendon, 1967).

Considering the importance of eye gaze in social interaction, it is not surprising that social gaze behavior has been studied in many robotic platforms (Imai et al., 2002; Yoshikawa et al., 2006; Mutlu et al., 2009). Mechanical and Android robotic platforms control eye gaze by using actuators in the eyeballs. These actuators, however, may not be fast or accurate enough to replicate movement of the human eyes. The movement of the human eye is controlled by three pairs of muscles and it can reach an angular speed of about 400°/sec with 200ms time to initiate (Pateromichelakis et al., 2014). Computer graphics animations, on the other hand, have a greater capability for producing natural-looking eye gaze (Cassell, 2000; Ruhland et al., 2014). However, it is known that the perception of 3D objects that are displayed on 2D surfaces is influenced by the Mona Lisa effect (Todorović, 2006). Hence, the lack of physical embodiment and physical presence may constrain the perception of virtual agents' eye gaze.

6.1. Related Work

Many studies in vision science have evaluated head-eye gaze, but only on telepresent faces (Baron-Cohen et al., 1995; Allison et al., 2000; Itier and Batty, 2009; Sweeny et al., 2012b). Although embodiment and presence have been studied individually, there is not a comprehensive study that distinguishes the role of embodiment and presence in gaze perception. Gaze perception of a physically present human agent and his video was studied on a TV set by Anstis et al. (1969). In this classic study, subjects were asked to report the point on a glass screen at which the agent (TV or a human) was looking. To simulate head rotation in the telepresent condition, the TV set was rotated. The agent's head was rotated to -30°, 0° and 30° angles. The study found that eye gaze was much better perceived on a physically present human agent than on its telepresent counter-

Table 6: Summary and overview of literature comparing perception of eye gaze in different conditions

Work	Agent	Condition*				EG†	Description	Results**
		CR	TR	VA	GT			
Anstis et al. (1969)	TV		✓		✓	✓	• A horizontal scale (ruler) was used • Video of a human used for TR • The agent's head was rotated with -30°,0° and 30° angles	• Errors were greatest when head rotation and eye rotation were incongruent.
Delaunay et al. (2010)	LightHead	✓	✓		✓	✓	• A grid with 100 cells was used • Video of a human used for TR • Instead of head rotation, subjects viewed the Agent with 0° and 45° angles	• CR performed better than TR • GT performed significantly better than other conditions, in both frontal and side view situations
Al Moubayed and Skantze (2012)	Furhat	✓			✓		• A grid with nine cells was used • Vergence, parallel eyes, static and dynamic eyelids	• Perception of gaze was significantly worse when the head was moving compared with eye movement alone. • No significant difference between gaze with and without vergence.
Al-Moubayed et al. (2012)	Furhat	✓		✓			• Mona Lisa effect studied on five subjects sitting around a circle. • Only eye rotation studied	• Gaze was perceived more accurately on CR
Misawa et al. (2012)	LiveMask	✓		✓			• Photos of a person looking from -30° to 30° • Instead of rotating the head, subjects' view angle was changed	• CR was significantly better than VA • The Mona Lisa effect occurred in VR
Mollahosseini et al. (2014)	Expressionbot	✓		✓			• Mona Lisa effect studied on five subjects sitting around a circle	• Discrimination of eye gaze was better on CR
This work	Ryan	✓	✓	✓	✓	✓		

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.
†EG stands for Emergent Gaze which is defined as simultaneous movement of head and eye-gaze.
** Only the relevant finding from the original papers are reported in this summary.

part, and the perception of gaze was distorted with the rotation of the TV.

Delaunay et al. (2010) studied gaze perception on the LightHead robotic face, its telepresence, and the gaze of a human agent. A vertical glass screen with a 10x10 grid was placed between the agents and the subjects, and subjects were asked to report the gaze point when viewed from a frontal and 45° angle. Since asking a human to hold his/her head steady in a 45° position was not possible and chin/forehead rests did not allow horizontal rotations, to study the effect of head rotation, subjects were instead moved to a position with a 45° angle with respect to the agent. Under these conditions, subjects judged gaze from the video and the robot in both frontal and 45° view situations with equal sensitivity.

Al Moubayed and Skantze (2012) compared the perception of eye gaze on Furhat robotic face with a human agent in different conditions (i.e., presence of vergence, static/dynamic eyelids, etc.). They took a different approach by asking the agents to look at nine points on a table between the agent and the subjects. In this case, there was no significant difference between gaze with vergence and without vergence. Furthermore, head movement appeared to be more effective for influencing judgments along the horizontal axis while eye movement dominated judgments along the vertical axis. Regardless of conditions, gaze from the human agent was perceived better than gaze from the robot.

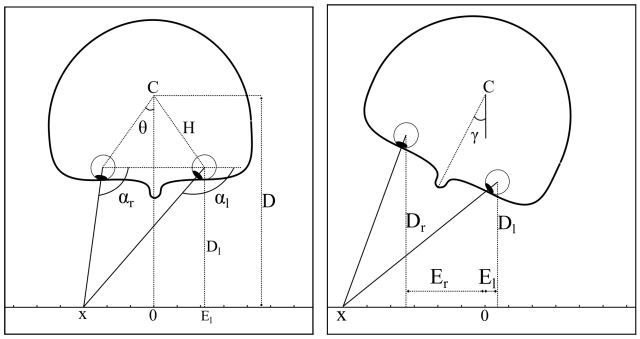
Studies show that virtual agents suffer from the Mona Lisa effect (Misawa et al., 2012; Al-Moubayed et al., 2012; Mollahosseini et al., 2014), in which the eyes in a picture appear to be looking at the viewer regardless of their location in front of the picture. For example, Al-Moubayed et al. (2012) studied the Mona Lisa effect on a virtual agent and its 3D projection on Furhat robotic face. Five subjects were simultaneously seated around the agent, each of whom was asked to report their perception of the agents' eye gaze direction. The results showed a clear Mona Lisa effect in the virtual agent since many subjects perceived a mutual gaze at the same time.

Table 6 summarizes several studies on eye gaze perception and their most relevant findings. The majority of these studies report that physical presence plays a greater role in perception of an agent's eye gaze than physical embodiment. Presumably, having a 3D view of the nose direction, the eye position and their composition help viewers to perceive eye gaze direction more accurately. Additionally, few studies have explored emergent gaze. Emergent gaze occurs when the visual system integrates global information about the rotation of the head with local information about the rotation of the eyes, to compute a distinct metric of gaze present in neither feature alone (Wollaston, 1824; Cline, 1967; Kinya and Mitsuo, 1984; Langton et al., 2004; Kluttz et al., 2009; Otsuka et al., 2014; Sweeny and Whitney, 2017). This approach to measuring gaze perception has been surprisingly underutilized in robotics work.

The present study evaluates the perception of emergent gaze, while at the same time comparing the roles of embodiment and presence of the robot. One of the reasons that emergent gaze has not been studied extensively both with humans and robots is the difficulty inherent in controlling the movements of a human agent. Rotating a human's head and eyes to an exact position requires special apparatuses, and it complicates the experiment process. Hence, most studies of gaze either do not include a condition with a human agent, or they use a typical chin/forehead rest to fix the human's head in place, which precludes examination of emergent gaze.

6.2. Methodology

To evaluate the accuracy of agents' eye gaze in the current investigation, the agent looked at a particular point on a glass divider located between the agent and the subjects. A horizontal



(a) Head facing forward (b) Head rotated by γ

Figure 7: Schema and the variables used in the calculating eye gaze angle (Drawing not to scale).

line with fifty-one equidistant points was drawn on the glass. The agent looked at a point on the glass screen and subjects were asked to report their perception of the agent's gaze direction.

In order to precisely set eye gaze toward a target point, we needed to rotate the agents' eyeballs such that the pupils were directed towards the target point. In this study, the target points were at agent's eye level, hence we only needed to change the yaw angle for the eyes. Assuming the face is frontal (rotated zero degrees), the yaw angle for right and left eyes (α_r and α_l, respectively) is calculated as:

$$\alpha_r = \frac{\pi}{2} - \arctan\frac{x + E_r}{D_r} \tag{3}$$

$$\alpha_l = \frac{\pi}{2} - \arctan\frac{x - E_l}{D_l} \tag{4}$$

where $x \in [-75cm, 75cm]$ is the target point on the glass screen. E_r and E_l are the distance of right and left eye from the center of the glass screen in the x-Axis, and D_r and D_l are the distance of the right and left eyes from the glass screen in the y-Axis, calculated as:

$$E_r = E_l = H \times \sin(\theta) \tag{5}$$

$$D_r = D_l = D + H \times \cos(\theta) \tag{6}$$

where H is the distance of the head pivot point (C) to the center of the eyes, θ is the angle between the eyes and the head pivot point, D is the distance of the head pivot point to the glass screen. Figure 7a shows the schema and the variables used in these calculations.

When the head is straight and not rotated, $D_l = D_r$ and $E_r = E_l$. If the head is rotated by $\gamma°$ (Figure 7b), the values of E_r and D_r in Equations (3) and (4) are changed as follows:

$$E_r = H \times \sin(\theta + \gamma) \tag{7}$$

$$E_l = H \times \sin(\theta - \gamma) \tag{8}$$

$$D_r = D - H \times \cos(\theta + \gamma) \tag{9}$$

$$D_l = D - H \times \cos(\theta - \gamma) \tag{10}$$

In the above equations, we assumed that the agent does not have any facial curvature in the eye area (Figure 8-left). If the face has an angle (ϵ) in the eye area (Figure 8-right), Equations (3) and (4) will change as follows:

$$\alpha_r = \frac{\pi}{2} - \arctan\frac{x + E_r}{D_r} - \epsilon \tag{11}$$

$$\alpha_l = \frac{\pi}{2} - \arctan\frac{x - E_l}{D_l} + \epsilon \tag{12}$$

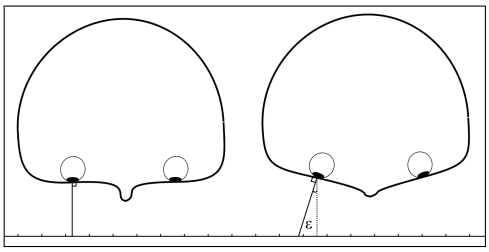


Figure 8: Mask with flat eye region (left) and with angled eye region (right)

6.3. Eye Gaze Experiment

We examined the perception of eye gaze with 23 subjects 7 female and 16 males, with age range of 21-40 years (Mean= 28.4, SD=5.5), each of whom had normal or corrected to normal vision. To evaluate the role of embodiment and presence in perception of agents' eye gaze, four conditions (VA, CR, TR, and GT) were examined in this experiment. In each condition, the agent looked at a particular point on a glass divider located between the agent and the subjects. The subjects were then asked to report their perception of where the agent was looking.

The subjects were seated in front of the glass screen, and then asked to keep their head still on a chin-forehead rest and look straight at the agent at a distance of 120cm. To simulate the most accurate head rotation and avoid a Mona Lisa effect, which is common when viewing a face on a flat screen, in the VA condition we presented rotations of the animated head itself rather than rotations of the screen portraying the head. Figure 9 illustrates the eye gaze evaluation setup.

Fifty-one points, three centimeters apart from each other, were marked by letters and numbers on the glass. However, the agents looked at only five points located at -39, -21, 0, 21 and 39 centimeters (with zero as the middle point of the glass divider). Hereafter, these points are referred to as *A, B, C, D* and *E*, respectively (shown in Fig. 9). Subjects were not aware of the agent's restricted gaze targets, and they were instructed that the agent may look at any points on the glass. Figure. 10 shows photos of different conditions viewed from the subject's position.

We examined the emergent perception of eye gaze (i.e., the integration of head rotation information with eye position). In particular, there were five possible head rotations (-30°, -16°, 0°, 16°, and 30°), and in each head position, the eyes were shifted toward the five points on the glass screen. An example of this condition is shown in Fig. 9, where the agent's head is rotated toward +16° and the eyes are directed at point *B*.

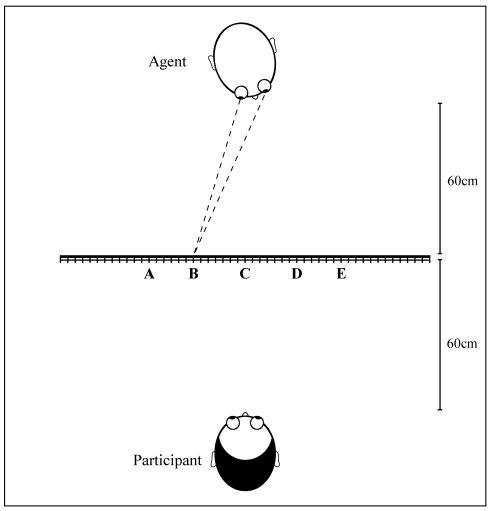
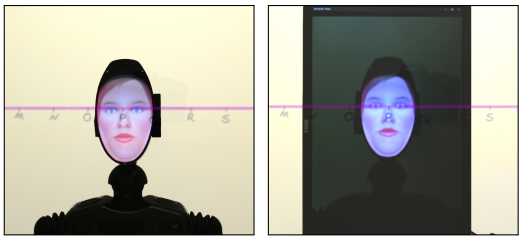
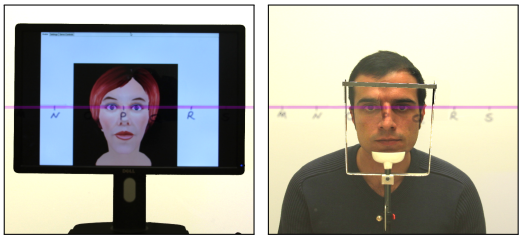


Figure 9: Perception of eye-gaze setup. Fifty-one points with three centimeters distance from each other were marked on the glass. The agents looked at only A, B, C, D, and E points located at -39, -21, 0, 21 and 39 centimeters from the center respectively.



(a) Copresent Robot (b) Telepresent Robot



(c) Virtual Agent (d) Ground-Truth

Figure 10: Eye gaze different conditions

The method described in Section 6.2 was used to calculate the angle for the agent's eyes in CR and TR scenarios. The dimensions of the robot head for CR and the 3D model for VA were measured, and depending on the target point on the glass screen, the eyes of the robot/3D model were rotated toward the target point. The measurement used in CR was: $D = 73cm$, $H = 13.35cm$, $\theta = 13°$, and the measurement used in VA was: $D = 70cm$, $H = 10.45cm$, $\theta = 17°$. Since a mask with a flat eye region was used in CR and a flat screen was used in VA, the value of ϵ was set to 0°.

A Canon EOS 80D DSLR camera was used to take pictures of the robot from the point of view of the subject. The captured pictures were calibrated to the size of the robot head. Using this method, from the point of view of the subject, the agent in both CR and TR had the same size and proportions, and in theory, the same direction of eye gaze (if we took a picture from the subject's point of view, it would look the same). The difference was that the TR condition featured a 2D representation of the CR condition.

To keep the human agent's head in an exact head rotation angle consistently during the GT experiments, we modified a chin/forehead rest to rotate and then stabilize in 1° increments. In the GT condition, a human was seated in the place of the agent and looked at the points on the glass, while keeping his head still on this chin forehead rest and his shoulders facing directly forward.

In all four conditions, first, the agent's head was rotated to one of the five angles (-30°, -16°, 0°, 16°, and 30°) randomly. Then at each of these head angles the eyes were rotated to gaze at one of the 10 points on the board (two trials for the five targets *A, B, C, D* and *E*) randomly. The subject was asked to close his/her eyes between each trial to eliminate any effect of seeing the agent adjust his head and eyes. In total, each subject reported 50 gaze directions (5 angles × 5 points × 2 trials) for each condition. Each condition was run in a block lasting five minutes and the subjects were asked to leave the room for two minutes until the room was setup for the next condition.

Four different agent conditions (VA, TR, CR and GT) were presented in random order to the subjects, and subjects were asked to report their perception of the point at which the agent was looking. Accuracy was calculated by measuring the error in each subject's reports of eye gaze. Gaze perception error was defined as the absolute distance between the point that the subjects reported and the actual target point at which the agent was looking.

6.4. Eye Gaze Results

We performed a 5 (head rotation) × 5 (eye gaze) × 4 (agent conditions: CR, TR, VA and GT) ANOVA with agent condition, head rotation and target point as within-subject factors. The dependent variable was gaze perception error. This analysis revealed a significant main effect of agent condition [$F(3, 66) = 134.55$, $p < .0001$]. We also found main effects of head rotation [$F(4, 88) = 70.25$, $p < .0001$] and eye gaze [$F(4, 88) = 31.39$, $p < .0001$]. This analysis also revealed an interaction between agent condition and head rotation [$F(12, 264) = 11.17$, $p < .0001$], but the interaction between the agent condition and eye gaze was not significant [$F(12, 264) = 95.16$, *n.s*]. Figure 12 shows the estimated marginal means of gaze perception error for different agents, head rotation angle and target points. As shown, differences between the agent conditions depended on head rotation, but not eye gaze.

Table 7 shows the average and standard deviation of error for each condition and proportional error with respect to human ground-truth. The results indicate that eye gaze was better perceived on CR than TR and VA, with 13.21% and 32.23% lower proportional error, respectively. Figure 11 shows the average error (cm) in the perception of different agents' eye gaze for different head rotation and target points. As Fig. 11-(a) shows, when the eye gaze was directly toward the subject's face (point C), the perception of eye gaze had a relatively negligible

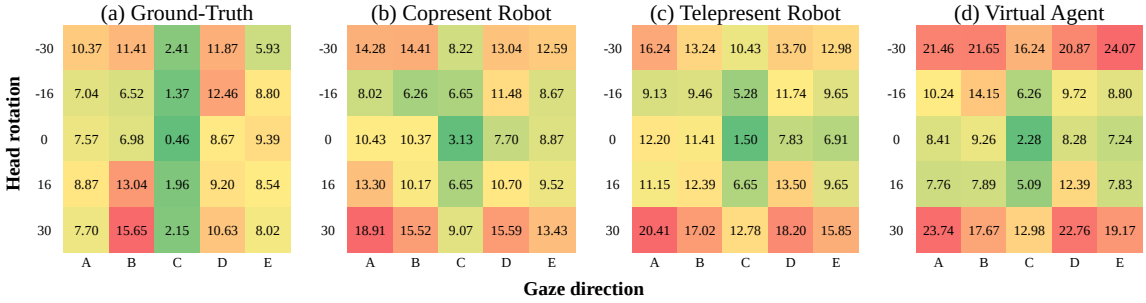
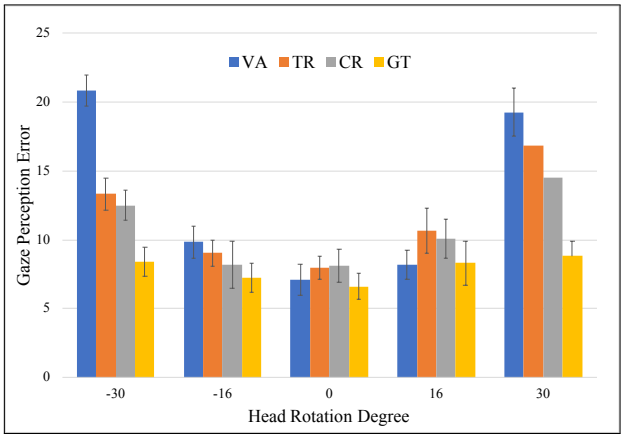


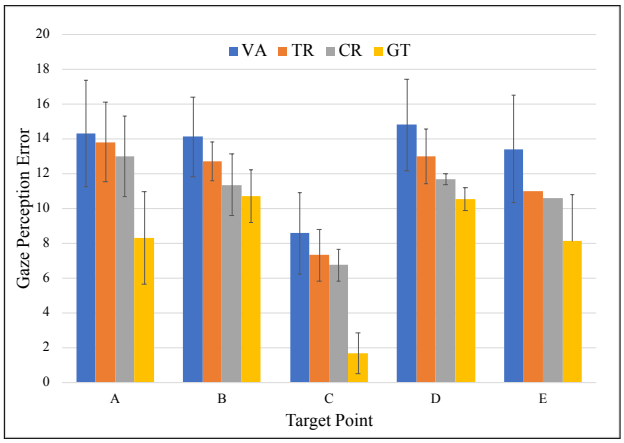
Figure 11: Average absolute error of gaze perception in different conditions [best viewed in color].

Table 7: Average and proportional error with respect to human ground-truth for different agent conditions.

	Average Error ± STD (cm)	Proportional Error to GT
GT	7.88 ± 2.90	-
CR	10.50 ± 3.11	33.26%
TR	11.04 ± 3.16	46.47%
VA	13.04 ± 2.88	65.57%



(a) Head rotation



(b) Target Point

Figure 12: Estimated marginal means of gaze perception error for different agents and (a) head rotation angles and (b) different gaze target points. The target points A, B, C, D corresponds to -39, -21, 0, 21 and 39cm from the center, respectively.

amount of error. In other words, subjects were able to recognize mutual eye contact with high precision on the human agent. The same pattern emerged in the CR and TR conditions. Interestingly, subjects discriminated mutual eye gaze poorly in the VA condition, especially with incongruent head and eye rotations.

Notably, when the head was rotated to its extremes (-30° and 30°), perception of gazes directed toward points *B* and *D* had higher error than gazes directed toward points *A* and *E*. This suggests that subjects had difficulty recognizing gaze direction accurately when the rotation of the head was incongruent with that of the eyes. Hence, subjects may have guessed a point at the far end of the glass screen, which gave them more room for error at points *B* and *D*.

As shown in Fig. 11, eye gaze of the virtual agent was seen with a notable amount of error (~24cm) when combined with a strong head rotation. This could be because the animation lacked binocular depth cues by virtue of being present on a flat screen. This could have made the perception of head rotation more difficult, while the embodiment of the robot helped subjects to recognize the head angle better.

In order to more directly measure the effect of agents' embodiment and presence, we removed human GT from the analysis and performed a 5 (head rotation) × 5 (eye gaze) × 3 (agent conditions: CR, TR, VA) ANOVA with agent condition, head rotation and eye gaze as within-subject factors. This analysis revealed main effects of agent [$F(2, 44) = 8.740$, $p = .001$], head rotation [$F(4, 88) = 64.95$, $p < .001$] and eye gaze [$F(4, 88) = 16.39$, $p < .0001$]. Similar to previous analysis, and as shown in Figure 12, there was a significant interaction between the agent condition and head rotation [$F(8, 176) = 8.75$, $p < .0001$], but the interaction between the agent condition and eye gaze was not significant [$F(8, 176) = 23.98$, *n.s*].

Since there was an interaction between the agent condition and head rotation, we performed pairwise two-tailed t-test comparisons between agent conditions at different head rotations. Table 8 shows pairwise *p*-value and Cohen's *d* effect-size between agent conditions. As shown, embodiment (Research Question 1) improved the perception of eye gaze at -30° and 30°, as indexed by significant differences between TR and VA conditions ($p < .001$ and $p = .023$ with large effect sizes $d = 1.22$ and $d = 0.69$ respectively). Physical presence did not improve the perception of eye gaze (Research Question 2), as the differences between TR and CR conditions were not significant at any head angle. There were also significant differ-

Table 8: Pairwise comparison (LSD p-value) and Cohen's d effect size of users' perception of eye gaze at different head rotations. Significant pairs are shown in bold.

	TR vs CR		VA vs CR		VA vs TR	
Head Angle	p	d	p	d	p	d
-30°	.660	0.13	**<.001**	**1.49**	**<.001**	**1.22**
-16°	.479	0.21	.190	0.39	.5484	0.17
0°	.890	0.04	.278	0.32	.269	0.32
16°	.599	0.15	.116	0.47	.158	0.42
30°	.217	0.36	**.004**	**0.89**	**.023**	**0.69**

ences between CR and VA at -30° and 30°, both $p < .001$ with large effect sizes $d = 1.49$ and $d = 0.89$ respectively (Research Question 3). Because TR and VA were both significantly different at these head angles, we conclude that improvement in the perception of eye gaze compared with CR is mainly due to embodiment rather than presence of the robot. And in particular, embodiment of the robot highly affected the precision of the gaze perception combined with extreme head rotations in a frontal situated setting.

These findings are congruent with previous studies showing that the perception of a robot's eye gaze is more accurate than that of a virtual agent (Al-Moubayed et al., 2012; Misawa et al., 2012; Mollahosseini et al., 2014). There was no difference in perception of gaze when seen on a robotic agent or its telepresence, which is consistent with a study performed by Delaunay et al. (2010). We also did not observe a significant difference between gaze perception on the telepresent robot and virtual agents—a comparison which has not been addressed in previous studies.

7. Conclusion

This work examines the role of social robots' embodiment and presence in users' perception of facial cues using a quantitative approach. Understanding how people respond to physical and virtual agents is an important factor in designing successful social agents. Three research questions as the effect of physical embodiment (Q1), physical presence (Q2), and the joint effect of physical embodiment and presence (Q3), on human perception of agents' facial cues (visual speech, facial expressions and eye gaze) were studied in this research. To study these effects, we leveraged three different agent conditions (i.e., copresent robot, telepresent robot, and virtual agent) as well as human ground truth to evaluate the optimal case in our settings. The results of this study indicate that:

1. There was no evidence that embodiment or presence improves the perception of visual speech, regardless of syntactic or semantic cues in sentences.
2. Both embodiment and physical presence improve the perception of certain facial expressions in emotive agents.
3. The combination of embodiment and presence (and mainly embodiment) highly affects the precision of eye gaze perception in a frontal situated setting.

Comparison of our findings with previous studies also indicates that the type of embodiment is important. We used a retro-projected robotic head in this study, which has some limitations (e.g., the mask is static, the jaw and lip movements are only optical). We believe that the limitations of embodiment can highly affect the perception of social cues. For example, the static jaw and optical lip movement may affect the perception of visual speech on the retro-projected robotic, and hence there was no significant effect of embodiment or presence in visual speech. Also, since the jaw does not move, the perception of a surprised expression on the video of the robot was significantly lower than the virtual agent. In addition, the eye movement on the retro-projected robotic head is also optical and the eyeballs do not rotate on the static mask. Our results showed a significant interaction between the agent condition and the head rotation, while the interaction of the agent condition and the eye gaze was not significant. This might be again due to the limitation on embodiment, which can be validated by comparing two different embodiments (e.g., an Android robot v.s. a retro-projected robot) in future studies.

Naturally, each type of embodiment has its own limitations. For instance, mechanical or Android robotic heads are limited by the number of actuators used in their face preventing them from showing accurate visual speech or certain facial expressions. Therefore, the findings of any investigations on the role of embodiment and presence cannot necessarily be generalized to other types of robotic embodiments, without considering the characteristics of the embodied agents.

Acknowledgments

This work is partially supported by the NSF grants IIS-1111568 and CNS-1427872 and a DU Professional Research Opportunity for Faculty (PROF) grant. The 3D animation system used to control the Ryan model was developed jointly by Boulder Learning Inc. (http://www.boulderlearning.com) and the Computer Vision and Social Robotics Laboratory at the University of Denver.

References

Al Moubayed, S., Beskow, J., Skantze, G., Granström, B., 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In: Cognitive behavioural systems. Springer, pp. 114–130. 5

Al-Moubayed, S., Edlund, J., Beskow, J., 2012. Taming mona lisa: Communicating gaze faithfully in 2d and 3d facial projections. ACM Transactions on Interactive Intelligent Systems (TiiS) 1 (2), 11. 12, 16

Al Moubayed, S., Skantze, G., 2012. Perception of gaze direction for situated interaction. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. ACM, p. 3. 12

Al Moubayed, S., Skantze, G., Beskow, J., 2013. The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. International Journal of Humanoid Robotics 10 (01), 1350005. 1, 2, 5, 6, 7

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the sts region. Trends in cognitive sciences 4 (7), 267–278. 11

Anstis, S. M., Mayhew, J. W., Morley, T., 1969. The perception of where a face or television'portrait'is looking. The American journal of psychology 82 (4), 474–489. 11, 12

Bainbridge, W. A., Hart, J. W., Kim, E. S., Scassellati, B., 2011. The benefits of interactions with physically present robots over video-displayed agents. International Journal of Social Robotics 3 (1), 41–52. 1, 2, 3

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., Walker, J., 1995. Are children with autism blind to the mentalistic significance of the eyes? British Journal of Developmental Psychology 13 (4), 379–398. 11

Bartneck, C., Lyons, M., 2007. Hci and the face: towards an art of the soluble. Human-Computer Interaction. Interaction Design and Usability, 20–29. 4

Bartneck, C., Reichenbach, J., Breemen, v. A., 2004. In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In: Proceedings of the Design and Emotion. pp. 32–51. 8, 9, 10

Becker-Asano, C., Ishiguro, H., April 2011. Evaluating facial displays of emotion for the android robot geminoid f. In: 2011 IEEE Workshop on Affective Computational Intelligence (WACI). pp. 1–8. 7

Beer, J. M., Prakash, A., Mitzner, T. L., Rogers, W. A., 2011. Understanding robot acceptance. Georgia Institute of Technology, 1–45. 7

Bilger, R. C., Nuetzel, J., Rabinowitz, W., Rzeczkowski, C., 1984. Standardization of a test of speech perception in noise. Journal of Speech, Language, and Hearing Research 27 (1), 32–48. 6

Biocca, F., 1997. The cyborg's dilemma: Progressive embodiment in virtual environments. Journal of Computer-Mediated Communication 3 (2), 0–0. 3

Bolanos, D., 2012. The bavieca open-source speech recognition toolkit. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, pp. 354–359. 5, 6

Breazeal, C., 2003. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies 59 (1), 119–155. 7

Breazeal, C., 2005. Socially intelligent robots. interactions 12 (2), 19–22. 1

Breazeal, C. L., 2000. Sociable machines: Expressive social exchange between humans and robots. Ph.D. thesis, Massachusetts Institute of Technology. 4

Bruce, A., Nourbakhsh, I., Simmons, R., 2002. The role of expressiveness and attention in human-robot interaction. In: Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on. Vol. 4. IEEE, pp. 4138–4142. 7

Cassell, J., 2000. Embodied conversational interface agents. Communications of the ACM 43 (4), 70–78. 3, 7, 11

Chen, Y.-C., Yeh, S.-L., 2012. Look into my eyes and i will see you: Unconscious processing of human gaze. Consciousness and cognition 21 (4), 1703–1710. 11

Cline, M. G., 1967. The perception of where a person is looking. The American journal of psychology 80 (1), 41–50. 12

Cohen, J., 1977. Statistical power analysis for the behavioral sciences (revised ed.). 10

Cronbach, L. J., 1951. Coefficient alpha and the internal structure of tests. psychometrika 16 (3), 297–334. 2

Dautenhahn, K., 1998. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. Applied artificial intelligence 12 (7-8), 573–617. 2

Dautenhahn, K., 2001. Socially intelligent agents-the human in the loop. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 31 (5), 345–348. 3

Dautenhahn, K., 2007. Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences 362 (1480), 679–704. 1

Delaunay, F., de Greeff, J., Belpaeme, T., 2010. A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In: Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction. IEEE Press, pp. 39–44. 1, 2, 12, 16

DreamFace-Tech., 2015. Social robotics. Last checked: 01.20.2017. URL http://dreamfacetech.com/ 4

Ekman, P., Friesen, W., 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto. 4, 8

Elliott, L. L., 1995. Verbal auditory closure and the speech perception in noise (spin) test. Journal of Speech, Language, and Hearing Research 38 (6), 1363–1376. 6

Emery, N. J., 2000. The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews 24 (6), 581–604. 11

Friesen, W. V., Ekman, P., 1983. Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco 2, 36. 8

Fujimura, R., Nakadai, K., Imai, M., Ohmura, R., 2010. Prot—an embodied agent for intelligible and user-friendly human-robot interaction. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. IEEE, pp. 3860–3867. 1, 2, 3

Goffman, E., 1963. Behavior in public place. Glencoe: the free press, New York. 3

Guizzo, E., 2010. World robot population reaches 8.6 million. IEEE Spectrum 14. 1

Hartholt, A., Gratch, J., Weiss, L., et al., 2009. At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In: International Workshop on Intelligent Virtual Agents. Springer, pp. 500–501. 1

Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., Picard, R. W., 2013. Mach: My automated conversation coach. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. ACM, pp. 697–706. 1

IDC, 2016. International data corporation (idc) press. Last checked: 05.14.2017. URL http://www.idc.com/getdoc.jsp?containerId=prUS41046916 1

Imai, M., Kanda, T., Ono, T., Ishiguro, H., Mase, K., 2002. Robot mediated round table: Analysis of the effect of robot's gaze. In: Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on. IEEE, pp. 411–416. 11

IPA-Handbook, I. P. A., 1999. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press. 5

Itier, R. J., Batty, M., 2009. Neural bases of eye and gaze processing: the core of social cognition. Neuroscience & Biobehavioral Reviews 33 (6), 843–863. 11

Ju, W., Sirkin, D., 2010. Animate objects: How physical motion encourages public interaction. In: International Conference on Persuasive Technology. Springer, pp. 40–51. 1, 2, 3

Kajita, S., Nakano, T., Goto, M., Matsusaka, Y., Nakaoka, S., Yokoi, K., 2011. Vocawatcher: Natural singing motion generator for a humanoid robot. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, pp. 2000–2007. 4

Kalikow, D. N., Stevens, K. N., Elliott, L. L., 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. The Journal of the Acoustical Society of America 61 (5), 1337–1351. 6

Kätsyri, J., Sams, M., 2008. The effect of dynamics on identifying basic emotions from synthetic and natural faces. International Journal of Human-Computer Studies 66 (4), 233–242. 8, 9

Kendon, A., 1967. Some functions of gaze-direction in social interaction. Acta psychologica 26, 22–63. 11

Kendon, A., Harris, R. M., Key, M. R., 1975. Organization of behavior in face-to-face interaction. Walter de Gruyter. 2

Kidd, C. D., Breazeal, C., Sept 2004. Effect of a robot on user perceptions. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566). Vol. 4. pp. 3559–3564 vol.4. 1, 2, 3

Kiesler, S., Powers, A., Fussell, S. R., Torrey, C., 2008. Anthropomorphic interactions with a robot and robot–like agent. Social Cognition 26 (2), 169–181. 2, 3

Kinya, M., Mitsuo, E., 1984. Illusory face dislocation effect and configurational integration in the inverted face. Tohoku Psychologica Folia 43 (1-4), 150–160. 12

Kluttz, N. L., Mayes, B. R., West, R. W., Kerby, D. S., 2009. The effect of head turn on the perception of gaze. Vision research 49 (15), 1979–1993. 12

Kopp, S., Gesellensetter, L., Krämer, N. C., Wachsmuth, I., 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In: International Workshop on Intelligent Virtual Agents. Springer, pp. 329–343. 1

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., Nehaniv, C. L., 2009. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. Advanced Robotics 23 (14), 1951–1996. 1, 2, 3

Langton, S. R., Honeyman, H., Tessler, E., 2004. The influence of head contour and nose angle on the perception of eye-gaze direction. Perception & psychophysics 66 (5), 752–771. 12

Lazzeri, N., Mazzei, D., Greco, A., Rotesi, A., Lanatà, A., De Rossi, D. E., 2015. Can a humanoid face be expressive? a psychophysiological investigation. Frontiers in bioengineering and biotechnology 3. 8, 9, 11

Lee, K. M., Jung, Y., Kim, J., Kim, S. R., 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. International Journal of Human-Computer Studies 64 (10), 962–973.

1, 2, 3

Lester, J. C., Voerman, J. L., Towns, S. G., Callaway, C. B., 1999. Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. Applied Artificial Intelligence 13 (4-5), 383–414. 5

Li, J., 2015. The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human-Computer Studies 77, 23–37. 2, 3

Li, J., Chignell, M., 2011. Communication of emotion in social robots through simple head and arm movements. International Journal of Social Robotics 3 (2), 125–142. 2

Lin, C.-Y., Cheng, L.-C., Shen, L.-C., 2013. Oral mechanism design on face robot for lip-synchronized speech. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, pp. 4316–4321. 4

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, pp. 94–101. 9

Luo, R. C., Chang, S.-R., Huang, C.-C., Yang, Y.-P., 2011. Human robot interactions using speech synthesis and recognition with lip synchronization. In: IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society. IEEE, pp. 171–176. 4

Ma, J., Cole, R., 2004. Animating visible speech and facial expressions. The Visual Computer 20 (2-3), 86–105. 4, 5

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748. 4

Milgram, P., Takemura, H., Utsumi, A., Kishino, F., 1995. Augmented reality: A class of displays on the reality-virtuality continuum. In: Photonics for industrial applications. International Society for Optics and Photonics, pp. 282–292. 3

Misawa, K., Ishiguro, Y., Rekimoto, J., 2012. Livemask: A telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. In: Proceedings of the international working conference on advanced visual interfaces. ACM, pp. 394–397. 12, 16

Mollahosseini, A., Graitzer, G., Borts, E., Conyers, S., Voyles, R. M., Cole, R., Mahoor, M. H., 2014. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In: Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on. IEEE, pp. 1098–1103. 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 16

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N., 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction. ACM, pp. 61–68. 11

Nishio, S., Ishiguro, H., Hagita, N., 2007. Geminoid: Teleoperated android of an existing person. INTECH Open Access Publisher Vienna. 4

Oh, K.-G., Jung, C.-Y., Lee, Y.-G., Kim, S.-J., 2010. Real-time lip synchronization between text-to-speech (tts) system and robot mouth. In: RO-MAN, 2010 IEEE. IEEE, pp. 620–625. 4

Otsuka, Y., Mareschal, I., Calder, A. J., Clifford, C. W., 2014. Dual-route model of the effect of head orientation on perceived gaze direction. Journal of Experimental Psychology: Human perception and performance 40 (4), 1425. 12

Ouni, S., Massaro, D. W., Cohen, M. M., Young, K., Jesse, A., 2003. Internationalization of a talking head. In: Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain. pp. 286–318. 5, 6

Pateromichelakis, N., Mazel, A., Hache, M., Koumpogiannis, T., Gelin, R., Maisonnier, B., Berthoz, A., 2014. Head-eyes system and gaze analysis of the humanoid robot romeo. In: Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on. IEEE, pp. 1374–1379. 11

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., Jeeves, M., 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. Proceedings of the Royal Society of London B: Biological Sciences 223 (1232), 293–317. 11

Pfeifer, R., Scheier, C., 1999. Understanding intelligence. MIT press. 3

Ruhland, K., Andrist, S., Badler, J., Peters, C., Badler, N., Gleicher, M., Mutlu, B., Mcdonnell, R., 2014. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In: Eurographics State-of-the-Art Report. pp. 69–91. 11

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L., 2008. Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management. Springer, pp. 47–56. 8

Siciliano, C., Faulkner, A., Williams, G., 2003. Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In: AVSP 2003-International Conference on Audio-Visual Speech Processing. pp. 205–208. 5, 6

Sweeny, T. D., Guzman-Martinez, E., Ortega, L., Grabowecky, M., Suzuki, S., 2012a. Sounds exaggerate visual shape. Cognition 124 (2), 194–200. 4

Sweeny, T. D., Haroz, S., Whitney, D., 2012b. Reference repulsion in the categorical perception of biological motion. Vision research 64, 26–34. 11

Sweeny, T. D., Whitney, D., 2017. The center of attention: Metamers, sensitivity, and bias in the emergent perception of gaze. Vision Research 131, 67–74. 12

Todorović, D., 2006. Geometrical basis of perception of gaze direction. Vision research 46 (21), 3549–3562. 11

Vala, M., Sequeira, P., Paiva, A., Aylett, R., 2007. Fearnot! demo: a virtual environment with synthetic characters to help bullying. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. ACM, p. 271. 1

Van-Breemen, A., 2004. Bringing robots to life: Applying principles of animation to robots. In: Proceedings of Shapping Human-Robot Interaction workshop held at CHI 2004. Citeseer, pp. 143–144. 7, 8

Wainer, J., Feil-Seifer, D. J., Shell, D. A., Mataric, M. J., 2007. Embodiment and human-robot interaction: A task-based perspective. In: Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on. IEEE, pp. 872–877. 2, 3

Walker, J. H., Sproull, L., Subramani, R., 1994. Using a human face in an interface. In: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp. 85–91. 5

Wollaston, W. H., 1824. On the apparent direction of eyes in a portrait. Philosophical Transactions of the Royal Society of London 114, 247–256. 12

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., Miyamoto, T., 2006. Responsive robot gaze to interaction partner. In: Robotics: Science and systems. 11

Zhao, S., 2003. Toward a taxonomy of copresence. Presence: Teleoperators and Virtual Environments 12 (5), 445–455. 3