# Gaze Tracking for Human Robot Interaction

Oskar Palinko

A thesis submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Supervisors: Alessandra Sciutti, Giulio Sandini


Doctoral Course in Cognitive Robotics, Interaction and Rehabilitation Technologies

Doctoral Program in Bioengineering and Robotics

Università degli Studi di Genova

&

Robotics, Brain and Cognitive Sciences Department

Istituto Italiano di Tecnologia

March 2017.

## Abstract

Humans communicate with each other very naturally. They are very much able to sense and interpret even the subtlest cues which then modulate the flow of the interaction between them. This allows for very natural and seamless human-to-human interchanges. For example we are very capable even to detect distress from the trembling of someone's voice, or happiness from a multitude of body movements.

Unfortunately, robots are not as proficient in communicating with humans as other humans are. An important cause for this is that robots are not as good at sensing subtle cues as humans are. Sometimes they can't even perceive the signals or they can't interpret the perceived signals. For example robots have a very hard time detecting pupil size, which to humans signifies increased attention or arousal by the collocutor. On the other hand, a robot can perceive a certain arrangement of facial features of the human partner but it might not be sophisticated enough to interpret the displayed emotion. This is especially true with robots perceiving implicit communication signals like gaze. For this reason we are proposing to provide them with the ability to read the gaze of humans by implementing gaze tracking algorithms on humanoid robots.

In order to enable the widespread use of gaze in human-robot interaction (HRI) it would be very beneficial to have affordable ocular-visual systems on robots. We addressed this issue in Study I which describes a fairly advanced visual subsystem using affordable off-the-shelf components [1].

One of the very important gaze cues is mutual gaze: when two agents are looking at each other's eyes. In Study II we set out on the path of implementing a full gaze tracker by first building its subset: a mutual gaze detector [2]. We successfully used this detector to enable an iCub robot to dictate text to its students (experiment participants) with gaze-contingency. We found that reacting to gaze can be a very useful tool for a teacher robot to adapt to its students' needs, thus making the interaction more pleasant [3].

Commercially available gaze tracking systems are mostly created for human computer interaction (HCI) environments. Since robotic platforms have quite specific constraints, these regular HCI gaze tracking solutions can be used on humanoid robots only with limited results. This is why in Study III we set out to design an eye tracker specifically created for robotic applications, focusing on mobility, embeddedness, human-likeness, lightning invariance, low

cost, and user independence [4]. We found that the newly designed system performs well enough to successfully support collaborative HRI scenarios like building a tower out of toy building blocks.

Once the eye tracking system was implemented, in Study IV we decided to explore how useful such a system was by testing the empirically established concept in HRI literature that eye gaze can be easily replaced by its first proxy, head pose, without excessive loss of information [5]. For this purpose we created two conditions in our tower building scenario: one activated by the participants' eye gaze and another by their head pose. It was found that eye gaze worked much better both according to objective and subjective measures. This proved to us that HRI would definitely benefit from a suitable gaze tracking system as compared to head pose tracking only.

Finally in Study V instead of a collaborative we focused on a competitive task to see how humans and robots perform in such environments [6]. Namely we included multiple humans and robots in playing a gaze-based social game called the "Wink Murder". It was found that robots can be programmed to win at playing this game against novice humans. We also observed many different behavior patterns of humans while competing with robots. This kind of a scenario has the potential to further enrich our understanding of gaze communication between humans and humanoid robots.

In summary in this dissertation we are shedding light on the possible role of gaze tracking in aiding and augmenting communication between humans and robots. Our newly implemented robotic head, mutual gaze detector and eye tracker are the software and hardware tools for achieving this goal, while our collaborative and competitive HRI scenarios and experiments provide the data for our claims. Finally we find that robots enabled with gaze reading do provide a more efficient and natural interaction between humans and robots.

# Table of Contents

# 1. Introduction

Humans are very efficient collaborators, able to rapidly coordinate with each other, often with no need of detailed verbal instructions. This efficiency derives from the use of a wealth of communication cues to guide interaction, both explicit (as for instance gestures or speech) and implicit (as some elements of gaze). Implicit communication signals are those, which are not intended to carry information, but they do anyways and are fundamental for effective communication. For instance, when humans want to reach for an object, their gaze anticipates their hand on target. This implies that keen observers can predict the goal of their partner even before the beginning of the hand motion, just by looking at their eyes. When people turn their gaze to gather information, their eyes immediately give off where their visual attention is focused at and hence which object they want to take.

Communication between people during daily activities relies on a series of multimodal cues, as speech, gestures, pointing, etc. among which gaze is one of the most important [7]. This is particularly evident when eye gaze processing is atypical, as in individuals with autism spectrum disorders, where such impairment is linked to social and communicative deficits [8]. In fact, during social interaction we are not always aware that besides acquiring visual stimuli (sensing) we are also transmitting information with our eyes (acting). This information is used by our partners to detect the focus of our attention and to modulate turn taking for example. Let's look at an example: Mary and Jane are talking to each other. At one point Mary looks up for a short time (gaze aversion). Jane realizes that Mary needs additional time to think, thus waits patiently. At a different moment during the conversation Jane looks at a magazine in front of her, which causes Mary to realize that the object of attention has shifted towards the observed item (gaze pointing, joint attention). If Peter joins the conversation, Jane will be able to understand when she is addressed as opposed to the new collocutor if Mary's gaze is fixated on her instead of on Peter (mutual gaze). All these interaction examples are facilitated by observing the gaze of others. Robots could greatly benefit from understanding the gaze of their human partners. On one hand, understanding such an implicit communication cue makes robots become more aware of others' intentions; on the other hand, it provides the partner with immediate evidence that the robot interprets the interaction correctly. Both abilities concur to promote a more natural relationship. This is particularly true for humanoid robots, since the humanoid shape might induce humans to automatically assume that the robot's visual perception will be similar to their own. A similar, unconscious assumption will increase human expectations that the robot will appropriately

respond to usual gaze behaviors, for instance by following their gaze toward the object they are attending to.

Recently the importance of communication through gaze has been acknowledged also in robotics, even beyond the boundaries of purely social applications. For instance, in the field of small scale manufacturing, one of the key selling points of Baxter (Rethink Robotics) is its ability to seamlessly communicate its focus of attention thanks to its "eyes", which make it easily understandable by non-trained collaborators. However, while the opportunity of using robot eyes to communicate has been already applied in the market, the possibility for a robot to observe humans' eyes to anticipate their needs and intentions has not been widely used yet.

One reason for this lack of gaze tracking in robots might be the need for specific camera properties to calculate eye gaze direction. In particular, high resolution, narrow field-of-view images are ideally required for such a computation, while robots are in general equipped with wide field-of-view cameras to be able to move and interact in a large environment. Some robots are also limited to lower resolution cameras for different reasons: for example when network bandwidth utilization is prioritized for real time behavior (walking, balancing, etc.), and not for visual processing. Light and shadow can affect this calculation as well. An alternative possibility is to use ad hoc hardware. However, standard tabletop gaze tracking devices are usually designed to be static, observing just one spot in space: the user in front of a display screen. On the other hand robots often need a system that can deal with agents moving in space. The common solution adopted in experimental settings is the use of head-mounted systems worn by the human partner. These are moving with the subject, but are intrusive and require that anyone wanting to interact with a robot wear special glasses or a helmet. This approach often limits the adoption of eye tracking in open environments (e.g. airports, shopping malls, hospitals, etc.) where robots could be required to interact with people with no prior preparation of the human partners.

Because of the above problems concerning retrieving gaze in human-robot interaction (HRI) a number of authors (see Chapter 2) have resorted to using the so called "head gaze" instead of eye gaze. This choice was mostly made because head orientation is easier to compute as it is a much larger image feature than the eyes. But the problem is that eye gaze does not always coincide with head gaze. People can make short glances at objects without moving their heads (e.g. checking the time on a wrist watch or glancing at a secondary screen). Indeed, eye gaze contains more information than head orientation only. Also for humans it has been proved that actual gaze direction estimation is significantly more precise when it is based also on eyes with respect to head only [9]. Moreover, in natural collaborative scenarios the objects are often close to each

other and people tend to switch their focus of attention just by moving their eyes, yielding to minor or null head movements. The inability to read actual eye movements could then make the robot miss important information for an efficient interaction, like which object the human collaborator attends to.

This dissertation is organized into chapters based on the studies we completed in order to investigate the use and relevance of gaze in HRI. Study I (An affordable active robot head for gaze tracking) looks at how to design a gaze tracking head system which addresses the specific needs of humanoid robot platforms [1]. Study II (A gaze-contingent robot for a dictation scenario) discusses a dictation scenario between a human and a humanoid which is facilitated by mutual gaze detection [3]. Study III (Design and verification of an eye tracker) addresses the issue of how to design a gaze tracking system specifically aimed at human-robot interaction [4]. Study IV (Comparing eye gaze to head pose for HRI) explores if gaze tracking would be more useful than simple head pose tracking for completing a collaborative task [5]. Study V (A gaze-based social game for humans and robots) explores how gaze can be used in a competitive social game scenario with two humans and two android robots [6].

## 1.1 Problems

The general problem we are addressing in this thesis is that the communication between humans and robots is not as natural as between humans. Robots today are unable to read some subtle cues which humans use in their everyday interaction. One of these cues is the gaze. The more specific problems we are addressing are as follows:

- Humanoid robot head solutions which can accommodate for gaze tracking of humans are either not available or they are very expensive, preventing the wider acceptance of gaze reading in HRI.
- Readily available eye tracking solutions do not address the specific needs of human-robot interaction environments.
- Robots can have problems in deciding when to take initiative in turn taking while interacting with humans, as for example in a tutor-student scenario.
- It is not clear if the replacement of gaze tracking with its first proxy, head pose tracking, would negatively influence the efficiency of communication between humans and robots.
- It is not known whether gaze cues could be used also to coordinate interactions involving more than two partners and in competitive, rather than collaborative, situations.

## 1.2 Hypotheses

The general hypothesis we are proposing is that human-robot interaction can be improved by enabling robots to read humans' gaze. The more specific hypotheses as related to specific studies are as follows:

- An affordable humanoid robot head could make gaze reading a pervasive technology in human-robot interaction.
- A new calibration-free, visual light, monocular gaze tracking algorithm could solve the lack of gaze reading in HRI.
- Gaze reading, and more specifically mutual gaze detection could improve the performance of robots in turn taking situations.
- Gaze tracking will provide more information and a smoother interaction between humans and robots than using head pose tracking only.
- Gaze tracking can be used in social game scenarios to enable an engaging interaction between humans and humanoids which could shed light on different communication strategies between these two types of agents.

## 1.3 Goals

The general goal of my research is to make human-robot interaction more efficient and more natural by enabling robots to read the gaze of humans. The specific goals as related to the separate studies are:

- Designing a robot head which would allow robots to track gaze in an affordable way.
- Designing a humanoid robot specific gaze tracking system which would allow the robots to read humans' gaze. Testing the accuracy of such a system.
- Assessment of improvement in HRI between a gaze-contingent robot behavior and a rhythmic behavior of the robot for turn taking. For this we developed and tested a dictation scenario in which the robot is the teacher while the human is the student.
- Assessment of potential improvements in HRI using gaze tracking compared to head tracking only. For this we design a tower building scenario and test human subjects with the robot.
- Designing and testing a gaze-based social game scenario between robots and humans with the goal of identifying robot performance and human behavior in such a novel environment.

## 1.4 Research platforms

We used three separate robotic platforms in the research we are reporting on in this thesis: the proof-of-concept humanoid robot head created in Study I, the iCub humanoid robot and the Actroid-F android robot. The first system will be explained in detail in Chapter 3, while the other two will be presented in the following section.

### 1.4.1 iCub

The iCub is a child-sized humanoid robot [10]. It has sophisticated ocular-visual and motor systems. It is capable of performing complex movements with its eyes, head, torso, arms, hands, fingers and legs. The control can be based on direct or inverse kinematics [11]. Its movements can also be actively compliant by using its force sensors for feedback control located in its arms [12]. The iCub has especially sophisticated hands and arms with 5 actuated fingers and 9 degrees of freedom (DOF) per hand. The hand and forehand has additional 3 DOFs in both arms.

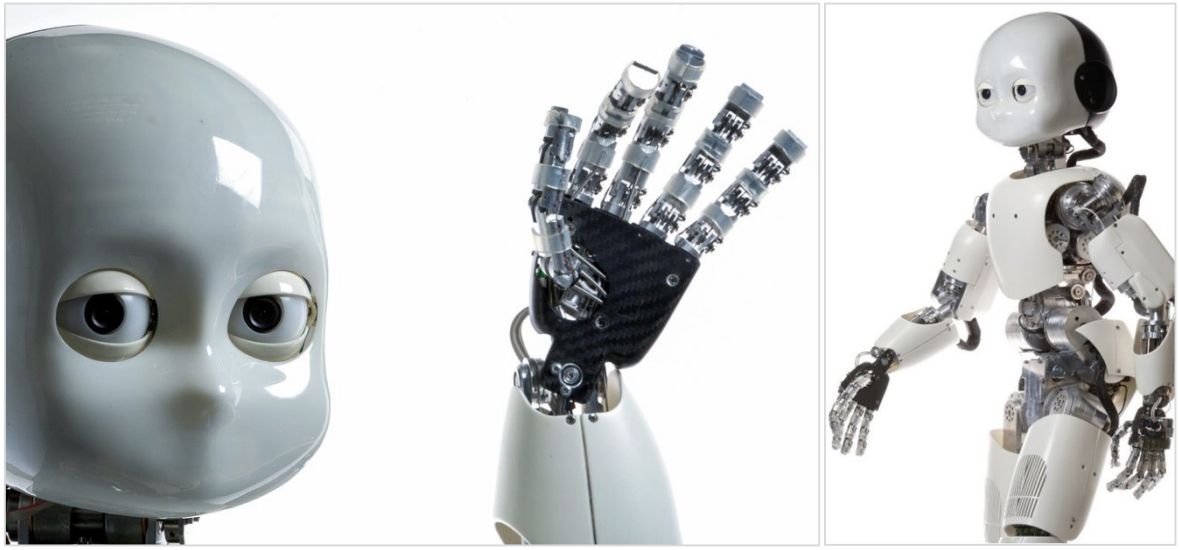

**Figure 1. The iCub humanoid robot.**

The ocular motor system features 3 DOFs for both eyes: pitch, yaw and vergence. The head has pitch, yaw, and roll rotations which is an additional 3 DOFs [13]. Each joint can be controlled directly with minimum jerk trajectory or inversely by giving the robot's visual system a 3D target of gaze [14]. The robot is able to emulate the vestibulo-ocular reflex which makes its movements seem even more natural.

The original iCub camera system consists of two PointGrey Dragonfly 2 cameras with VGA resolution (640 x 480 pixels) built into the eyeballs of iCub. This system was used in our Study II, where the iCub was detecting mutual gaze with humans. For our eye tracking purposes in Study III and Study IV these cameras were of too low resolution, thus we needed to switch to a

camera system of the same type but with higher resolution which were also installed in the eyeballs of the robot.

**Table 1. Technical characteristics of iCub's built-in cameras used in three studies.**

| | Study II | Study III & Study IV |
|---|---|---|
| *Manufacturer* | PointGrey Inc. | PointGrey Inc. |
| *model* | Dragonfly 2 | Dragonfly 2 |
| *Resolution* | 648 x 488 | 1032 x 776 |
| *Frame rate* | 30 FPS | 30 FPS |
| *Focal length* | 4mm | 4mm |

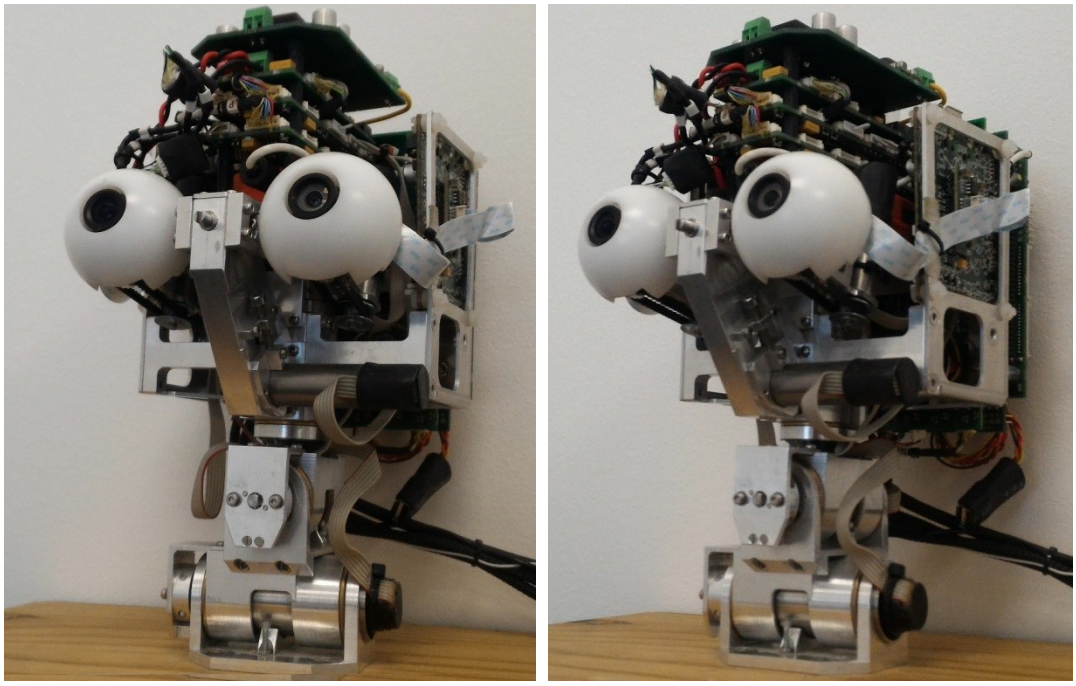

**Figure 2. iCub's head and ocular system (as seen from two angles).**

### 1.4.2 Actroid-F

The Actroid-F is an android platform, which means that the robot is extremely human-looking [15]. It has human body size proportions, realistic looking skin, human hair wig, etc. It is actuated pneumatically, which provides a passive compliance to the robot. This means that it is able to withstand external force disturbances with no need of an ad hoc control, i.e. if somebody pushes it, the pushed DOF will passively retract to avoid breaking. The Actroid-F does not have feedback control of its joints, thus joint positions are not as accurate as in the iCub. These androids have limited body movement capabilities with only 2 DOFs: torso pitch and right arm elbow pitch. On the other hand they have sophisticated head and face movements: 3 DOFs of head, eye pitch, eye yaw, mouth open-close, mouth stretch (smile) and eyebrow up-down. There are two robots of this exact type at the Robot Innovation Center, AIST, Tsukuba, Japan: one female and one male. They were both used in our experiment in Study V.

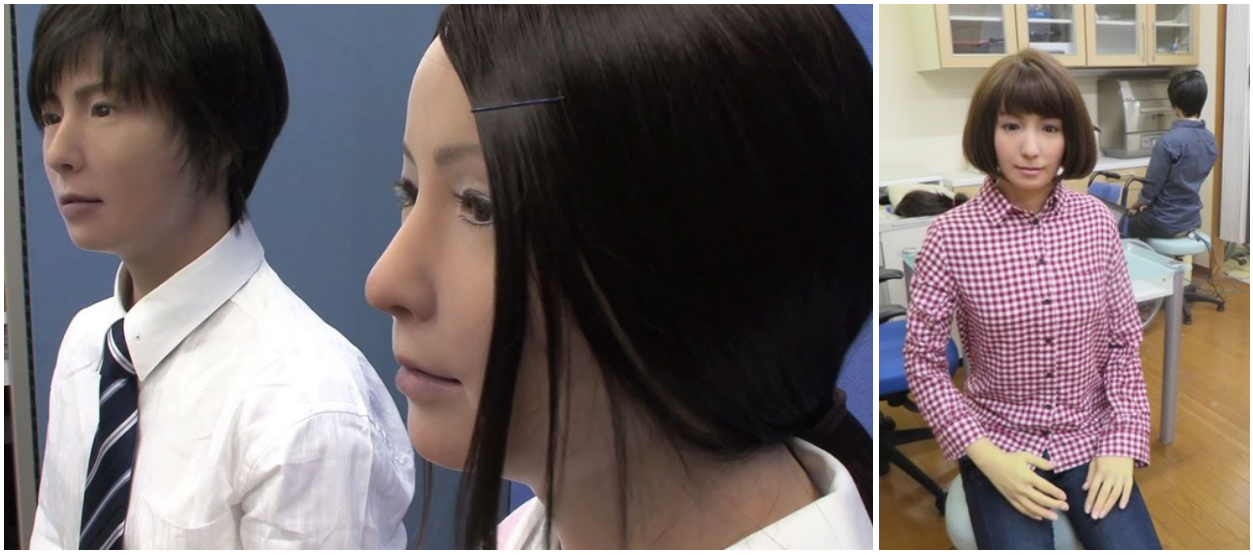

**Figure 3. The Actroid-F android robots.**

The visual system of the Actroid is based on two NCM13-J cameras installed also in the eyeball of the robot. The eye movements of these robots are very human like, but the accuracy of the movement can vary, because of the pneumatic drive. Unlike the iCub, the cameras of the Actroids are located behind a plastic transparent cover of the pupils (see Figure 4), which leads to some degradation of the camera images. This mostly appears as non-homogeneous blurring. The androids' cameras work quite well in VGA resolution (640 x 480 pixels) at 30 FPS. For gaze tracking purposes at an interaction distance (80cm-120cm) this resolution is not satisfactory. On the other hand the next higher resolution setting (1280 x 1024) produces much lower framerate: around 10 FPS. Additionally, the acquisition of these larger images produces considerable motion blurring. In this situation, our gaze tracking algorithm was still able to detect and track faces, but iris center detection suffered more, because of the small apparent size of the eyes in the camera's images. Yet another limiting factor for this camera was its relatively narrow field of view (FOV): 43 degrees in the horizontal plane. This is much less than the human peripheral FOV. We needed more human-like FOV for our task of tracking multiple people.

**Table 2. Technical characteristics of Actroid's built-in cameras.**

| | |
|---|---|
| *Manufacturer* | Nippon Chemi-con |
| *Model* | NCM13-J |
| *Technology* | CMOS |
| *Sensor size* | ¼ inch |
| *Resolution* | 1280 x 1024 640 x 480 |
| *Frame rate* | ~10 FPS ~30FPS |

**Figure 4. Eye of the Actroid-F**

Because of the above mentioned technical limitations of the Actroid's camera system we had to very unwillingly resort to using a single outside camera, the Logicool (Logitech) C920. The benefits of this high-end webcam are:

- High resolution (1920 x 1080 pixels)
- High framerate (30 FPS) thanks to its H.264 hardware data compression
- Wide horizontal field of view (~80 degrees)

All these characteristics made the C920 the best choice for our task. One negative aspect of it is its size which doesn't allow it to be installed in the eyeball of any humanoid robot. Historically, webcams were not the preferred choice for use in robotics because of their mostly inferior optics quality (causing image distortions) and sensor size/quality compared to industrial cameras. On the other hand, since the wide acceptance of an easily available computer library, OpenCV [16], it has become much easier to correct the distorted images using simple image calibration procedures (checker board). This still does not provide extremely accurate images, but in situations like ours, where the points of interest are located near the center of the sensor field, the use of webcams can be warranted.

**Table 3. Technical characteristics of the external webcam.**

| | |
|---|---|
| *Manufacturer* | Logicool (Logitech) |
| *Model* | C920 |
| *Resolution* | 1920 x 1080 |
| *Frame rate* | 30 FPS |
| *Horizontal FOV* | ~80 degrees |

# 2. General background

Humans are great at communicating with each other. They can even notice subtle variations in each other's behavior, e.g. raised voice because of anger, erratic eye movement caused by anxiety, rapid involuntary facial expressions, etc. [17] Robots are not as perfect as humans in communicating with other people. One of the reasons for this is their insensitivity to implicit interaction cues, which include: body motion, speech modality, social eye gaze, etc. We are addressing this problem by enabling humanoid robots to detect human gaze cues and test how this helps making their social interaction more natural.

## 2.1 Implicit interaction in HRI

Implicit communication is a kind of interaction where transferring information is not the purpose of the communicator, but information is still conveyed [18]. For example, an increased speech volume can be unintentional, but it still conveys information to the listener about the emotional state of the speaker (anger, annoyance, excitement, etc.) Implicit interaction augments explicit communication by adding more information content to it [19]–[22]. Explicit communication without implicit communication tends to be less effective and increase the possibility for errors, whereas using implicit cues may add redundancy, thus making communication more robust [23][19]. Implicit signals supplement communication with additional modalities which enhance interaction quality as noticed in military literature, where communication is of essential value [24]. Often, implicit behavior is linked to the communicator's own goals [25]. These goals can be transmitted more easily through implicit interaction than in direct ways. As it is based solely on the perception of not obvious cues, the efficacy of implicit communication does depend on the person's perception skills [26]. One of the important implicit cues is gaze, which has been studied not only in human-human interaction but also on humans interacting with e.g. computers and interactive displays [27].

## 2.2 Social gaze

Gaze by definition is "to fix the eyes in a steady intent look" as defined by Merriam-Webster Dictionary or "to look at someone or something for a long time" as defined by Cambridge Dictionary. Thus, gaze itself is produced by the eyes. However social gaze besides the eye direction is also affected by head and body orientation [28][29]. The direction of attention is determined by eyes if the subject is close and by head orientation if the subject is far away, i.e.

when the eyes are not visible well [30]. In studying social gaze, there is a number of concepts which are of particular importance for defining interaction between humans or between humans and robots, which will be defined in the following sections.

### 2.2.1 Mutual gaze

Mutual gaze is an important social cue from early developmental age, which eventually leads to more complex behaviors like joint attention [31][32]. Mutual gaze happens when the gaze of two agents interlock, which has a special importance in human communication [30][33]. This is also called eye contact. This phenomenon has a special prosocial value since very early in human development [34]. If there is a lack of eye contact between two interlocutors, then they might not even feel as fully participating in the conversation [35]. Mutual gaze is also used as a regulator of intimacy between people [36]. This phenomenon does not depend only on one person, but on both collocutors simultaneously [37]. For example, if a professor during class keeps maintaining eye contact with her students, she will create a stronger bond with the pupils thus allowing a better teaching experience. This beneficial process can be blocked either by the professor not looking at the students or the students themselves not looking back at their teacher. If the mutual gaze is broken, the understanding and engagement between the two sides might be hindered [38]. In human-robot interaction mutual gaze is a very important gaze signal. For example mutual gaze between a human and a robot can be used by the robot to advance turn taking [2].

### 2.2.2 Joint attention

Recent studies showed that joint attention in infants might develop as a consequence of mutual gaze [39]. Joint attention is produced when two agents focus their visual attention on a single object and understand that the other individual is looking at the same object too. More specifically, joint attention is “a coordinated and collaborative coupling between intentional agents where the goal of each agent is to attend to the same aspect of the environment” [40]. For example if two people become aware of an object which fell on the ground, this situation is not joint attention as the source of attention was coincidental. True joint attention occurs for example when 1) both persons are attending to the same object; 2) both persons know that the other is attending too; 3) they both attend to the same specific feature of the object.

In human-robot interaction joint attention can play a very important role. For example, when a robot needs to hand over an object, it is required that joint attention and engagement are established with the receiving human, thus ensuring a safe handover [41].

### 2.2.3 Gaze aversion

Gaze aversion is the process where turn taking is modulated by gazing away, thus holding the floor in a conversation. When a conversation is continuous, there is no need for gaze aversion, but as soon one of the speakers needs more time to process the information heard he might resort to the social cue of gaze aversion, which signifies that the person needs more time to understand what was said and/or to generate an answer to the potential question [42].

It has been studied how robots can generate these cues, but less so how they can "read" them [43]. Authors have managed to generate these aversion cues even with robots which did not have articulated eyes (Nao) and successfully used them in human-robot interaction studies [44].

## 2.3 Gaze tracking approaches and systems

In this paper we will use the terms "eye tracking", "gaze tracking" and "gaze estimation" interchangeably when referring to the same principle of determining the direction of the eye gaze of an observed person in a camera feed.

In the following, we will give a comparison of existing eye tracking system types, highlighting their advantages and disadvantages for implementation on a humanoid robot.

### 2.3.1 Head-mounted or remote systems

Highest precision gaze tracking can be achieved with head mounted systems. These setups have cameras mounted on a helmet or glasses-like structure near the subject's eyes. The negative aspect of these systems is that they might be cumbersome to wear. In HRI, wearing such devices could not only inconvenience the subjects but it can also affect the interaction, by forcing subjects to be aware of their own gaze. Also, in real life scenarios, it cannot be expected from people to wear glasses just to be able to communicate with a robot using gaze. Remote eye trackers provide less precision than head-mounted ones but with the added benefit of being unobtrusive to the user. They are usually mounted either under computer displays or on the dashboards of vehicles for automotive use, thus exhibiting low mobility. Embedding a remote eye tracker into the visual system of a mobile robot would help to overcome the limited working area of the tracking device.

### 2.3.2 Active or passive systems

Active eye trackers usually emit infrared light to a) break any shadows on the face of the subject, but more importantly to b) cause a reflection of light off the lenses of the eye (Purkinje images). Active systems locate these reflections and significantly enhance their gaze estimates based on

knowing the locations of the light sources compared to the cameras. As they operate in the infrared spectrum, the cameras used to record images are equipped with IR-pass filters. Passive eye trackers instead operate in the spectrum of visual light and usually do not use additional sources of light, avoiding also the risk of causing discomfort during interaction by drying out the partner's eyes. This makes the passive trackers much more natural to use and also better suited for HRI. On the other hand they struggle with imperfect lighting conditions and the lack of glints on the cornea that could allow easier gaze estimation.

### 2.3.3 Appearance vs. feature-based trackers

Appearance-based gaze trackers feed the image of the eye to a black box method (usually a convolutional neural network) to acquire an estimate of the gaze [45]. Their advantage is that they can operate even on noisy images, but they require training. Feature-based eye tracking algorithms [46] focus instead on extracting eye/face measures. They find features of the eye region using machine vision techniques that will allow the estimation of gaze. These features, which include corners of the eyes, outline of the iris and pupil, center of the pupil, outline of the eyelids, etc. are usually easy to extract from images with fair lighting.

## 2.4 Gaze tracking for human-humanoid interaction

As mentioned before gaze reading is particularly interesting for humanoid robots, as it is a natural human ability that might be expected from human-like robots. It has been often studied how robots can generate the variety of gaze behaviors which directly support communication, as mutual gaze [7], joint attention [47], gaze aversion [48], etc., but less so how they can "read" them [43].

Interestingly, due to technical limitations, many authors have substituted eye gaze with its closest proxy, head orientation (e.g., [49][50][38]), because of the increased complexity associated with the calculation of eye gaze. Although head and eye orientations are often in line, in certain types of interactions the eye movement is more relevant. For instance, in a scenario where one collaborator is using gaze aversion to "ask" for thinking time, the head can remain still, while only the eyes glance up. By looking only at head orientation, the observing partner could miss this important non-verbal cue. More in general, Borji and colleagues [9] found that the probability of guessing the actual gaze direction by human subjects is significantly higher when the eyes are visible in an image, compared to when the eyes are not visible. On the other hand head pose can be very useful when the eyes are not visible at all.

Among the studies which focused on eye gaze in human-human interaction and human-robot

interaction, several so far have been conducted with head mounted eye trackers [51][52]. This type of research provides a very detailed insight on eye-gaze behavior, but the systems used are too obtrusive for real-life interaction.

Finally a number of researchers have looked at using remote eye tracking systems built in humanoid robots. Matsumoto and Zelinsky for instance described a design that allowed the estimation of gaze using a remote system [53] while Ido and colleagues implemented it in the HRP2 humanoid robot [54] to enable it to detect face and gaze direction of its human partner. This information was then used for turn-taking, i.e. to figure out when the human collaborator was addressing the robot. The above mentioned systems were used to detect eye contact which is a subset of general gaze tracking which would in turn include gaze detection on different objects and people in the robot's vicinity. Also, so far no extensive use of eye gaze tracking has been done in human-humanoid interaction, in particular in cooperative scenarios that combine turn-taking and joint attention tasks. These are some of the improvements we are aiming for with our current research. Moreover, we intend to improve gaze tracking for HRI by adopting a modular software approach in which different sub-algorithms could be easily substituted by better ones and by using new and more precise techniques of facial features detection (see Sections 5.2.2 and 5.2.3 for details).

# 3. Study I – An affordable active robot head for gaze tracking

## 3.1 Introduction

Gaze plays an important role in human-human interaction. People exchange many glances while communicating with each other. However, our eyes do not only provide us with visual information but they also serve as tools for implicit interaction: in a busy administrative office, the attending clerk needs only to glance at the next client's eyes to initiate the transaction. We are also very much able to guess about someone's object of attention by just observing the eyes: a customer's glance reveals which item s/he is interested in and tells the seller which product's presentation to focus on. It is thus evident that robots could also benefit from knowing their human collaborator's gaze direction. But from the technological point of view, the robot camera systems are often not as sophisticated as the human eye: they do not have the necessary spatial resolution nor sensor distribution of the human eye. The robot needs a wide field of view (FOV) in order to locate potential human collaborators, but then it needs a narrow field of view to zoom in on the face of the detected person to figure out their gaze. This is an approximation of how human peripheral and foveal vision works. Finally, today's robots have a very limited data bandwidth over their inner networks, which puts a limit on the spatial and temporal resolution of the imaging system.

We propose a low-cost active camera system that addresses the before mentioned issues: we use a high definition (1080p) stereo pair of webcams in a robot head setup (see Figure 5) which performs pan and tilt movements [1]. Such a system could:

a) provide a wide field of view (64°) low resolution (VGA) image feed for locating collaborators,
b) use its unused pixel density to digitally zoom in to the face (30° FOV) of the detected person to perform eye tracking,
c) use pan and tilt for keeping focus on the face.

After discussing related work and a technical description of the system, we will report on a short experiment which validates the proposed camera system's ability to detect mutual gaze better than regular non-zoom camera solutions.

## 3.2 Study background

Since their appearance, eye tracking systems have found many applications in human-machine

interaction [55]. Remote eye trackers started becoming more used than head-mounted ones, as they are more convenient and less intrusive. The benefit of remote systems has also been recognized in human-robot interaction studies [56]. Instead of expensive commercial systems, our work proposes an active vision system developed by using affordable webcams. Drawing inspiration from the work of Atienza and Zelinsky [57], we present an active vision system with zoom capabilities, to cope with a wide interaction space and moving subjects. Our work expands on previous ideas by using affordable modern technology and by validating the benefits of a zoom system in a human subject pilot study. An important goal for eye tracking in HRI is the detection of mutual gaze, the exploration of which is also one of the goals of our research.

## 3.3 Methodology

The robot head system consists of the visual and actuator system, see Figure 5 and Figure 6.

### 3.3.1 Visual system

The visual system is a stereo mount of two Microsoft LifeCam Studio webcams. These cameras were selected for their high resolution (HD, 1080p, 1920x1080 actual pixels), compact size (for installing them in a humanoid robot), auto-exposure and auto-focus capabilities.

**Figure 5. Human-agent interaction.**


**Figure 6. Camera and motor setup.**

Because of the limited data bandwidth of the communication networks on modern robots we decided to constrain our system to VGA resolution of the cameras (640x480). Such a setup provides a relatively wide field of view (64°) that is adequate for recognizing people's faces in the robot's environment (using the Viola-Jones face detection algorithm in OpenCV [58]). Precise eye tracking is enabled by the cameras' digital zoom capability. This allows narrowing the field of view to 30°, while using the cameras actual sensor pixels instead of interpolation as

in non-HD webcams. This dual purpose narrow-wide FOV operation mode addresses the necessity to perceive details of what is fixated rather than out-of-focus elements: we see much more precisely in a narrow cone of our eyes called the fovea, while we have lower resolution outside of it, i.e. our retina is a space-variant sensing surface [59].

### 3.3.2 Motor system

The pan and tilt movements are performed by two Dynamixel AX-12 digital servo motors. They are mounted in the neck of the robot. Additional servos will provide head roll, eye vergence, and eye tilt in future implementations. The motors are controlled in a closed loop: compensation movements are generated so that the human collocutor's face is kept in the center of the camera image even when the person moves around.

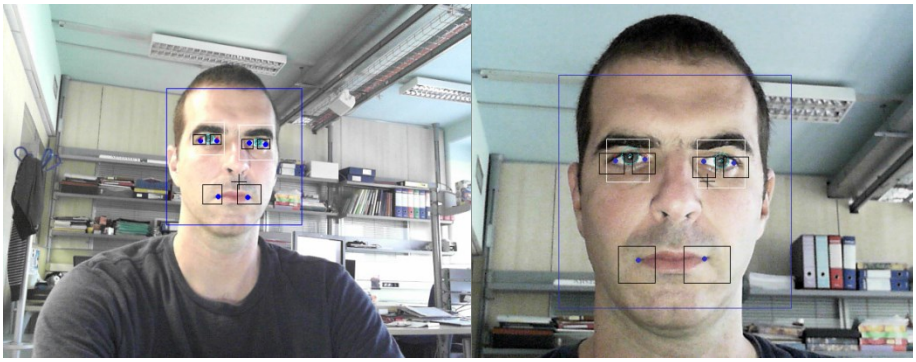

**Figure 7. Zoomed out (left) and maximum zoomed in (right) images.**

### 3.3.3 Operation procedure

The robot head starts its operation by panning left and right. When a face is detected, the motors start to track it, keeping it in the middle of the camera images. At the same time the cameras zoom in on the face and change the zoom level continuously to keep its size in the image nearly constant even when the subject moves closer or further away. This way facial features occupy a large area of the camera images, thus providing enough pixels for eye tracking. Once a face is detected, the Viola-Jones algorithm is again used for roughly detecting the eyes. Within this area the corners of the eyes are found using template matching and the iris is located by a circle fitting algorithm (Hough transformation), Figure 7. If the cameras lose the face for more than a second they automatically zoom out and start looking for a new face to detect, effectively restarting the process (see a video[1] of how the system works). It is worth mentioning that the proposed system uses visual light without Purkinje image tracking. It also does not need supervised face model learning for each subject.

[1] https://youtu.be/9xzQWAUCFW8

## 3.4 Experimental results

We designed and ran a pilot study to verify some of the benefits of our system: namely, we were interested to see if mutual gaze could be more precisely detected and tracked using our pan/tilt/zoom mechanism compared to a non-zoom system. For this task only the right webcam was used, because of bandwidth considerations. Three subject completed the test. They were asked to look either straight at the camera (mutual gaze) or 5 and 10 degrees to the left of it, as we made ten angle calculations for each offset. The distance of observation was either 40cm (near) or 80cm (far). The near condition didn't require any zooming, because the subjects' faces already occupied most of the camera image. The far condition had two options: using zoom and not using zoom. In the first one we let the previously described algorithm enlarge the face (Figure 7, right) while the latter condition did not use any zoom (Figure 7, left). The pan/tilt face tracking algorithm was operational, but images were taken only after the system stopped moving (reached a stable state). Gaze direction was calculated as the angle between straight ahead position (baseline) and the detected position of the eyeball, by assuming an eyeball diameter of 24mm. Figure 8 shows the absolute error between real and detected gaze directions, averaged over all three subjects and all three angle positions. It can be noticed that the error is quite low for the "close" and "far zoom" conditions while it's much higher for the "far no-zoom" option. We performed a Friedman ANOVA test on the absolute errors for each subject and found that differences were highly significant ($p<0.001$). This confirmed our expectation that when the subject is far away from the robot and the face is not zoomed in, the results will have high rates of error. These levels of error effectively prohibit from detecting mutual gaze in systems with only a wide FOV. Indeed, if we assume a threshold of 4° (which is about the width of an eye from 80cm away) for discriminating between mutual gaze or not, then the zoom enabled system was 90% accurate on average while the no-zoom system's performance was only 42%. Hence, mutual gaze can be detected more easily using a system like ours.

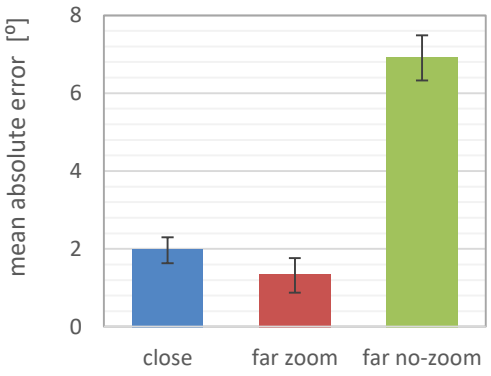


**Figure 8. Mean absolute error.**

Figure 9 shows detailed results for each subject and each angle. It can be seen that for the "close" and "far zoom" options, the angle estimates for all subjects cluster around their nominal values (e.g. orange markers around 5 degrees). At the same time the "far no-zoom" option shows erratic results (markers of different colors are mixed around different nominal gaze offsets) which confirms that mutual gaze detection is very difficult when not zooming in on the face.

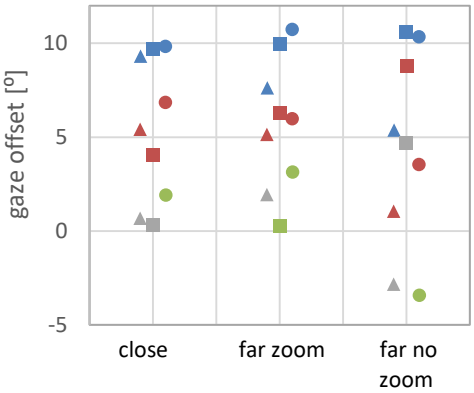


**Figure 9. Average gaze points for each subject, each gaze offset and each condition. Triangular markers denote Subject01, squares Subject02 and circles Subject03. Gray markers denote straight ahead gazes, orange ones 5 degree offset, while blue stands for 10 degree offsets.**

## 3.5 Conclusions and discussion

In this work we presented an active camera system that is designed to facilitate eye tracking for use in human-robot interaction. The pan/tilt mechanism allows the robot to a) scan its environment to find interaction partners while it's zoomed out and b) track a detected face while it is moving around in zoomed in mode. The amount of zoom was automatically controlled to keep the face maximized in the camera image. The digital zoom lets the robot perceive more details about a subject's face features, e.g. the eyes and the mouth. These details enable more precise eye tracking compared to a system without zoom. This advantage becomes evident in situations when the subject is more than 40cm away from the robot. Since many interaction scenarios involve distances greater than 40cm, such a system would benefit most robots. A digital zoom system can be faster and much cheaper than an optical zoom lens. The 40cm distance is very close for interacting with human-sized robots but it can be adequate for tabletop robots.

The proposed active vision system is very affordable as it uses off-the-shelf web cameras and servo motors (less than 250EUR total), thus allowing wider and quicker dissemination. As web cameras' performance rapidly increases with every new generation, they could slowly replace

much more expensive systems for robot applications, also thanks to the OpenCV library, which makes it simple to calibrate these low-cost cameras for manufacturing imperfections.

# 4. Study II – A gaze-contingent robot for a dictation scenario

## 4.1 Introduction and background

As robots are making their way from factory floors into our everyday lives, the design of their interaction with humans is becoming more and more important. For example in Japan even today it is possible to find robotic greeters when entering electronics stores or mobile service providers (e.g. Pepper). Thus, it is important to pay careful attention to how these new entities will communicate with humans. The basis of interaction is for the robot to respond to the actions of the person: when customers come in, greet them with a smile. As humans naturally use eye contact to establish and modulate interpersonal communication [7] it could be beneficial if robots could also do so while talking to people [60]. Even though a reactive (contingent) approach of the robot is usually favored in conversational turn-taking [61][62] it is still a question if this holds if the role of the robot and its human partners is changed, i.e. what if the robot assumes a leading role, as for example in teaching? Would it be appropriate for it to react to the needs of the "students" or to try to keep a predefined pace not paying attention to reactions? Or more precisely, what amount of contingency is appropriate for a robotic tutor? For example, if we look at a dictation scenario, often present in second-language learning schools, should a potential robot teacher follow the pace of students who are taking notes or should it impose a pace on them? Moreover, to what amount should the robot use eye contact to establish the timing of the interaction?

In our opinion gaze is an important implicit element of communication. For example, humans and more recently even robots can hold their ground in a conversation by just averting their gaze, signaling that they are thinking about what to say or that they are disengaged [43]. Even more importantly, mutual gaze (i.e. eye contact) detection plays a very important role in turn-taking [63]. In this paper we explore the benefits and drawbacks of augmenting a teaching scenario with implicit gaze communication. We compare a contingent behavior, where the robot reacts to its partner's glances to a purely rhythmic one, where the pace of interaction is preset.

With this task we aim to address two main questions. First, can a mechanism as simple as the detection and the response to subjects' gaze be enough for controlling the turn-taking process in a dictation task, with no explicit instructions given to the participants? In other words, can a simple assumption about an automatic interactive behavior - as gazing at the robot to get more information - lead to a working turn-taking system? Second, will the adoption of a responsive or

adaptive behavior lead to a more efficient and time-effective interaction avoiding idle times or will it lead to a slower task completion, as participants will tend to slow down when their timing is not regulated by the teacher?

## 4.2 Methodology

In this experiment [2][3] subjects assumed the role of students whose task was to write down on a whiteboard what their robotic teacher dictated (see Figure 10). The robot dictated two sets of 32 short sentences, one in English and one in the participants' mother tongue (Italian). Two different dictating strategies were adopted and presented to the subjects as procedures alpha and beta. In the alpha condition – hereafter Rhythmic – the dictation progressed at a predetermined fixed pace. In the beta condition – hereafter Contingent – the robot pronounced a sentence only when it detected that the subject was gazing at it, assuming that establishing mutual gaze would signal the readiness of the subject to continue writing. The idea of using two languages for the dictation - one mother tongue (Italian) and one not - was made on purpose to change the task difficulty within subject. This way we could test whether the contingent ability of the robot could make it adapt to the different temporal needs experienced by each of the subjects when confronted with tasks of different difficulty. In the following sections, we will provide more details about the system, the different conditions, the subject sample and the data analysis.

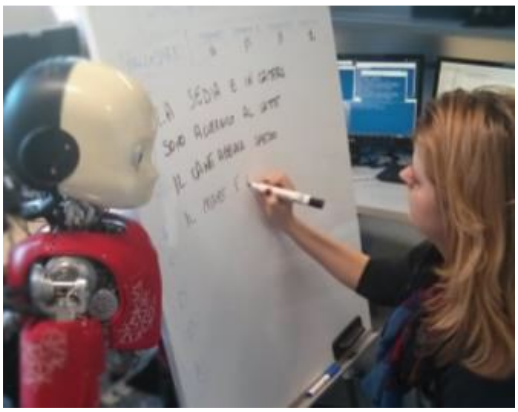

**Figure 10. iCub dictating.**

### 4.2.1 The setup

The robot used in this experiment was the humanoid robot iCub [10]. Our setup leveraged on the use of some existing iCub modules as well as the development of new ones. Figure **11** gives an overview of the system architecture.

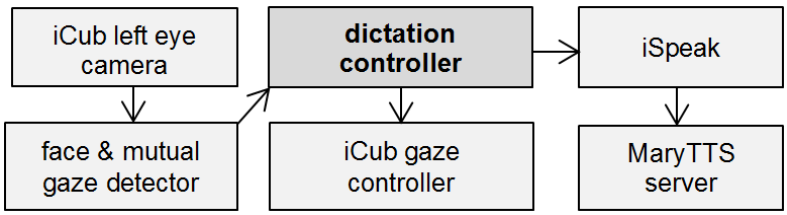


**Figure 11. System architecture.**

In the following, we will describe each of the elements mentioned above.

1) iCub left eye camera: We used the standard iCub camera-grabbing module to acquire videos for our scenario in VGA resolution (640x480) from the iCub's left eye camera (PointGrey Dragonfly2). The speed of image acquisition was around 20 frames per second throughout the experiment.

2) iCub gaze controller: This block represents another standard iCub module, iKinGazeCtrl [14], which provides an interface to adjust the robot's gaze direction towards any given point in the camera's image. When a new gaze direction is set, first the eyes perform a saccadic movement towards the goal and then the head turns too, so that the eyes are back to their straight forward direction as much as possible. This capability is employed in our system to track the subject's face with dual intent: a) to keep the face always in the field of view of the camera and b) to provide a more human-like behavior of the robot by directing its gaze towards the subject.

3) Face and mutual gaze detector: For detecting the location of the subject's face and the potential mutual gaze in the camera image, we used the face and mutual gaze detection tool, described in [3]. The module was initially verified in [2], and now exploited in the current experiment. The iCub's own built in camera was used.

The location of human faces is detected by using an open source face detector [64]. Face detection alone is not sufficient to detect mutual gaze. Features corresponding to the location of the pupils need to be extracted from the face region. Furthermore, features correlating to the head pose are required to compensate for different head orientations. First, facial features are extracted using dlib's implementation of [65] which provides robust facial feature detection even on partially occluded faces.

Subsequently the facial features are used to extract the eye regions of each face. On each region, a pupil detection algorithm is applied. The algorithm uses multiple heuristics to generate a candidate map of the most probable pupil location even in presence of low resolution and noise. First, a gradient-based approach is used to extract center candidates [66]. Second, the region is

adaptively categorized into two luminance classes. The weights for the map corresponding to darker areas are increased. The assumption here is that darker areas more likely correspond to the pupil. Finally, the location of the maximum in the candidate map provides the desired pupil location. A subset of the facial features and the pupil coordinates are used to build a 6-dimensional feature vector. To detect mutual gaze we trained an epsilon-insensitive support vector regression model to estimate the horizontal gaze direction[2]. As training data we used the Columbia gaze database [67]. Using a regression model has the advantage that it is now possible to detect mutual gaze simply by applying a threshold to the estimated horizontal gaze direction. In our experiment, we set a mutual gaze threshold of ±10°.

Once the features are extracted and classified, the mutual gaze status is sent together with the coordinates of each face to the Dictation Controller, which in turn provides the center coordinates of one face to the Gaze Controller module every second, to ensure that the robot keeps a steady gaze on the human participant.

4) iSpeak: iSpeak is a standard iCub module which provides speech synthesis functionality to the robot. In our setup it receives textual sentences from the Dictation Controller and passes them on to the MaryTTS module. At the same time it produces simulated lip movements using the LED lights representing the robot's mouth.

5) MaryTTS: It is an open-source text-to-speech platform which transform textual sentences into speech using different voices [68].

6) Dictation Controller: The DC module was specifically created for the current experiment. It accepts as input the location of the subject's face and the presence or absence of mutual gaze. As output it sends textual sentences to iSpeak for execution and also tells the iCub gaze controller which way to turn the robot's visual attention. In the Rhythmic condition of the experiment the robot does not react to mutual gaze, rather the sentences are sent out to iSpeak with fixed timing. None the less, the robot waits for the issuance of the next sentence proportionally to the length of the previous sentence, which is being written down by the human subject. The waiting time was selected to simulate an average writing time of about 26 words per minute [69]. On the other hand, during the Contingent condition, the next sentence is not started until the subject glances back at the robot, after finishing writing. This glance back is the mutual gaze signal sent by the Mutual Gaze Detector module. We require the mutual gaze signal to be present continuously for at least 150ms for it to be recognized as a gaze back event. As an

[2] More details about the system can be found at: https://github.com/lschilli/gazetool

additional constraint we disabled reactions to gazes back at the robot in the first 5 seconds after the end of speech, in order to suppress false positives, as it was impossible to finish writing within such a short period.

### 4.2.2 Subjects

Eight subjects (6 women and 2 men, ranging in age from 26 to 33 years, mean age 28 years) took part in the experiment. All subjects were healthy and did not present any neurological, muscular, or cognitive disorder. All participants gave written informed consent before testing. The study was approved by the local ethics committee and all experiments were conducted in accordance with legal requirements and international norms (Declaration of Helsinki, 1964).

### 4.2.3 Procedure

The whole task consisted of the dictation by the humanoid robot iCub of four paragraphs, each composed of 8 short predefined sentences (e.g., "The flowers are red."), in two sessions: one in English and one in Italian. In total subjects had therefore to write 64 short sentences. In particular, for each language, participants encountered two blocks of each condition (Rhythmic and Contingent) in counterbalanced order (i.e., either R-C, C-R or C-R, R-C). Also the order of language (Italian or English) presentations was counterbalanced among participants to control for order effects. Subjects were instructed to listen to each sentence and then write it down, while leaving blank spaces for any word that they did not understand. The sentences were chosen so that, in each paragraph, the average length was about 19 characters, both for the English and for the Italian sessions. The difference between conditions was that in the Rhythmic condition, the robot waited for a fixed time after each sentence, while in the Contingent condition, the robot did not initiate a new sentence until the subject gazed at it. In both conditions the robot moved its head and eyes to look at the subject. The task lasted on average about half an hour per subject and was fully recorded both through the camera in the robot's left eye and through an external camera. After the experiment subjects were requested to complete a short questionnaire where they had to rate each of the two procedures (alpha or beta) on three 7-point scales with respect to the perceived probability to make an error, the pleasantness of the procedure and its difficulty. Then, they were asked to indicate which of the two would have belonged to a more advanced language course and to briefly explain which the actual difference between them was.

### 4.2.4 Data analysis

The video recordings of all subjects were annotated in ELAN[3], to individuate the timing of subjects' writing, robot dictating and potential system failures or subjects' strange behaviors (e.g.,

[3] ELAN - The Language Tool, https://tla.mpi.nl/tools/tla-tools/elan/

a posteriori corrections of previously written sentences) [70]. The annotations were then imported in MATLAB through the SALEM Toolbox [71] where they were further analyzed. The main variables for the analysis were Task Duration – the time interval between the beginning of the dictation of two subsequent sentences; Wait Time – the time interval between the completion of writing and the beginning of the dictation of the next sentence; the Writing Speed and the Number of Errors in the writing. Writing Speed was computed as the number of characters written by the participant in the time interval between the beginning and the end of the writing of each sentence, then averaged across all sentences belonging to the same condition and language. Number of Errors was also transcribed from the video by counting up the syntactical errors the subjects committed. Furthermore from the responses to the questionnaire we derived a measure of the perceived pleasantness of the two procedures and an evaluation of how clear the understanding of robot behavior in the two conditions was.

## 4.3 Experimental results

### 4.3.1 System errors

During the execution of the experiment there were three cases when a technical error in the Dictation Controller algorithm caused the robot to pronounce two sentences one right after another. Since each time two sentences were affected, we needed to eliminate 6 sentences out of the total 512 (1.17%) from the final analysis. Furthermore, the Gaze Detection algorithm sometimes caused false positives (detected mutual gaze when the subject was not looking at the robot) and false negatives (didn't detect when the subject was in mutual gaze). False positives occurred 7 times (1.37%), while false negatives were recorded 9 times (1.76%). Except one time, these false detections did not cause automatic cancellation of the sentences, as they were not disruptive to the process. It was assumed that the face of the subject will always be kept in the field of view of the robot. This was true except one time when a subject moved out of the robot's FOV, thus the robot gaze had to be manually redirected back to the subject, which caused one sentence to be eliminated.

### 4.3.2 Subjective evaluations

The first goal of our experiment was to assess whether subjects could perform the dictation in the Contingent condition without any explanation of how it worked. All subjects but one automatically adapted appropriately to the task, naturally gazing at the robot after finishing writing. The single exception, who initially stared continuously at the whiteboard, started looking back at the robot after being invited by the experimenter to "interact with iCub", and from that moment on established an appropriate gaze pattern for the rest of the experimental

session. To evaluate whether participants had explicitly understood how the system worked in the two different conditions (called generically alpha and beta during the experiment), in a questionnaire we asked them to describe the difference. Of the 8 subjects only two realized that such difference consisted in how the robot timed its utterances. Of the other 6, three did not describe any difference, while three erroneously thought that a difference existed in the type of sentences used or in the robot's voice. It is important to note that although most subjects did not realize that robot behavior in the beta (Contingent) condition was responsive to their gaze, they naturally exhibited a gaze behavior which was appropriate to guarantee the continuation of the task.

The second question we were interested in was whether a contingent, more adaptable robotic behavior could result in a more pleasant interaction for the human partner. To address this question we asked subjects to evaluate separately the two conditions alpha and beta at the end of the experiment, choosing a value on 7-point scales for the probability to make an error, the pleasantness of the task and the difficulty of the condition. Although most subjects could not detect the actual difference between the two conditions, 5 out of 8 participants found beta (Contingent) less difficult and less likely to cause errors, and 2 of them found it also more pleasant. The other participant rated the two conditions equally. Accordingly, when asked which of the two conditions would have been part of a more advanced language learning course, the majority (5 over 8) picked alpha (2 indicated beta and one indicated both).

### 4.3.3 Quantitative analysis

To be sure that we were not causing additional difficulty to the task with the introduction of the Contingent behavior, we compared the number of errors in writing (misspellings and blanks) as a function of the condition and the language of the dictation. As it can be seen in Figure 12 left, the number of errors was significantly higher for English than for Italian ([$F(1, 7) = 35.48$, $p<0.001$], Two-way Within Measures ANOVA, with Language and Condition as factors), but no difference was present as a function of Condition ($p=0.54$), nor any interaction between Condition and Language ($p= 0.38$). Therefore, the English dictation qualified as a more difficult task than the Italian one (mother tongue) for our sample, while gaze contingency had no effect on the number of errors. This is confirmed also by an analysis of the average writing speed (Figure 12, right), which appears to be significantly slower for English than Italian [$F(7, 1) = 10.51$, $p = 0.014$], but stable across conditions (Condition: $p=0.35$; interaction: $p= 0.54$, Two-way Within Measures ANOVA, with Language and Condition as factors). Here we wanted to create an easier and a harder task, on purpose, to evaluate whether the increase in difficulty made the Contingent

system more advantageous or not. Hence the reduced writing speed and the increased amount of errors in English are a successful manipulation check, demonstrating that our prediction works: English is more difficult for our sample than Italian.

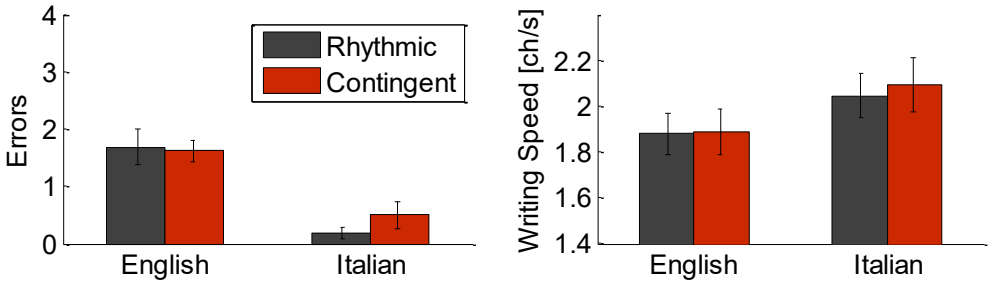


**Figure 12. Left: average number of errors per condition; Right: average writing speed per condition. Error bars represent sample standard error (SEM).**

Since subjects adopted different strategies to cope with difficulty at understanding the dictation in both the Contingent and the Rhythmic conditions, with some immediately leaving a blank and some spending a long time thinking to the possible completion, we decided to remove the sentences containing errors from the following analyses to reduce inter-individual variability.

A further important question that we wanted to address with our task was whether leaving the possibility to the subjects to – implicitly – control the timing of the dictation could have led to slacking, i.e., to the adoption of a slower pace, especially for those subjects who naturally tend to be slower at writing. To verify this we measured for each subject the time to complete a single sentence (Task Duration), as the time between the beginning of the dictation of one sentence and the beginning of the next. In the Rhythmic condition this value was fixed to a predetermined average value (see Section 4.2). In the Contingent case, it was determined by when the participant looked back at the robot.

In Figure 13 we plotted individual Task Durations in the Contingent condition averaged over each block of 8 sentences (top panel – Italian, bottom panel – English). From the graphs we can derive two observations. First, on average task duration in the Contingent condition did not differ significantly from the reference (Rhythmic) value. This is confirmed also from a Two-way Within measures ANOVA on the Task Duration averaged between the two blocks, with Language and Condition as factors, where neither factors nor interaction reached significance [$F(1, 7)$, $p = 0.46$, $p = 0.12$ and $p = 0.27$ respectively]. Hence, even when implicitly allowed to freely pace themselves, subjects maintained on average the fixed timing predicted by assuming an average writing speed. The second observation regards instead the difference between the first and second blocks of sentences. Indeed, a few subjects (three for Italian and one for English) exhibited a clear adaptation between the two blocks, with a substantial decrease in average Task

Duration, up to 140% in Italian and 224% in English. Therefore, at least a subset of subjects needed a few trials to get entrained in the appropriate behavior for the Contingent condition to run.

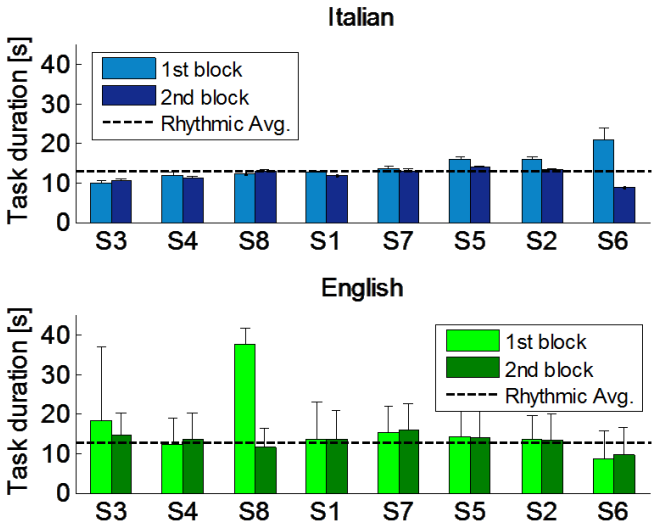


**Figure 13. Average Task Duration in the two blocks of the *Contingent* condition for each subject, subject, sorted with increasing Task Duration in the 1st Italian block. Error bars represent SEM.**

To discount the impact of training on performance, in the following analyses we considered only the second block for all conditions.

Another useful variable to compare the performances and strategies adopted by participants in the two different conditions is represented by the Wait Time, i.e. the time between the completion of the writing of a sentence and the beginning of the dictation of the next. In the Rhythmic condition, the dictation timing was fixed, therefore the Wait Time indicates how appropriate the chosen velocity for the subject was, with negative Wait Time implying that the dictation was too fast (the robot started dictating before the subjects completed their writing) and large positive Wait Time indicating that the Rhythm was too slow, potentially leading to boredom and loss of time. In Figure 14 (left panel) we have plotted individual Wait Times during the last block of each Rhythmic condition, as a function of subject's average writing speed. As expected, Wait Time tends to increase with subjects' speed (linear fit slope: 2.10 ± 0.82 (SD), $R^2$ = 0.52 for Italian, slope: 4.51 ± 1.90 (SD), $R^2$ = 0.48 for English). However, for most subjects it is positive and not too long (about 2 seconds), indicating that the timing selected for our Rhythmic condition was reasonable for the task at hand.

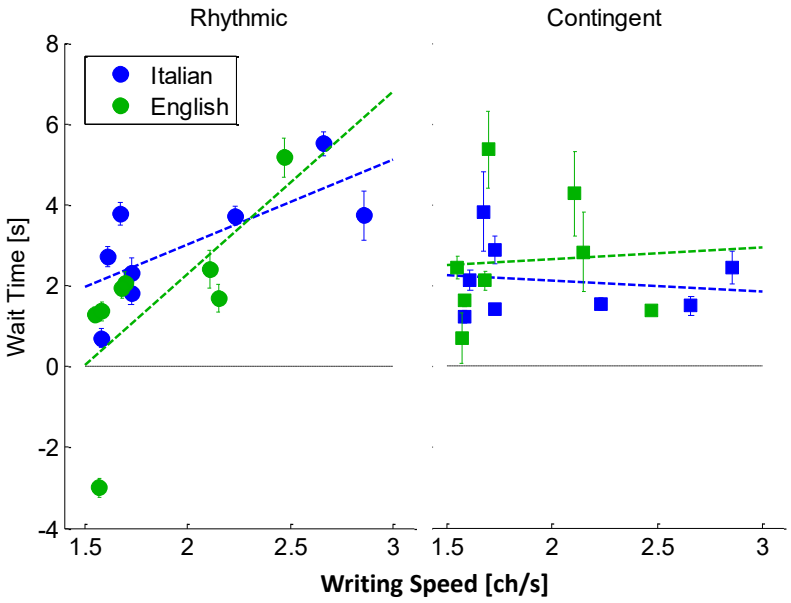


**Figure 14. Individual Wait Times as a function of Writing speed. Error bars represent SEM over the last block of each**

We then moved to check what happens when the dictation rhythm is not fixed but depends on subjects' gazing. Will slower subjects take more time to process and check what they have written? Or will faster subjects compensate the short time spent writing by a lengthier check of their sentences? Figure 14 (right panel) seems to suggest the opposite, i.e. a tendency to converge on average to the same Wait Time (again around two seconds) independently on the average subject's writing speed (linear fit slope: -0.27 ± 0.70 (SD), $R^2 = 0.02$ for Italian, slope: 0.29 ± 1.82 (SD), $R^2 = 0.004$ for English). This suggests for instance that the faster participants exploited the contingent scenario to accelerate the rhythm of the dictation. The slowest participant, instead, who was often interrupted in his writing in the Rhythmic condition, in the Contingent case exhibited a slightly slower rhythm that guaranteed him at least a brief time between one sentence and the next.

As a last analysis we evaluated whether task difficulty had an impact on how the same subject dealt with the possibility to control the turn-taking in the interaction. To this aim we compared the Wait Time each subject adopted in the Contingent condition with respect to the Wait Time he or she exhibited in the Rhythmic one (Delta = Wait Time Contingent – Wait Time Rhythmic), both when the session was easier (i.e., in Italian) or more difficult (in English). In Figure 15 these differences are plotted for each participant, and participants are ordered as a function of their average writing speed. From the graph it emerges that most subjects (5 of 8) exploited the contingency in the easier condition to reduce the Wait Time (i.e., most blue bars are negative). However, all but the fastest subject showed the opposite tendency in the more difficult (English) task. So there is a significant difference in the strategy and relative timing adopted as a function

of task difficulty, even within the same subject (Pair sample t-test on Delta with Language as factor [t(7)= -3.31 p = 0.013]).

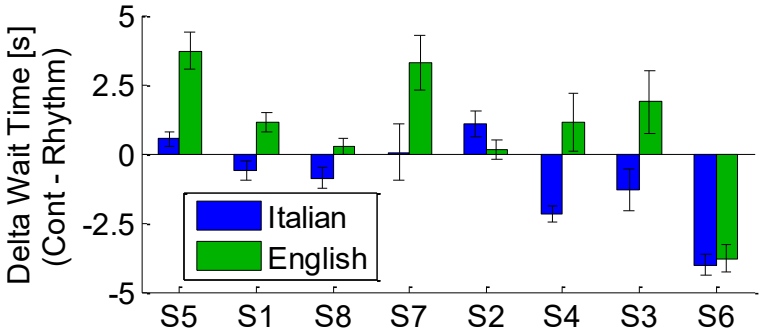


**Figure 15. Average differences in Wait Time between the last block of the Contingent and the Rhythmic conditions for the different subjects. Error bars represent standard deviations (SD) of the difference.**

## 4.4 Discussion

The aim of this study was two-fold. On one hand we wanted to demonstrate the importance for the robot to read an implicit communication signal as the establishment of mutual gaze to regulate the interaction. On the other hand we aimed at assessing under which conditions a gaze contingent, personalized response could lead to a more efficient or more pleasant interaction.

To begin with the first question: was the possibility for the robot to monitor humans' gaze signal important to establish a natural and seamless interaction? The answer that comes from our study seems to be positive. Most subjects did not realize that the difference between the two experimental conditions was the gaze–dependency of one of them. That notwithstanding for most of them the interaction proceeded successfully both in the Rhythmic and in the Contingent case, suggesting that looking at the robot when the writing was completed came as a quite natural attitude. Only one subject needed an incitement by the experimenter ("Please consider that the robot is waiting for your interaction before continuing the dictation"), but also this subject after this first comment started a natural turn-taking with the robot. Hence, reading subjects' gaze was an efficacious "trick" to appropriately time the robot's actions, leveraging on a natural human attitude, i.e. looking at a silent speaker to get more information.

But was mutual gaze the better signal to indicate readiness in this setting? Alternative approaches could have been monitoring the hand of the writer to be able to anticipate when he was decelerating toward sentence completion. Although such an approach could have guaranteed a higher degree of anticipation in the contingent session and hence a higher responsiveness, it

would have increased also the complexity of the system. In particular establishing when the subjects are satisfied with their writing and ready to pass to the next sentence can be an insidious task. Should the right time be at the end of writing of the dictated sentence? And what if the subject feels the need to re-read or correct a misspelled letter or to add punctuation? The use of gaze as implicit signal actually moves the responsibility of the choice of when to pass to the new sentence directly to the subject, who freely and unconsciously decides when he is ready. Moreover, monitoring the gaze of multiple people at the same time is already possible with our system, while monitoring multiple people's writing could represent a more challenging and error-prone task. This consideration could become relevant in view of possible applications or robot teaching groups of people, for instance at a school.

Moving to the second main question of our work, establishing whether a contingent behavior is advantageous in a robot also when its role is that of a leader (and potentially a pace maker) requires a more complex evaluation. From a subjective point of view, the answer seems again positive: no participants preferred the Rhythmic condition to the Contingent one, and five of the 8 subjects felt that the Contingent condition was slightly easier and less error-prone.

From a quantitative evaluation of the performance however the reply must be more cautious. Although no significant decreases in performance appeared when subjects were – implicitly – allowed to pace the interaction, neither a significant improvement (e.g., faster task completion) appeared on average. Moreover, a few subjects needed some trials before getting entrained with a steady-state rhythm of interaction, which led to highly variable behaviors at the beginning of the Contingent session (e.g. compare first and second block of trials in S6 and S8 in Figure 13, top and bottom panel respectively). Furthermore, choosing a contingent approach implies also the increase in the risk of system errors that a simpler rhythmic system does not face. So, a trade-off must be evaluated between the advantages yielded by the contingency and the inherent risks of errors (in our case false positives or false negatives in the detection of subject's mutual gaze – not detecting the readiness of the subject).

On the other hand, the subjects who were at the extremes of the writing speed distribution – the slowest and the fastest – could actually take advantage of the robot's contingent behavior. Only in such condition the former could complete writing without being interrupted by the next dictation and the latter could accelerate the dictation process at her own pace. So, if the Rhythmic condition is good enough for the average subject, the Contingent case makes a real difference mostly for faster and slower participants. This trend is visible also within the same subject when faced with tasks of different difficulty. Indeed, most participants exhibited the

opposite strategy when dealing with an easier or a more complex task: they decreased their Wait Time when writing in their mother tongue and increased it for the dictation in the foreign language (see Figure 15). So, a contingent approach makes the robot dictation suitable to cope with variability among different subjects and also within the same subject, if dealing with tasks characterized by different levels of difficulty.

To sum up, although a contingent behavior in our dictation context had clear subjective advantages and did not disrupt the appropriate rhythm of the interaction, a case to case evaluation is required to quantify the advantages and drawbacks that such an approach might determine. Indeed, contingency might lead also to the need for an initial adaptation to the turn-taking and to a larger variability in subjects' performances as a function of task difficulty. However, if the implementation of a contingent system is sufficiently simple and robust, there are situations in which it should be preferred. In particular this holds true when an average estimate of human behavior is not a good predictor for the performance of the individual human partner involved in the interaction, as for instance when working with kids or special populations.

## 4.5 Study conclusions

"Taking dictation requires choreography between speaker and listener" [72] and such a choreography can be achieved through a rhythmic leading of the teacher or through an adaptive, gaze-contingent interaction between speaker and listener. We have shown that this latter approach makes the interaction more comfortable and consequently preferable to the majority of the subjects. However, the larger quantitative benefits are not for all of them but rather for participants with writing speeds more different from the average population, as the contingent approach allows them to exploit (or cope with) their specific characteristics. Therefore, a principle as simple as detecting the establishment of mutual gaze becomes for a robot an efficient mean to seamlessly interact with human partners with different needs in a turn-taking task.

# 5. Study III – Design and verification of an eye tracker

## 5.1 Introduction

Communication between people during daily activities relies on a series of multimodal cues, as speech, gestures, pointing, etc. among which gaze is one of the most important [7].

We hypothesize that enabling a robot to understand its human partner by exploiting gaze reading would substantially enhance the effectiveness and naturalness of the interaction. Indeed, gaze reading cuts times and delays in the interaction because it provides information in parallel to other forms of communication (as speech or gesture) reducing the complexity of the information transferred across those channels. For example, gaze can support speech understanding, by grounding otherwise ambiguous references to objects in the scene (e.g.: this/that). Similarly, in dialogs, turns are "directed" by the eyes, while contents are transferred through verbal communication and there is no need to stop the verbal flow of information to communicate "now it is your turn".

Our general goal is therefore to endow robots with an unobtrusive and natural gaze tracking system that would enable effective human-robot interaction (HRI) by augmenting the robot's perception. In the current study we introduce a geometric feature-based passive gaze estimation system for the iCub humanoid robot [10] to allow successful exploitation of gaze information in communication with humans. The ability is introduced as a software module integrated in the software framework governing the humanoid robot iCub, but our approach is general enough to be easily portable to any other robotic system.

First, for a background on existing eye tracking approaches, please see Section 2.3. Then we will provide a technical description of our system, followed by a validation of its performance. Finally, we will present the results of a realistic human-robot interaction task which benefits from our gaze estimation system. The study will be concluded with a discussion of future potential improvements and applications.

## 5.2 Implementation

Our goal is to implement, verify and exploit a monocular gaze tracking module for a humanoid robot, with the benefit of providing a flexible, cheap and easy-to-use solution, able to support a natural interaction with human partners. In this study we will implement the system on the iCub

platform [4].

The iCub humanoid robot has a highly sophisticated visual system. The eyes have three degrees of freedom (pan, tilt and vergence), while the robot's neck has three more DOFs [13]. The integrated control mechanisms allow saccade-like motions and reproduce the vestibulo-ocular reflex.

In this work we used PointGrey DragonFly 2 cameras with XGA (1024x768) resolution which are compatible with the iCub visual system. However we will provide evidence that the system works also for the standard iCub visual system which consists of the same camera type but with VGA resolution (640x480). We note that the used cameras employ fixed-focus lenses, thus we identified the operational distance for the tasks of our interest to be in the range from 60cm and 100cm from the camera and we set the focus of the camera accordingly.

Gaze tracking can be divided into several subtasks, which will be introduced in detail in the following paragraphs.

### 5.2.1 Image acquisition

We acquire the images from the iCub robot's software architecture. Two high resolution images (XGA) can be acquired at about 25 frames per second in the Bayer color format. Our algorithm uses the right eye's image. We opted for using the images of only one of the two cameras because 1) all necessary parameters can be calculated even from one image and 2) because of performance considerations: processing two images would significantly slow down the algorithm rendering it hardly usable for real time HRI.

### 5.2.2 Face and face features detection

To accomplish this step we rely on the dlib library [64] which implements Khazemi and Sullivan's algorithm for precisely detecting a 68-point face feature set using ensembles of regression trees [65]. This algorithm starts from the Viola-Jones [73] rough detection of the location of faces in the image. Once this is determined, the 68 feature points are aligned to specific features of the human face: the contour of the jaw, nose, eyes, eyebrows and mouth, see Figure 16.

**Figure 16. Left: face features detection as seen by the iCub. Right: eye area detection (top right) and extracted eye area with detected pupil center as seen by the iCub (bottom right).**

### 5.2.3 Eye area and pupil center extraction

Both eyes are bound by a 6-point polygon as seen in Figure 16. These areas are masked and extracted to allow the detection of the center of the pupil. Our approach compensates for imprecision in the focus settings of the camera and the relatively low resolution of the eye image. We solved these potential issues by locating the averagely darkest area in the masked image of the eye which almost always corresponds to the iris.

### 5.2.4 Head orientation detection

In order to detect head orientation, a 3D model of the face is needed. The previously described face feature localization algorithm (Section 5.2.2) provides only a 2D representation of face points. Thus we adapted the constrained local models (CLM) approach by Baltrusaitis et al. [74]. This solution gives us not only the 2D face features, but also the 3D model of the face complete with head position and orientation information. We decided not to employ CLM also for locating the contours of the eyes, as the previously mentioned algorithm by Khazemi and Sullivan seems to provide a more reliable estimate of these important points. Nonetheless, CLM provides us with the needed head orientation which we will use in the subsequent steps of gaze estimation. We assume that the head is kept in the middle of the camera image by the face tracking algorithm.

### 5.2.5 Eye model geometry

In detecting the gaze we opted for a feature-based approach with using an eye model (cf. Section 2.3). The method which we are adapting from Ishikawa et al. [46] can also be called geometric, as it maps eye features (center of pupil, corners of the eyes) to a 3D model of the eyes. The eye model is not to be confused with the previously mentioned 3D model of the whole face, which is used only for determining head pose. In order to simplify the approach, the eye model approximates the visual axis of the eye (i.e., the line from the center of the pupil to the fovea) with the optic axis (i.e., a line connecting the center of the anterior curvature of the cornea with that of the posterior curvature of the sclera) and also considers the eye to be spherical. In this sense, to find a human's gaze direction we connect the center of the eye with the center of the

pupil, which is located on the surface of the eye. The angle between this line and the line connecting the camera with the center of the eye is the angle of gaze (see Figure 17). The center of the eye is not moving compared to the head, so it is possible to express it in relation to some other immobile point on the human face which is close to the eye, e.g. the middle point between the two corners of the eye. Therefore, we determined the 3D displacement of the center of the eye compared to this midpoint. Knowing these distances, as well as the radius of the eye, allows us to estimate the gaze of a person.

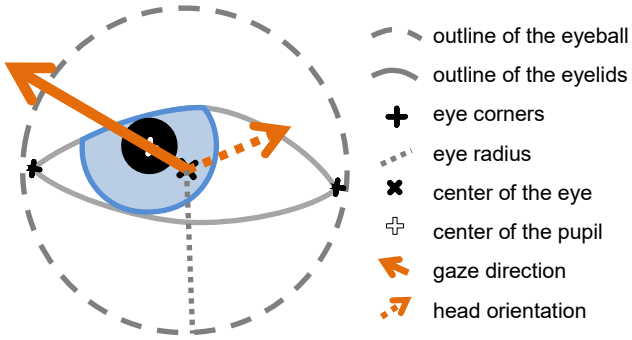


**Figure 17. The model of the eye.**

### 5.2.6 Averaging eye model values

To compute the radius of the eye and the distances of the eye center from the midpoint between eye corners it would be necessary to run a calibration process, during which the subject would have to look at a number of points on a screen with known angular distances (details in [46]). However, repeating the above described calibration process for each subject would be cumbersome and could negatively affect the interaction with the robot. The calibration instructions indeed could induce participants to monitor explicitly their otherwise unconscious gaze behavior also during the task, making it unnatural. Thus, we decided to make our gaze estimator subject-independent, by calculating a priori an average value of the above mentioned eye model parameters. It is clear that such approximation will lead to some degradation of the estimation, but it adds much to the naturalness and rapidity of the interaction. These average measures could also be considered as an initial guess, which could be improved during the interaction period (see Section 5.5).

**Figure 18. Gaze tracking estimates using our system (subject distance ~80cm).**

Averaging is made possible by exploiting the fact that the distance between the eyes of people (inter-pupillary distance) shows surprisingly little variance: mean of 62.3mm for women and 64.7mm for men with a standard deviation of 3.6mm and 3.7mm respectively, according to [75]. Thus, we normalize each face by the distance between the two outer corners of the eyes, which is a very similar measure as the inter-pupillary distance. Once this assumption is made, we can average the normalized values of the eye model and obtain a general model that "fits all". For this averaging process, we chose as training corpus the Columbia gaze data set, that includes high resolution and high quality images of 56 subjects looking at predetermined angles with different head rotations [67]. Twelve subjects of this data set were eliminated from training as the face feature detection algorithm struggled with recognizing some of their images, mostly due to reflections on their eyeglasses. We did not eliminate all subjects with glasses, as on some of them the algorithm was sufficiently reliable. The performance of the developed system is visualized in Figure 18.

## 5.3 Validation experiment

Once the generalized parameters of the eye model were calculated, we proceeded to verify their usability on data acquired on the humanoid robot, which is our target platform. We chose two interaction distances: 60cm and 100cm, representing near and far values within the social distance for interaction [76]. We created a board with markings at -20, -10, 0, 10 and 20 degrees both in horizontal and vertical for the near and far distances, with the exception of -20 and 20 degrees vertical for the far case, because of space constraints (available board size). This board was put in front of the iCub with a small hole in the middle for the robot's eyes, while the subjects sat on the other side at the two selected distances and looked at the marked angles sequentially. Subjects were first required to look at the board keeping their head straight toward the robot, and then rotated by 15 degrees to the right and to the left. The validation thus consisted of 25 points x 3 head rotations for the near distance and 15 points x 3 head rotations for the far distance, yielding to a total of 120 points per subject. We chose +/- 15 degrees of head rotation because both eyes are still clearly visible under these angles. We tested 8 participants, 6 men and 2 women between the ages of 23 and 36 years. One person wore glasses. We collected images of the subjects from the robot's eye and estimated gaze using the above described method and by using the CLM algorithm to acquire head pose. To reduce the noise in the results eye gaze was averaged between the left and the right eyes.

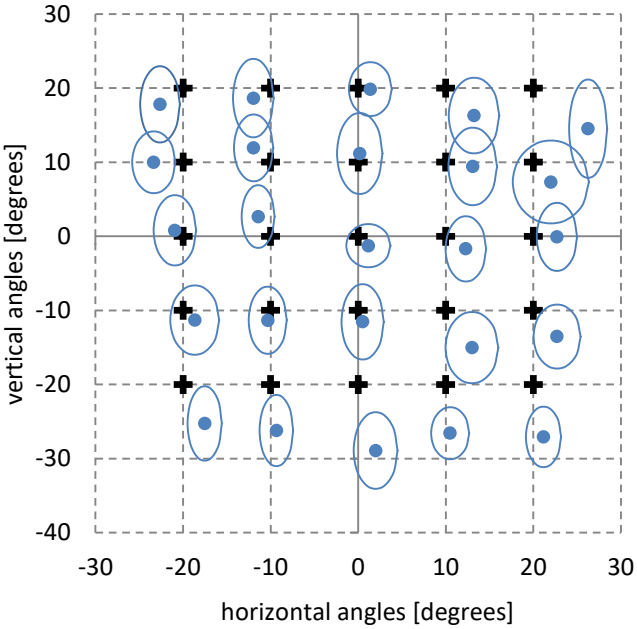


**Figure 19. Nominal (plus marks) and estimated gaze locations (blue dots) with 1 standard error ellipses for the validation experiment at 60cm distance.**

Figure 19 shows the distribution of gaze around points of gaze for the near interaction distance, averaged across the three head orientations. The black plus marks represent the position of points on the calibration screen while the blue dots stand for the average values of the corresponding gaze directions estimated by the robot. The blue ellipses indicate standard error of the estimate. The shape is elliptical because the variances, and thus the standard errors, are different in the vertical and horizontal directions. A more compact view of the results is reported in Figure 20, where the estimated angles averaged across all head rotations and both the viewing distances are plotted against the corresponding nominal angles separately for the horizontal and vertical components.

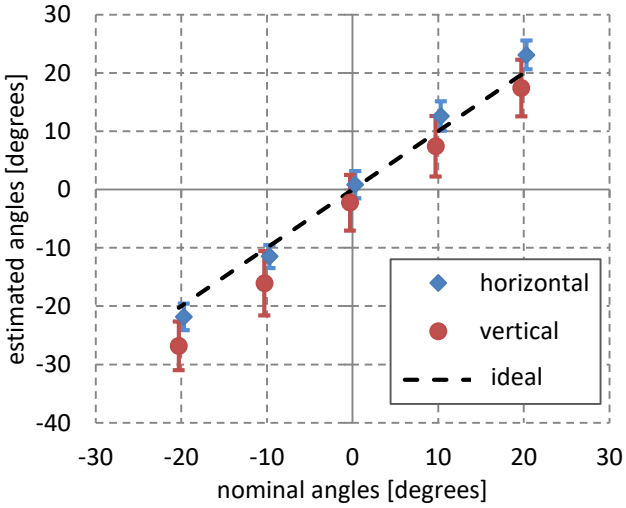


**Figure 20. Average estimated gaze angles (+/- 1SE) along the horizontal (blue) and vertical (red) directions. Average across all subjects, distances and head orientations. Values for -20/+20 vertical are computed only on the near distance.**

From both figures it emerges that the estimates of the horizontal component of the gaze were more accurate and less variable than those of the vertical gaze (see also Table 4). It is interesting to note that the performance of our system in the horizontal plane seems to be in a similar range as that of a human observer, as measured in a slightly different task by Al Moubayed and Skantze [77].

The lower accuracy in the vertical plane was caused by the imperfect detection of eye corners, since the estimation of eye corners position is lower than the actual location when subjects look down and higher when subjects look up. To counter this effect, we computed the average offsets in the training dataset and introduced a correction factor in the eye corner detections. However, even with this correction, the variance in the vertical data was considerably higher than in the horizontal ones. One reason for this additional variability is the fact that when people look down the upper eyelid covers a big part of the eye's surface, making precise pupil detection difficult. This is one of the reasons why commercial eye trackers are usually positioned below the level of the eyes, thus giving the cameras a better view of the eye when the person is looking down. In interaction contexts, however, the robot is often at the same level as its human partner, therefore we kept this relative positioning in our validation scenario, although it resulted in gazing down being harder to detect.

**Table 4. Average absolute errors of gaze estimation.**

| Angular error [degrees] | Near distance (60cm) | Far distance (100cm) |
|---|---|---|
| *horizontal error* | 4.96 | 5.33 |
| *vertical error* | 9.65 | 13.56 |

Considering interaction distance, the average absolute errors were larger for the 100cm distance than for 60cm, as expected. In fact, at 100cm distance the face and eyes appeared much smaller in the camera images and the algorithm had less information for estimating the center of the pupil and thus the gaze.

We evaluated also how such a gaze tracking system would work with the standard iCub cameras (VGA resolution at 640x480 pixels). To this aim, we simulated a lower resolution by down-sampling our original images to VGA. The analysis results for the near condition showed similar performance as the far condition with the higher resolution cameras, because the size of the face in pixels was very similar in these two cases (where avg. horizontal abs. error was 5.32° and vertical 13.4°). Hence, gaze detection for the near condition degraded by 33.1% when switching

from higher resolution cameras to VGA ones. This implies that in order to achieve similar performance as at 60cm distance with the high resolution cameras, the subject should be at a distance of about 37cm from the robot when using VGA cameras. Much of the imprecision in gaze estimation comes from the imperfect detection of eye corners and pupil centers, but part of it derives also from the averaging procedure, where the same eye parameters are applied to every subject. Nonetheless, the obtained accuracy is sufficient to enable the robot to distinguish with 75% probability which of two objects is looked at by a human in front of it at a 60cm distance, if the objects are 6.2cm apart horizontally, at half distance between the robot and the human. Hence, we deemed the system performance high enough to be useful in a real helping scenario and we proceeded with a HRI experiment to verify its usability.

## 5.4 Human-robot interaction experiment

Once the gaze estimator's performance was verified, we tested the designed system in a proof-of-concept HRI experiment. This collaboration scenario was set up to assess both the robot's ability to perform turn-taking, by detecting eye contact with the subject, as well as its ability to recognize the focus of the partner's visual attention on different objects. The experimental setup is portrayed in Figure 21.

Participants sat opposite of the robot and the experimenter and were required to stack up four numbered toy building blocks on top of each other in ascending order. Both the robot and the experimenter had one building block in each of their hands. Neither of them knew the numbers on the blocks, which were visible only to the subject. Each participant was instructed to get the building blocks in rising order one by one and stack them up on a nearby table. The only additional instruction was that the building blocks could not be taken before being passed by one of the "helpers". Subjects were free to adopt any possible communication strategy to ask for the blocks and were naïve towards the goal of the research.

**Figure 21. HRI study setup: the robot is offering a building block to the subject with its left hand. Lower right: gaze estimation by iCub on the subject before the offering action.**

The robot's behavior was programmed to make an offering movement with the appropriate arm (lifting its hand higher up and towards the subject for 2 seconds), when it detected the subject either looking in sequence at the object and then establishing mutual gaze or vice versa establishing mutual gaze and subsequently looking at the object, see Figure 22. During the whole interaction the robot was following the subject`s face by turning its eyes and head towards the middle of the detected face. The test was repeated 5 times with each subject with random positions of the numbered blocks (for an example of the experimental procedure see the video[4] accompanying [4]). The experiment was completed by 7 subjects (5 men, 2 women, between 25 and 32 years of age). Three subjects wore eyeglasses.

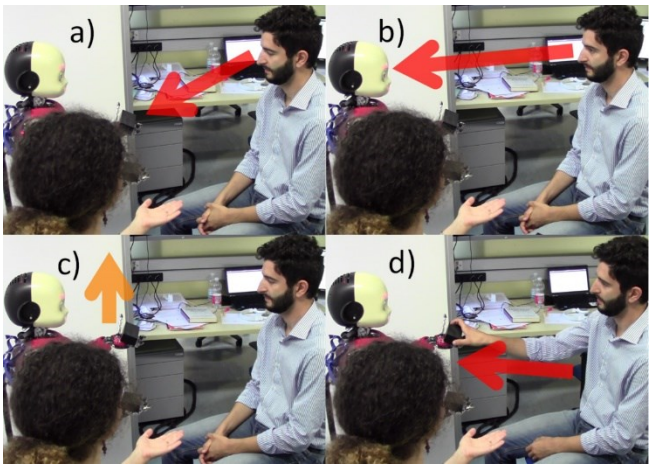


**Figure 22. Interaction sequence: a) participant looks at robot's hand, then b) at the robot's face. c) The robot lifts its arm and finally d) the subject takes the block.**

To get iCub's attention participants used a combination of speech commands, hand pointing and gazing at the blocks. As a result, all block stacking tasks were successfully completed, except one, in which there was a technical failure on the robot (97% of task completion rate). Even though participants did not know what triggered the robot's offering behavior, all of them

[4] https://youtu.be/KmB1QXiVkpM

succeeded in activating it. When interviewed at the end of the experiment, most subjects (5 out of 7) were convinced that either their verbal instructions or their pointing gesture alone directed the iCub and did not mention gaze. Over the 5 repetitions participants also tended to become more proficient at executing the task, as the average time to complete the stacking dropped from 42.7s for the first trial to 33.7s for the last one.

We performed a further analysis to assess robot performance in: 1) distinguishing its turn (i.e., being asked to perform a task, signaled by mutual gaze and glance at hand) and 2) during its turn, detecting the gaze either on the left or right hand. In the first task, the robot achieved a success rate of 83.0%, with 10.6% errors being false negatives (the robot did not react when it was gazed upon) and 6.4% false positives (the robot reacted when it was not gazed upon). Out of all the times, when the robot successfully detected its turn for action, it performed correct hand-over of the proper block 69.6% of the times, while 30.4% of the times it lifted the wrong hand. It should be noted that almost half of these errors occurred in the interaction with a single subject, for whom the robot could not determine the right selection repeatedly. These errors might have been caused by the subject's eyeglasses, even though our algorithm worked quite well for two other subjects wearing glasses. We hypothesize that this malfunction might have been caused by the type of glasses, since different shapes of eyeglass lenses can have different effects on eye tracking systems [78].

## 5.5 Discussion and conclusions

The validation and proof-of-concept scenarios proved that the proposed system is a viable low-cost, passive, calibration-free gaze tracking solution for humanoid platforms. The solution is low-cost as it can detect human gaze with any kind of camera which is at least in VGA resolution. Our system not only works with webcams, but could also benefit from some of the advanced options of these low-cost devices (high resolution, auto-focus, auto-white balance, hardware image compression). Traditional computer vision cameras provide higher quality optics, however in this scenario we compensate for lower quality optics of the cheaper devices by procedures for calibrating and rectifying images provided by OpenCV. The proposed gaze tracker also does not require additional infrared illumination pods, which makes it cheaper, more flexible and more natural.

We proposed an averaging procedure for the parameters of the eye model, which allows us to apply gaze tracking without calibration for each subject. This process certainly introduced more variance to our system, but it facilitated the interaction with naïve subjects. To mitigate the

problem in the future it would be possible to use "soft calibration" methods to increase the accuracy of the system, by adjusting the eye model parameters of an individual on the fly when we can assume that the robot could know the subject's gaze direction from the context, e.g. when the robot actively shows something to an engaged partner.

The estimation of vertical gaze could be further improved with better face feature recognition. Since our software is designed to leverage on modularity, substituting certain modules with more effective alternatives will be facilitated.

Our gaze tracking algorithm might not work well with certain types of very thick eye glasses, as mentioned before. Also it is not effective on great distances (>2m) because of the finite resolution of the cameras.

The benefits of a built-in gaze tracker in a humanoid can be various: it could improve turn-taking, joint attention and in general the processing of all the communicative gaze cues typical of human interaction. For instance, in conversations the robot will be able to detect events like mutual gaze and gaze aversion, which both can be used in naturally establishing turns in verbal exchanges. Moreover, the robot's ability to detect the partner's attention on objects can give it more "intuition" in knowing which object the human coworker is interested in. A first evidence of this claim comes already from the HRI experiment we presented (Section 5.4), where we demonstrated that even gaze alone sometime is enough to drive human-robot interaction to success. Furthermore, the robot could potentially be used for diagnosing early behavioral problems associated with gaze processing as Autism Spectrum Disorders, by monitoring subjects' gaze in real time. Indeed, a diagnosis based on gaze analysis has already been suggested to be promising [79] although so far it could be obtained only with a lengthy a posteriori manual annotation of video recordings of interactions. Our system would give the additional possibility to appropriately adapt robot reactions to special needs during the interaction, something that nowadays often requires human intervention or Wizard of Oz scenarios [80].

# 6. Study IV – Comparing eye gaze to head pose for HRI

## 6.1 Introduction

Humans are great at communication with their gaze, while robots mostly lack this ability and use head pose as a first approximation to gaze (for a more general introduction to the topic, please see Section 1). In natural collaborative scenarios objects of interest are often close to each other and people tend to switch their focus of attention just by moving their eyes, yielding to minor or null head movements. The inability to read actual eye movements could then make the robot miss important information for an efficient interaction, like which object the human collaborator is attending to.

In this work we add performance improvements to our calibration-free, visual light, monocular eye gaze tracking algorithm designed to work on humanoid robots as presented in the previous study. This system enables a robotic platform to catch the subtle communication signals associated with human eye motion during collaboration with no need of ad hoc hardware or high resolution, narrow field-of-view cameras. Using this system we then quantify which advantage an eye gaze sensitive robot could bring in a common interaction task with respect to the head gaze based solution commonly adopted. We consider a collaborative scenario in which human participants have to build a tower out of toy building blocks, in part handed over to them by the robot. This is the same scenario described previously in Section 5.4. We measure the performance of the interaction when the robot is programmed to monitor the eyes of the naïve subjects to detect which block they are interested in. Then we compare it with the performance in a condition in which the robot is only sensitive to head orientation.

The next section will give an overview of previous work on head and eye gaze tracking in robotics, and position our study in this field. Section 6.3 will describe our gaze tracking system. Section 6.4 will be dedicated to the experiment we designed to evaluate the efficiency of eye gaze tracking versus head gaze tracking alone. Section 6.5 will summarize our findings concerning the accuracy of the employed system. Section 6.6 will show behavioral results, Section 6.7 will show gender differences, while Section 6.8 will highlight subjective measures. Subsequent sections will discuss these results and propose conclusions.

## 6.2 Study background

For the discussion of different gaze tracking solutions please refer to Section 2.3.

The gaze tracking approach that we believe is most promising for robotics, and in particular for robot companions, is passive remote calibration-free gaze tracking. This choice indeed guarantees the highest degree of naturalness in the interaction. With such a gaze tracker a robot could read the user's gaze seamlessly, with no need for additional hardware (e.g. to produce infrared light), no physical encumbrance to potential interacting partners (e.g., helmet or glasses), nor need for preparation to the interaction (for calibration purposes). Thus in the rest of this section we will be focusing on this type of gaze tracking systems.

In the field of human-robot interaction it has become a common practice to replace eye gaze with its approximation, head gaze. Doniec et al. describe a method for learning joint attention by a robot [49]. In their approach the robotic agent observes the caregiver's gaze towards certain objects. However, eye gaze is replaced by head pose, because the authors claim that eye gaze was not possible to extract due to the low resolution of their cameras. Using a Radial Basis Function Network they were able to train the robot to recognize joint attention towards a number of objects on a table and then recognize the selection of objects based on head pose. They report a recognition rate of 95% when testing is done with the same person as training, but 62% when different people are used in training and testing.

Kim et al. reported on a robotic system capable of learning gaze following [50]. They used head pose estimation, because they did not have an eye gaze tracking system available. Their system was able to learn correct associations between the caregiver's head pose and corresponding motor actions using offline reinforced learning.

Ivaldi at al. presented an experiment on robot initiative during a collaborative task with a human [38]. In this publication eye gaze is replaced by head orientation, acquired using an RGB-D sensor. They claim to be able to detect head yaw with the accuracy of 93%. The authors underline that such an estimation of gaze is inaccurate, but that it has the advantages of not necessitating external eye tracking devices or high resolution cameras, keeping the interaction natural and non-invasive. In this study participants were at a distance of >1m away from the robot.

Sheiki and Odobez explore attention recognition in HRI [81]. They claim that most current systems approximate gaze with head pose because eye gaze estimation is often impossible to achieve. The authors use a Hidden Markov Model to dynamically decode the visual focus of attention. They propose using context to improve the detection of visual attention.

Nagai et al. conducted a study on how a robot could learn joint attention [31]. A neural network

was used to associate the visual appearance of a caregiver's face with object angular displacements. The term gaze is used throughout the publication, but the objects of learning are face images, with distinct head and eye rotations. The authors emulated visual development by de-blurring blurred images of the caregiver.

In the above publications [49][50][38][81] authors either used direct substitution of gaze with head pose or utilized contextual information for training their machine learning methods. Sheiki and Odobez [81] in addition to context also used dynamic mapping from body posture to gaze using head pose. As we aim to make a general system, independent of context, in this study we use head pose as a direct substitution for gaze and compare it to eye gaze itself.

Admoni et al. performed an experiment from the opposite point of view: humans observing a robot's gaze [41]. The humanoid in question was HERB which was programmed to hand over objects while performing different gaze actions: natural gaze, joint attention and mirrored gaze. The authors mention the robot's eye gaze, even though the robot's head consists of a camera and a microphone mounted on a platform with a pan/tilt mechanism. The authors find that it is possible to influence the conversation even with an approximation of a robot head.

All these results show that head orientation might provide a useful approximation for eye gaze in human-robot applications under certain conditions. However, it is often suggested that this solution is less accurate than having access to eyes gaze (e.g. [38]). The technical issues or the non-naturalness of the adoption of traditional eye trackers though forced the choice of this approximation, although no quantification of the information loss has been done so far, due to the strong dependence on the settings (task, distances, cameras, etc.)

On the other hand, there are already studies where actual eye gaze tracking is used in humanoid robots. As one of the pioneers of this field Matsumoto and Zelinsky developed an eye gaze tracking system [53] which was implemented on the HRP2 humanoid robot. They use a least squares method to align certain features on the user's face with a 3D face model. Once the face is tracked, an eye model is applied to the image of the eyes to estimate gaze direction. They used this gaze tracking system on a humanoid robot in a dialog scenario, to appropriately detect eye contact with participants.

On a similar vein our group has developed a mutual gaze detection system using which we were able to control a turn taking scenario on the iCub humanoid robot, see Study II in Section 4. In this study the robot waited for the user to glance back at it, before continuing to dictate sentences. We then expanded this detection algorithm into a full-fledged modular eye tracking system (see

Study III in Section 5), achieving accuracies of 5 degrees average absolute error in horizontal gaze estimation. With this system we completed a proof of concept HRI experiment in which we found that the robot was able to understand which object to hand over to its human partner by using only gaze information.

Our current study is a continuation of this line of research, where we present an improved version of our gaze tracking robot and we confront its performance (based on eye gaze) with its traditionally accepted approximation "head gaze" in a common HRI scenario.

## 6.3 Methodology

In the following we will shortly describe our original eye gaze tracking system first introduced in Study III [4] and emphasize the performance improvements we made since then.

We implemented a model-based, visual light, monocular, calibration-free, remote gaze tracking algorithm, using existing head pose and face feature tracking algorithms (see Figure 23). Head pose was calculated using the Constrained Local Models (CLM) approach implemented by Baltrusaitis [74]. It provided us with the head orientation that is directly used in the head gaze experimental condition as well as in calculating the eye gaze. Face features were found using the approach described in [65] and implemented by King in [64]. This provided us with robust tracking of locations like the corners of the eyes and mouth. Once the eye region was located we used averaging methods to find the center of the darkest area of the eye, which approximates the center of the pupil, due to the light color of the sclera. Averaging consisted of blurring the image and finding the darkest area in it. Once we found the locations of the eye corners, pupil center and head orientation, we applied these points to an eye model with the goal of calculating gaze angles. The parameters of this model were estimated in a least squares approach on the Columbia gaze dataset. The approach was adopted from [46]. This allowed us to create an eye model for a "generic subject", thus eliminating calibration for each new user. We verified the newly obtained eye gaze model by assembling our own gaze dataset using the iCub's eyes. We found that the mean absolute error in horizontal gaze was around 5 degrees, while for the vertical gaze it was 9 degrees. The larger error in vertical gaze detection stems from the fact that when people look down their upper eyelids covered most of their eyes, which in turn causes imprecisions in detecting the eye corners and pupil center. None the less the robot equipped with this system could reliably understand which objects are gazed upon by its interacting partner, as described in Section 5.

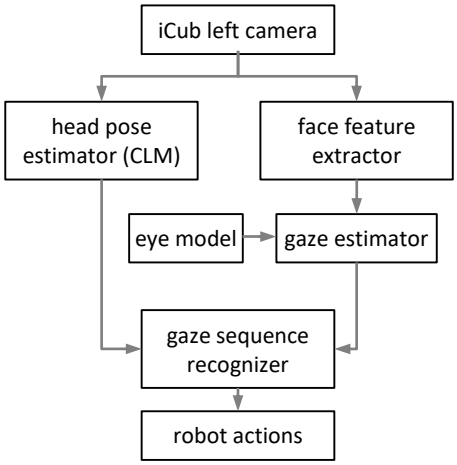


**Figure 23. Eye tracking system diagram.**

The original system had a slow throughput rate, at around 7 frames per second (FPS), mainly due to the computational complexity of the face features detection algorithm. In our current approach we were able to optimize its performance, by performing face detection only in the area where the previous frame contained a face. Once a face is lost, the face detection is performed again on the full frame (1024x768 pixels). This allowed us to achieve framerates of around 22 FPS (see Figure 24), which was adequate for capturing even faster eye movements. To deal with possible momentary glitches triggering gaze events in our robot, we applied temporal smoothing of the gaze signal with an averaging filter window of around 1s. This smoothing also allowed us to distinguish between glances down and blinks, which look like very short downward glances, increasing the robustness of the system. These changes made the estimated gaze signal more robust and precise.


**Figure 24. Example output of our eye tacking system.**
**White lines from eyes – eye gaze, black line from top of the nose – head gaze.**

## 6.4 Experimental setup

In order to evaluate whether eye gaze can be substituted by head gaze with no significant loss of information, we constructed a human-robot interaction scenario involving three agents: an experimenter, the robot and the subject, see Figure 25 [5]. The subjects started each experiment

facing the robot. Even though they could not move their legs freely (due to iCub's platform), during the experiment they moved their upper body and head to face both agents. As it can be seen in Figure 25, the separation between the robot's two arms was around 50 degrees as seen from the perspective of the subject. This separation was dictated by robot kinematics and the need to sit the subject as close to the robot as possible (~100cm) for more precise gaze tracking while maintaining a comfortable social distance.

We had two experimental conditions: eye gaze and head gaze. The task of the subject was to build a tower out of four toy building blocks, as described in Section 5.4. At the beginning of each build, the four blocks were located in each of the hands of the experimenter and the robot. The blocks were numbered from 1 to 4, but these stickers were visible only to the subject (to prevent them from asking for the block by number). The blocks needed to be stacked on the provided table in ascending order. The subjects were instructed that they had to take the blocks, which they needed to ask for, only after they were offered to them. We did not tell them what interaction modality to use to achieve this offering motion - a movement of the hand up and towards the subject.

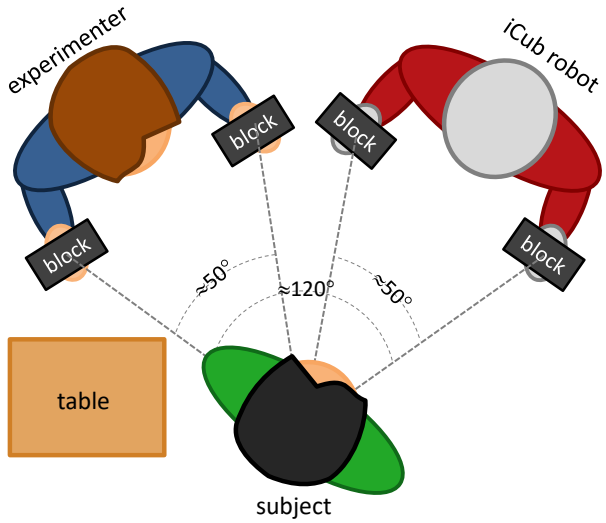


**Figure 25. Experimental setup.**

The robot was programmed to react either to the eye motion or head motion, namely eye gaze and head gaze conditions respectively. The sequence triggering the offering motion for the first condition was the subject's glance first at the face of the robot and then at one of the hands (Figure 26, left) or first at one of the hands and then at the face (Figure 26, right). This activation sequence was inspired by the definition of joint attention, which requires not only that two agents look at the same object, but also that they are aware of the attention target of the other, an awareness often obtained by establishing eye contact before and/or

after gaze following. In the head gaze condition, eye gaze was replaced by head movements from and to the face and hands.

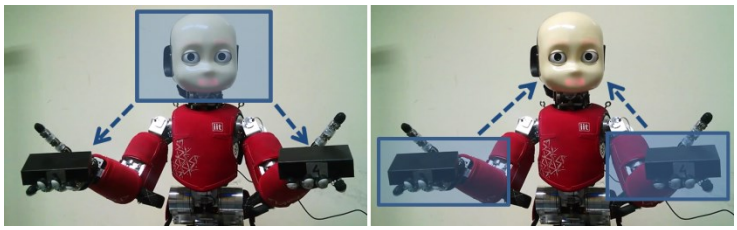

**Figure 26. Gaze actions triggering robot reaction: a) gazing at face then one of the hands and b) gazing at a hand then at face.**

Since our vertical gaze detection is less accurate than horizontal, we set fixed box boundaries of where the gaze originated from but did not do so for the end point. Rather we looked at how much of a vertical displacement over time there was from the initial upper (head) or lower (hands) gaze boxes, i.e. we looked at the vertical angular velocity. The thresholds for these velocities were selected based on testing with five pilot subjects in order to minimize both false positives and false negatives in triggering the offering actions in both conditions. For the gaze option the threshold was set to 20 degrees change per second, while for the head pose it was 2 degrees pitch over one second. For example in the head gaze option if the gaze started from the head box and pitched down at a rate greater that 2 degrees per second towards one of the hands the hand offering action would be triggered. The left hand was selected for movement if the pitch was towards down-left, the right if it was toward down-right. The actual vertical displacement needed to turn one's attention from the robot's eyes towards the block in the hand was around 30 degrees. The big difference between the triggering thresholds for eyes and head was due to the fact that all pilot subjects tended to cover this angular distance with larger movements of their eyes and a smaller rotation of their head.

Importantly, both in the eye gaze and in the head gaze conditions the robot was only sensitive to subjects' eye/head gaze and disregarded any other source of information (speech, pointing, and gestures). Subjects however were not aware of this limitation and hence behaved naturally toward the robot, as if it could perceive all these other signals. This approach allowed us to single out the effectiveness of eye/head gaze tracking alone, without any influence on the natural pattern of interaction chosen by human participants. In situations when the robot had only one block in its hands, we didn't apply the common sense logic and hand over the only available block when gaze tracking indicated the empty hand. We still offered the hand selected by gaze (even if it was empty) because we wanted to test the performance of our eye tracking system.

The subjects were asked to complete 5 towers for each of the two conditions. The order of the

conditions was counter balanced. The order of the blocks in the hands was pseudo-random, making sure even distribution of different numbered blocks in hands. The conditions were named Alpha and Beta for the subjects. Before the experiment the participants filled out an institutional consent form, while after the experiment they filled out an experiment questionnaire asking them to compare the two conditions and a personality questionnaire. The subjects received written instructions about what they were expected to do. Using text-to-speech the iCub explained once again the task. It was also saying sentences like "Let's do it again", "Let's build another tower" at the beginning of each task. At the beginning of each session iCub reminded the subjects if it was Alpha or Beta. At the end it thanked the participants for taking part in the experiment. During the whole experiment the robot was programmed to follow the subject's face with its gaze. The iCub platform performed saccades which approximated human oculomotor actions: first turning its eyes towards the participant's face and then following with the head, while reproducing the vestibulo-ocular reflex, thus providing a natural interaction experience (see the video[5] accompanying [5] to understand how the robot behaved).

## 6.5 Gaze tracking results

The experiment was completed by 10 subjects (3 women and 7 men) with a mean age of 34.6 years. Two of the subjects wore eye glasses, and one was wearing lenses. These seeing aids did not stop the eye tracking algorithm from inferring the participants' gaze.

First we looked at the success rate of task completion (Table 5). All subjects managed to complete all tasks using eye gaze, but 5 out of 50 tasks could not be completed in the head condition, because 2 subjects just could not trigger the robot reaction in some trials.

**Table 5. Task completion success rate.**

| | total tasks | completed tasks | completion rate |
|---|---|---|---|
| *head gaze* | 50 | 45 | 90% |
| *eye gaze* | 50 | 50 | 100% |

Secondly, we analyzed task completion time to see if subjects were quicker with any of the two interaction modes. It turned out that using eye gaze, tasks were done much faster than using head gaze, see Figure 27. Applying a paired t-test we found that the difference between the two conditions is statistically significant [$t(9) = 3.171$, $p = 0.011$]. All of our figures, where needed, will indicate statistical significances with a horizontal line and star, and +/-1 standard error with

[5] https://youtu.be/SZdRKXWONso

vertical lines.

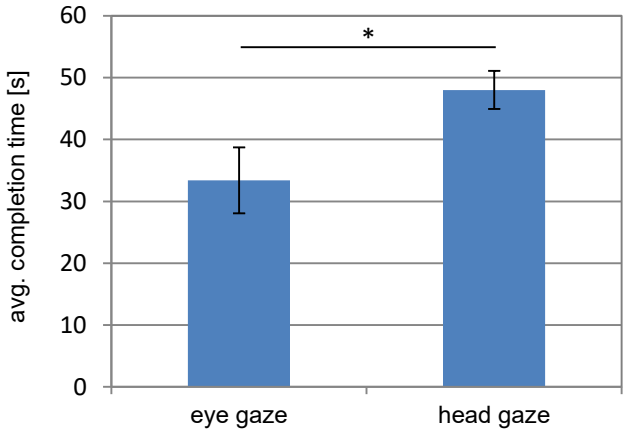


**Figure 27. Task completion time averages over all subjects.**
**Error bars represent +/- 1standard error. "*" indicates significant difference**

Next we looked at two measures of how successful the robot was in understanding human gaze behavior: 1) if the robot was able to detect if it was its own turn or the experimenter's turn to react, 2) within the successfully detected turns to react, how many times was the robot able to hand over the proper block. The results can be found in Table 6.

**Table 6. Average turn detection and handover accuracy percentage with standard error (also percentage) in brackets.**

| | turn detection rate | proper block handover |
|---|---|---|
| *head gaze* | 83.8% (3.4) | 60.6% (6.1) |
| *eye gaze* | 84.9% (4.2) | 81.5% (5.1) |

We must note that in our analysis we only counted situations in which the robot could be able to detect actions in its own interaction mode. For example, in head mode we only counted as false negatives cases when there was any detectable head movement from the subject but the robot did not react. The same procedure was applied for the eye mode. This was determined by the experimenter by reviewing and annotating the experiment videos.

From Table 6 it can be seen that in both conditions the robot was performing very similarly at detecting its own turn. The erroneous detections came mostly from transition periods: when the subjects were depositing the blocks on the table or when turning away from the robot to ask the experimenter for the next block. On the other hand there is a huge difference in success rates of handing over the proper block to the subject. In this measure the head gaze option was successful only around 60% of the time, while in the eye gaze option the robot successfully handed over about four out of five blocks. To further analyze these data, we looked at how many times the robot lifted its arm (attempt) before handing over the proper block once it accurately determined

its turn (see ). For the head gaze option there were more cases when the robot was not able to tell the subject's desired block repeatedly (15%). This means that the subjects performed the same actions but repeatedly got the wrong block. This number was only 4% for the eye gaze condition.

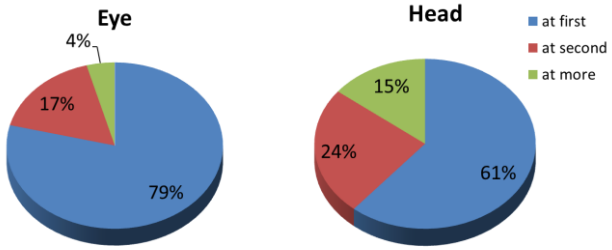


**Figure 28. How many offering motions it took before the robot handed over the proper block. "At first": the robot lifted the proper arm at once. "At second": the robot lifted its wrong arm before lifting the proper one. "At more": there were more than one wrong arm lifts before the proper one.**

The main reason for low numbers in proper block handover during head gaze lies in the fact that there was barely any head pitching from the subjects when they were asking for the blocks. They seem to have preferred to use their eye gaze, pointing and oral commands instead. There was also an interesting phenomenon on which we didn't count: some of the subjects rolled their heads while trying to get the objects. This was mostly the case when the robot did not react right away to their command. In this case they would roll their heads towards the block and keep it that way until the task was over. Neither of our eye gaze nor head gaze detection algorithms accounted for head roll, which might have caused a lot of the misrecognitions: we only expected pitching (nod) and yawing of the eyes and head.

In the experiment questionnaire we asked the participants to rate the likability, efficacy and smartness of the robot based on the two interaction conditions. We offered them a 7 level Likert scale to rate their answers. Figure 29 summarizes these subjective results. We found that although participants liked both modes of interaction (no significant difference in likability, paired t-test [$t(9) = -2.022$, $p = 0.074$]), they judged the robot as being significantly smarter and more efficient in the eye gaze condition [$t(9) = -2.689$, $p = 0.025$] and [$t(9) = -3.737$, $p = 0.005$] respectively).

When asked how the robot knew which object to hand over, four out of ten subjects didn't mention gaze in any way, four thought it was some combination of voice commands, gaze and gestures, while two were correct in thinking that it based its decision on gaze only.

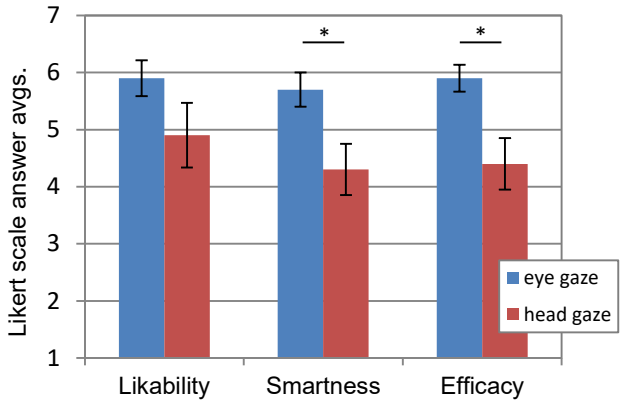


**Figure 29. Subjective ratings of the two interaction types.**
**Error bars represent +/- 1 standard error. "*" indicates statistically significant difference.**

## 6.6 Behavioral results

In order to delve deeper into the subjects' behavior with the robot, we recruited 22 more participants and administered the same experiment. Thus we ended up with 32 subjects in total, out of which 15 (46.9%) were women and 17 men. The average age of the group was 30.1 years with a standard deviation of 5.4 years. The majority of the subjects were Italian (75%) while the rest were: 3 Spanish, 1 Russian, 1 Belorussian, and 1 Swiss.

The larger sample was collected to validate the original results and to assess whether individual differences might interfere with the eye gaze based system. In particular, it might be hypothesized that introvert participants would refrain from exhibiting large eye gaze movements, potentially limiting the efficacy of the robot's contingent behavior. Indeed, extraversion has been found to be influencing gaze behavior in human-human interaction [82] which could also affect HRI [83]. Another hypothesis is that more neurotic participants could show more pronounced and noisy head movements, again interfering with the functionality of the tracking module, as neuroticism has been found to affect human-human communication [84].

Also, we wanted to assess whether the two gaze-based methods were associated to two different learning rates. If a progressive improvement was present for both head and eye gaze conditions, with the former being just slower, it might be pointed out that the benefits of the eye gaze based system would easily wear out after just a slightly longer training.

Finally, we were interested in evaluating whether gender differences might intervene already at the level of this very simple implicit communication exchange. Indeed, gender can also have a significant effect on gaze behavior, for instance women in general engage in eye contact more than men do [42], and this might impact also on their interaction with robots [85][86].

In this part of the analysis we looked at the following independent variables: gender, mode of interaction (eyes, head), and personality traits like extroversion, neuroticism and openness. The dependent variables we looked at were average task completion time, head movement distance, head azimuth variance, head elevation variance, mutual gaze percentage during task, eyes azimuth variance and eyes elevation variance. Average task completion time shows how quick each subject was to complete the tasks in the eyes and head condition on average. The shorter the completion time, the easier it was for subjects to complete the given task. It's a good indication of level of performance. Head movement distance is the angular distance traveled by the subject's head during the execution of tasks. Head azimuth variance gives us a measure of how much subjects moved their heads left and right, while head elevation variance measures up and down movement. The same applies for eye azimuth variance and eye elevation variance, only for the eyes, instead of the head. Mutual gaze percentage gives us the percentage the subjects spent looking at the robot's face in a task.

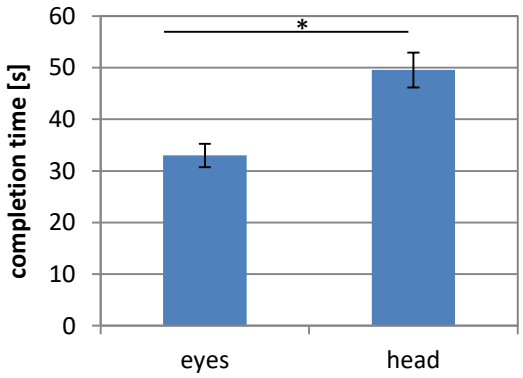


**Figure 30. Completion time between eyes and head condition. "*" indicates statistically significant difference.**

We have seen above that for 10 subjects the task completion time was already significantly lower for eyes vs. head. This result is confirmed also for the larger sample (Figure 30), with even higher significance (paired sample t-test [$t(31) = -4.84$, $p < 0.0001$]); the tower building is completed on average 33% faster when eye gaze is monitored (33.0s ± 12.8s SD for eye-gaze vs. 49.6s ± 19.0s SD for head gaze).

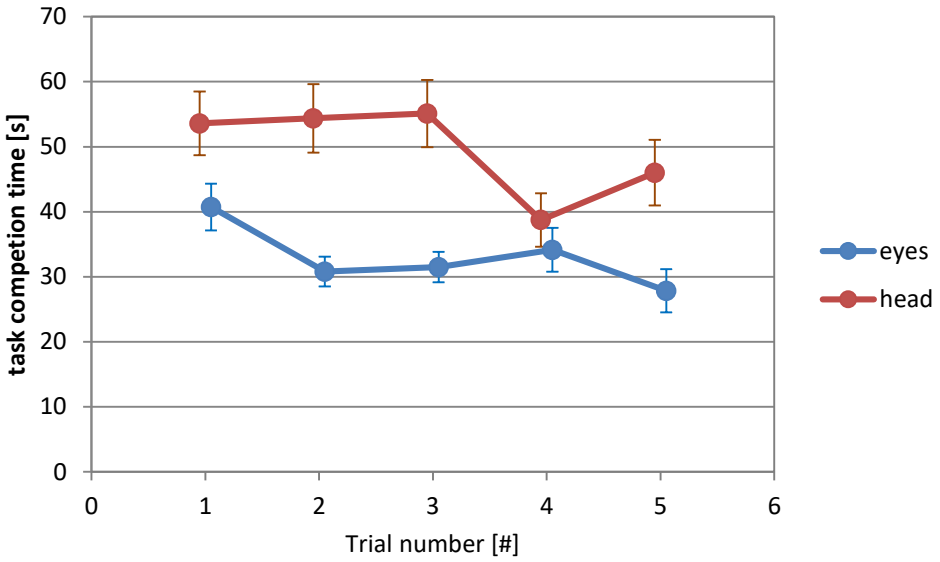


**Figure 31. Completion time over trials.**

To verify whether performance in the two conditions differed since the beginning of the experiment, or the discrepancy was due to a different learning rate, we evaluated how completion time varied as a function of trial number. Figure 31 shows that completion time was on average shorter in the eye gaze condition and it decreased progressively during the experiment. No evident decrease was instead present in the head gaze condition, which remained slower throughout the whole task. We ran a two-way Repeated Measures ANOVA with gaze-type (levels: eye gaze, head gaze) and trial (level: first, last) on the completion time. For the purpose of this analysis we replaced missing data – due to technical issues - with the completion time computed on the closest available trial. This happened on 7 of the 320 trials (2.19% of the data). Some subjects were not able to complete some tasks due to their inability to trigger the robot's actions. The completion time for these trials was replaced by the maximum allowed trial time of 2 minutes. We had 11 cases like this which represents 3.43% of all the data. The analysis revealed a significant effect of both factors [$F(1, 31) = 15.77$ , $p < 0.001$] for trial, [$F(1, 31) = 11.54$, $p = 0.002$] for gaze type, but interaction was not significant [$F(1, 31) = 0.801$, $p = 0.378$]. Bonferroni post-hoc comparisons evidenced that although completion time difference between the head and the eye gaze conditions during the first trial was not significant [$t(31) = 2.61$, $p = 0.08$], the difference was highly significant at the end of the experiment [$t(31) = 3.69$, $p = 0.005$]. Moreover, this change might be due to an improvement occurring for the eye gaze condition only: completion time tended to decrease between the first and last trial in the eye gaze condition (Bonferroni post-hoc [$t(31) = 2.62$, $p = 0.08$]), while this was not significant at all for head gaze (Bonferroni post-hoc [$t(31) = 1.54$, $p = 0.798$]).

The problem of missing data is particularly harsh in the context of Repeated Measures analysis.

Different strategies are used to cope with this issue - one of which is to replace the missing values with an approximate estimate of the value (imputation), which is the option we followed here. This has of course limitations (see for instance Chapter 25 in [87]). However, we are confident that the results in this case are robust to this manipulation. Indeed, since the missing data are due to technical issues, they can be seen as MCAR (missing completely at random) indicating that an analysis performed only on the subset of complete data would still be representative. That analysis actually confirms the same results as before: [$F(1, 24) = 10.33$ , $p = 0.003$] for trial, [$F(1, 24) = 6.19$, $p = 0.02$] for gaze type, but interaction was not significant [$F(1, 24) = 2.04$, $p = 0.17$].

## 6.7 Effect of gender

To evaluate whether the interaction unfolded differently as a function of gender, we replicated the analysis of the completion time by splitting the corpus in female and male participants. Interestingly the results seem to indicate that most to the difference between the performances in the eyes and head-based conditions comes from male subjects.

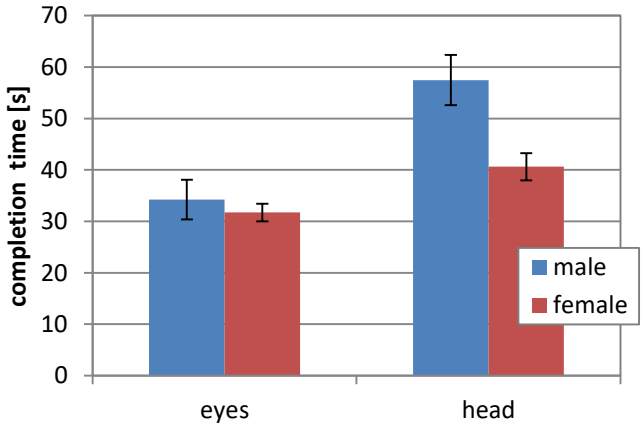


**Figure 32. Completion time with separated male and female subjects.**

In Figure 32 above it is visible that while for the eyes condition both men and women are similarly quick in completing the tasks, for the head option men are far slower than women. A two-way mixed-model ANOVA with type of gaze as within factor and gender as between factor reveals a significant effect of both gender [$F(1, 30) = 4.88$, $p < 0.035$] and gaze [$F(1, 33) = 24.98$, $p < 0.0001$], as well as interaction effect [$F(1, 30) = 5.01$, $p = 0.032$]. Post-hoc Bonferroni comparisons confirm that completion time in the head condition for male participants is significantly different from that of the eye condition (57.4s ± 3.9s SE vs. 34.1s ± 5.2s SE respectively, $p = 0.0001$), while for female participants this difference is not significant (40.6s ± 2.7s SE vs. 31.7s ± 1.7s SE respectively, $p = 0.41$). This post-hoc test also tells us that the

difference between men and women for head condition is significant ($p = 0.001$) while not so for the eyes condition ($p = 1$).

To understand the reasons behind such a difference between genders we analyzed different aspects of participants' gaze behavior in the task, to assess whether there were clear differences between groups. In particular we focused on the amount of eyes motion and head motion participants exhibited, quantified as the variance of the angular motions along the vertical (*Elevation*) and horizontal (*Azimuth*) axes, and on the percentage of time spent performing mutual gaze with the robot. We assumed a normal distribution of gaze angles both vertically and horizontally. Thus we calculated variance in both cases for eyes and for head motion, to get an idea how much each subject moved their heads and eyes during their tasks.

Figures 33-35 show how these different metrics varied as a function of gaze type and as a function of gender.

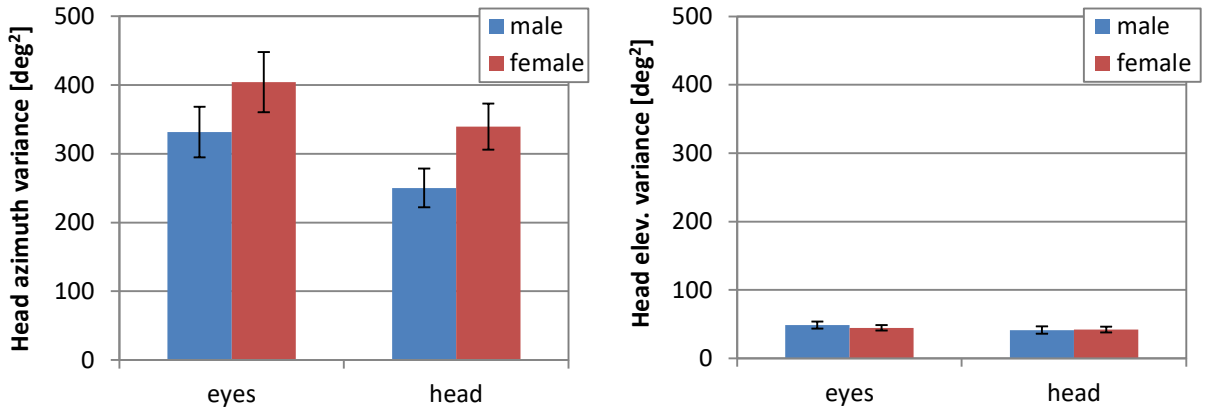


**Figure 33. Head azimuth (left) and elevation (right) variance for male and female subjects.**

Figure 33, left shows higher head azimuth variance for the eyes condition and higher variance for women vs. men both in eyes and head condition, i.e. women were moving their heads left-right more than men. Performing a Two-way Mixed Model ANOVA confirmed the significance of gaze-type [$F(1, 30) = 14.56$, $p = 0.0006$], but not gender differences. Head elevation variance in Figure 33, right did not show any differences regarding the observed variables, neither did eye azimuth variance in Figure 34, left. Eye elevation variance is shown in Figure 34, right. In both eyes and head conditions women show higher eye elevation variance levels than men, as confirmed by a Two-Ways Mixed Model ANOVA with gaze type and gender as factors, which revealed a significant gender effect [$F(1, 30) = 13.27$, $p = 0.001$], with no effect of condition nor interaction ($p = 0.37$, $p = 0.67$ respectively). This implies that women tended to move their eyes up and down more than men, regardless of the experimental task.

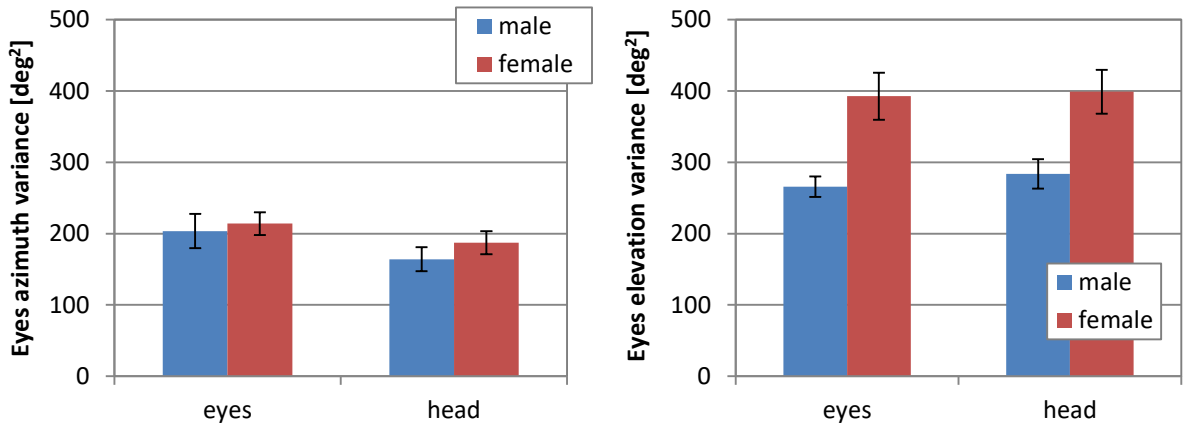


**Figure 34. Eyes azimuth (left) and elevation (right) variance for male and female subjects.**

Figure 35 shows that men tended to keep looking at the robot for somewhat longer time, but gender and gaze type differences were not significant when checked with a Two-way Mixed Model ANOVA.

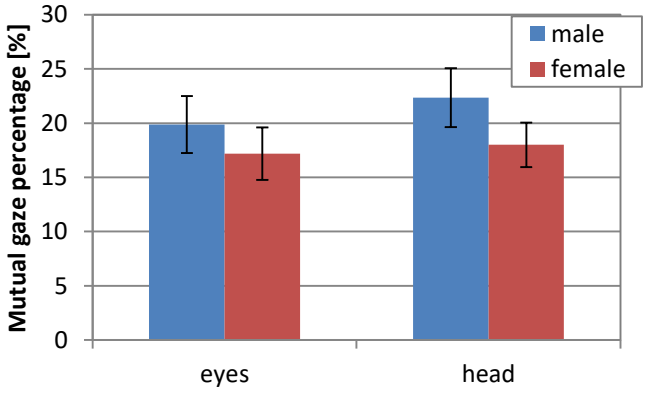


**Figure 35. Mutual gaze for male and female subjects.**

In this previous set of graphs we have found a tendency for women to move more their head and gaze with respect to men, which reaches significance especially in the up/down direction of their gaze. This higher mobility might be either a natural trait of the female gender or a sign that they were unconsciously trying to be more cooperative with the system by providing more ostensive cues, in all conditions – an effort that revealed itself particularly useful in the head condition, otherwise not very efficient. The current data does not allow us to reach a final conclusion if gender and/or gaze type affected head and eye behavior and thus the subjects' performance of the tasks.

## 6.8 Subjective measures

As mentioned in the previous section we administered a subjective opinion questionnaire to

gauge the participants' thoughts about the robot's efficacy, smartness and likability in the two operating conditions (eyes vs. head interaction). With additional subjects all differences became statistically significant using a two tailed Paired Two sample for Means t-Test, see Table 7 and Figure 36.

**Table 7. Subjective opinion statistics between eyes and head conditions**

| | likability | smartness | efficiency |
|---|---|---|---|
| *p-value* | 0.002715 | 0.000125 | 1.85E-05 |
| *t-value* | 2.039513 | 2.039513 | 2.039513 |

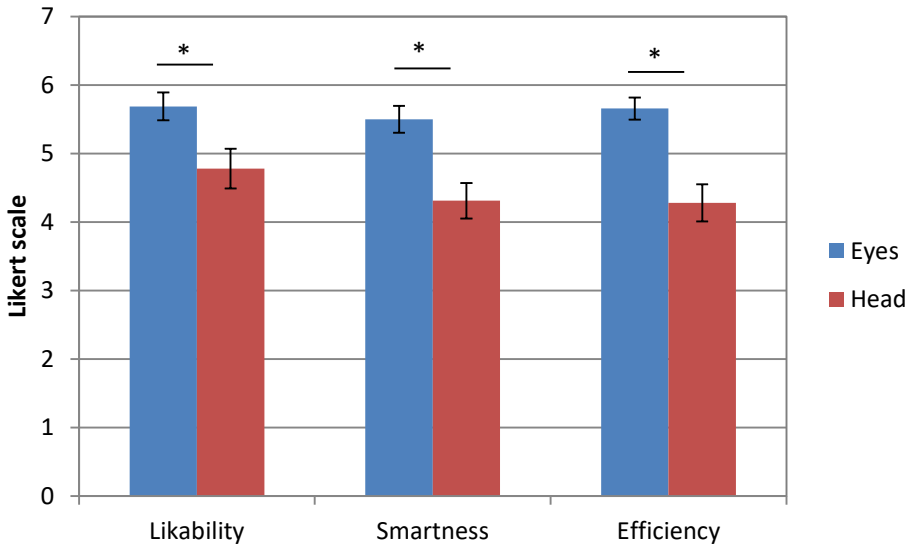


**Figure 36. Subjective opinion of participants between eyes and head condition. "*" indicates statistically significant difference.**

Knowing the previous differences in behavior between male and female subjects, we decided to split the subjective data up by gender too. It can be noticed in Figure 37 that all the differences between eyes and head for likability, smartness and efficacy are larger for men than for women. It can also be noticed that men constantly judged the head option with lower Likert values than women, while for the eyes option the opinions were more even. This signifies that men had a worse experience with the robot, while operating in the head mode thus they gave it a lower mark. This makes sense because we have seen in Figure 32 that men were much slower to complete the head tasks, as they were unable to activate the robot, contrary to women. This resulted in higher likability, smartness and efficiency rating for the head option for women compared to men.

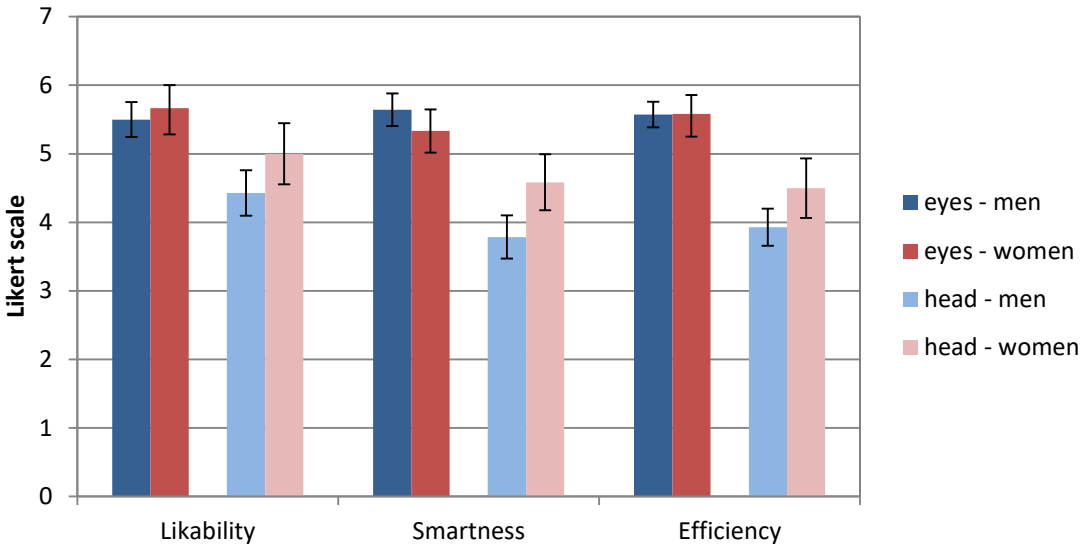


**Figure 37. Subjective measures between male and female subjects.**

To quantify these gender differences we performed a Two-Ways Mixed Model ANOVA with gaze-type and gender as factors for all three subjective factors. For Likability the only significant difference was for gaze type (head/eyes) [$F(1, 30) = 10.135$, $p = 0.003$]. This proves our finding from above, but does not show difference between women and men. The ANOVA for Smartness showed significance for gaze type [$F(1, 30) = 20.76$, $p = 0.00008$] and for interaction between gaze type and gender [$F(1, 30) = 5.96$, $p = 0.021$]. The Bonferroni post hoc test shows that the difference between head and eyes conditions is significant for men ($p=0.0001$) but not for women ($p=0.943$). This proves our previous analysis of Figure 37, where we saw that men prefer the eye condition more than the head condition because their performance was worse for head. Women's results for this are not significant as they performed much better using head gaze as activation method. For Efficiency the only significant difference is for gaze type [$F(1, 30) = 25.17$, $p = 0.0001$]. No effect for gender was found. Finally, based on the above analysis, we can say that although both men and women on average recognize that the interaction was more efficient and pleasant in the eyes condition, women (and women only) do no rate the robot as less smart in the head one.

In a further analysis it was checked if participants who completed their tasks quicker (lower completion time) had better subjective opinion of the robot and vice versa. For this we used regression analysis: Figure 38, Figure 39. Concatenating eyes and head options yielded a significant regression for efficiency, smartness and likability. This finding seems to indicate that the judgments of the robot in terms of its intelligence, likability and efficiency are tightly linked to its actual performance in the task.

**Table 8. Regression analysis for completion time vs. subjective measures.**

| | likability | smartness | efficiency |
|---|---|---|---|
| *R square* | 0.0648 | 0.1690 | 0.2655 |
| *p* | 0.0422 | 0.00074 | 0.00001 |
| *F* | 4.301 | 12.608 | 22.401 |

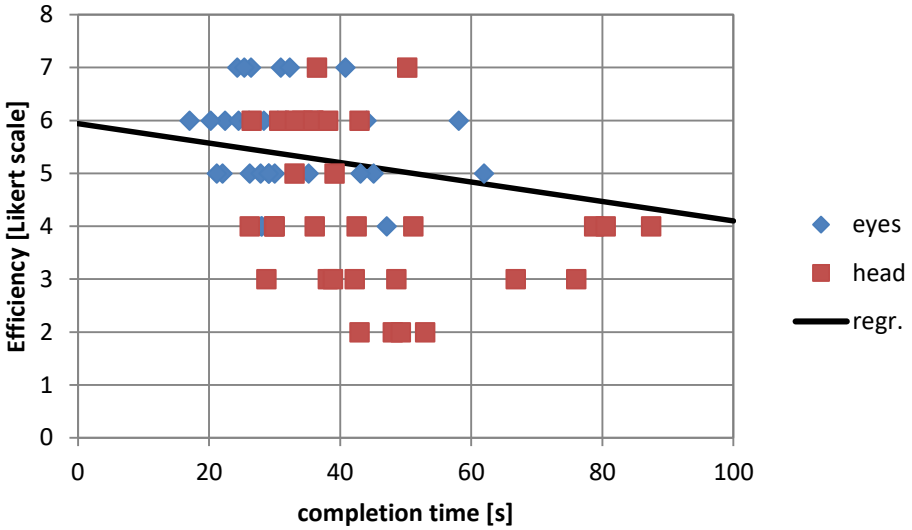


**Figure 38. Regression between efficiency and completion time.**

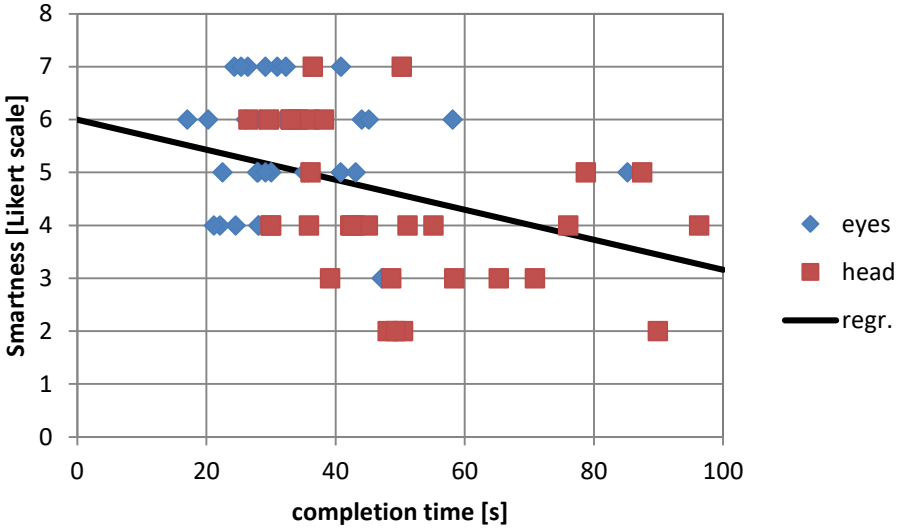


**Figure 39. Regression between smartness and completion time.**

## 6.9 Personality analysis

Since in human-human interaction it has been shown that personality traits might influence interactive behaviors and specifically gaze, [82] we wanted to assess whether personality could also have had an impact on the efficiency of our system, which relies on natural gazing behavior to make the robot understand and respond to the subjects' need. Thus we administered a personality questionnaire[6] to all 32 subjects. It is based on 50 questions and 4 categories: openness to new experience, neuroticism, extroversion and agreeability. Based on the 50 questions each subject gets a score estimate for each category named above. We were mostly interested in analyzing extroversion and openness to see how they affect people's interaction with the robot, i.e. we wanted to see if very introvert subjects would have difficulties communicating with an artificial agent.

First we wanted to evaluate whether personality traits (openness, extroversion, neuroticism) had any effect on actual participants' head and gaze behavior. To this aim we focused on subjects whose personality scores were more than one standard deviation away from the mean of the whole population. These subjects formed two groups (+1SD, -1SD). In the group of extroverts (+1SD) we had 6 people (1 woman, 5 men), while there were 5 introverts (3 women, 2 men). There were 5 people in the +1SD group for openness (1 woman, 4 men) and 5 in the -1SD group (4 women, 1 man). There were also 5 positive neurotics (3 women, 2 men) and 6 non-neurotics (2 women, 4 men).

Comparing the behavior metrics considered before, combining eyes and head conditions we found using two-sample t-tests that:

- More extrovert people moved their eyes left and right more compared to less extroverts [$t(20) = 2.309$, $p < 0.05$]
- More open subjects moved their eyes left/right more [$t(12) = 3.09$, $p < 0.01$], engaged in mutual gaze more ($t(14) = -2.27$, $p < 0.05$) and moved their heads up/down [$t(17) = 2.76$, $p < 0.05$] and left/right [$t(18) = 2.98$, $p < 0.01$] more than less open people.
- More neurotic subjects moved their heads left/right more [$t(20) = -2.27$, $p < 0.05$] compared to less neurotic people.

Next we analyzed how these two groups (-1SD and +1SD) performed at completing the given task to see if any of the personality traits would give an advantage or disadvantage for interaction with the robot. We have not found any statistically significant results (using a two-sampled t-

[6] The 50-item IPIP representation of Costa and McCrae's [96] five NEO domains

test) when comparing the two groups neither for extroversion [t(19) = -0.378, p = 0.35], nor openness [t(10) = -1.109, p = 0.14], nor neuroticism [t(19) = -0.062, p = 0.47], see Figure 40 below.

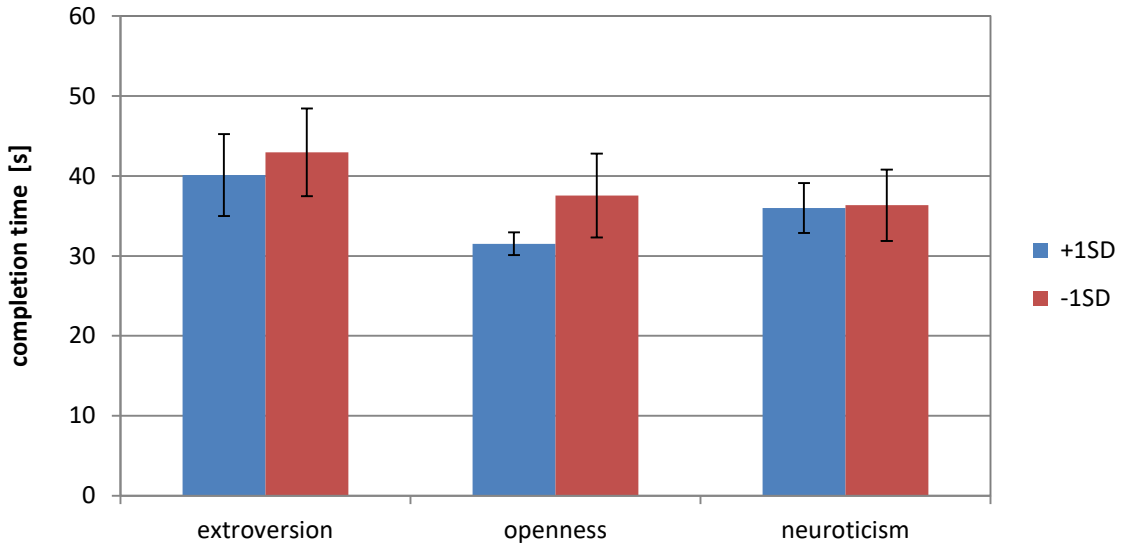


**Figure 40. Comparing -1SD and +1SD groups does not yield any statistical significance when looking at completion time.**

This negative result suggests that the system we propose can work well with individuals characterized by different personality traits. In particular, even if personality had an effect on the way people moved their eyes and head, this difference did not interfere with the intuitive gaze-based interaction with the robot. Also participants who scored low on extraversion and openness managed to complete the task not differently from subjects characterized by a more extrovert or open personality. This suggests that the gaze behavior on which our collaborative interaction was based is probably strongly rooted in human visuo-motor coordination and is preserved also in more neurotic and introvert people.

## 6.10 Discussion

The results of our collaborative tower building experiment show that enabling a robot to read eye gaze rather than approximating it with head gaze can bring significant advantages. When iCub monitored eye gaze, all but one trial were completed successfully, on average in less than 40s per trial. When relying on head gaze only, the interaction was slower and less effective: in seven cases the subjects were not even able to complete their task (4.4% failure of all interactions) with the head gaze option.

Moreover, from the pie charts in it can be noticed that the eye gaze option recognizes gaze behavior much better even at the first attempt, while for head gaze, there are much more cases of

repeated failures. This means that subjects will try over and over again the same thing and will grow quite frustrated by the robot's poor performance.

We believe that this is not due to a basic malfunctioning of the head tracking system or due to a wrong choice of its parameters (motion activation thresholds). Indeed, it must be noted that some of the subjects had no problem in completing the tasks when the robot reacted to their head pose and were almost as effective with this approach as with eyes. Moreover, even for the subjects who failed in some trials, the interaction was successful in the other repetitions. This finding convinces us that we selected the robot motion activation thresholds well, but head movement alone is less informative than eye movement. In particular, head motion was sometimes negligible (or non-existent), not providing the right trigger for robot action.

Looking at performance data, we found that women were more capable of activating the robot's offering motion even when the iCub reacted only to head movements. This seemed to be due to the fact that men move their heads less in interaction environments like ours. Analyzing personality data we have found that higher openness and extraversion result in more head movement and more time spent in eye contact with the robot. None the less, these behavioral differences between different personalities do not rise to the level to impede effective communication between humans and robots. In other words less open and more introvert subjects also found it easy to communicate with iCub using only their gaze.

An interesting finding is that while head gaze estimation seems good enough for turn change detection, it is not as good for object selection (left or right). We speculate that this result derives from the way humans naturally direct their attention. The change of turn implied a change of the selected interaction partner. In this process we usually make sure, even unconsciously, that the partners acknowledge to be targeted, to maximize their responsivity. Therefore, participants on average fully oriented themselves toward the current partner, with both head and eyes. Once the attention of the helper has been gained, it is assumed that they/it will easily understand what we need following our indications. Hence, head movements toward objects become more subtle or even disappear, not providing any more information.

Of course if subjects were instructed to guide the robot with their eyes or with their head movements, the performances would have been much better. However, our results show that the use of eye gaze detection allows for an efficient and smooth interaction even with fully naïve subjects, who behave with the robot as if it was a natural interaction partner. Using head gaze instead of eye gaze with unknowing subjects might lead instead to more unpredictable results.

In this study we have focused on whether the system detects the specific gaze gesture signal or not but we did not delve into whether the failure was due to a specific individual component of the sequence. Still the fact that at least for a subset of subjects (women) made the system work equally well in the two conditions suggests that there was not a critical general flaw in the head mode, which impeded its working at all.

Also the subjective results are quite informative of the fact that participants thought more positively of the better functioning option. The eye gaze option was evaluated to be much more efficient, likable and an indication of a smarter robot. It was interesting to see that in the eye gaze condition all but one tasks were completed successfully by all participants even though the majority (80%) of them did not realize that the robot responded to their gaze only.

Even though the eye gaze option performed well, there is still much room for improvement. One way to do this is to account for head roll in the algorithm. Another way would be to use some kind of calibration approach to increase the vertical accuracy of gaze estimation. To avoid a full gaze calibration, which would require a tedious ad hoc preparation to the interaction, a soft calibration is considered as a valid alternative. For example the robot could adjust its human eye model parameters at certain times when it knows from context that the subject is looking at its eyes. This is why we introduced the initial speaking phase of the robot. We assumed that while the robot speaks, the subjects would be looking at its eyes. At first glance at the data we can assume that this is mostly correct, but further in depth analysis is needed.

## 6.11 Study conclusions

A current challenge in robotics is to enable robotic companions in real life circumstances to communicate with their human counterparts more naturally. Our study shows how the ability to read human eye gaze represents a fundamental element to achieve this goal. In particular, we demonstrated that a humanoid robot enabled with eye gaze tracking abilities can successfully perform a collaborative building task with human partners completely naïve towards the reaction modality of the robot.

This work quantified also the impact of head-based and eye-based gaze tracking on the interaction. While for turn taking, monitoring the head motion provided similar results as monitoring the eyes, in an object selection task eyes provided an increase in efficiency of about 20 percentage points. As a whole, the eye gaze based interaction was on average 37% faster and was also perceived by participants as qualitatively more efficient. Therefore, it is not always a good idea to approximate eye gaze with head pose in human-robot interaction experiments.

We agree with some of the other authors (e.g. see Section 6.2) that if eye gaze is not available it might be sufficient to use head pose as the first proxy. For instance, head gaze detection might work if the head rotation angles are exaggerated, as for objects far apart or when subjects are instructed to use head motion. However, eye gaze tracking allows for a much finer scale of detection and does not require the human partners to change their natural interactive behavior.

We want to underline that with this research we don't imply that eye gaze should be the only cue a robot should use in interaction scenarios like the one we presented. Rather, we want to demonstrate that even gaze alone carries enough information to allow for task completion. The integration of the information derived from our eye gaze tracking system with the processing of other signals such as pointing and speech could lead to a very robust interactive robot.

To conclude, our eye tracking algorithm for robotics, based on standard cameras, could significantly improve the naturalness and efficiency of future robot companions, by eliminating the need of head-based approximations of gaze direction, which in certain contexts could lead to a much less efficient interaction.

# 7. Study V – A gaze-based social game for humans and robots

## 7.1 Introduction

The general problem we are addressing in this study is that human-humanoid interaction is not as natural as human-human interaction. We hypothesize that naturalness could be improved by exploiting gaze reading by the robot. Our goal was to improve our gaze tracking algorithms and to design a gaze tracking scenario in which we could employ this eye tracking solution to evaluate whether it could also be helpful in multiparty scenarios. The requirements for this scenario were: a) to include multiple humans and multiple robots and b) to be engaging enough for humans so they would not become bored quickly. These criteria would ensure a wealth of diverse gazing behavior. We decided that a social game would be the ideal scenario which would fulfill these requirements. Therefore we chose the so-called "Wink Murder" party game as the basis for our experiment. This game can be played by multiple humans and multiple robots, thus it can produce complex gazing behavior. The game also employs gaze detection as well as gaze actions (selection with gaze, mutual gaze) which is an added benefit. For the most part the game does not require verbal communication, thus it is easy to port it to different cultural groups. To the best of our knowledge, such a complex game with multiple humans and multiple humanoid robots was not yet implemented anywhere, thus using this setup could generate novel results, as we also show in our pilot experiment. Our main goals were to create a natural gazing scenario to validate our gaze tracking system and to explore human gazing behavior in interaction with androids.

## 7.2 Study background

As mentioned before gaze recognition is particularly interesting for humanoid robots, as it is a natural human ability that might be expected from human-like robots. There are two aspects of gaze behavior considering robots: gaze generation and gaze understanding. The first approach deals with how robots should display their own gazing. A good overview of this field can be found in [88]. Our work focuses both on humans reading the robot's gaze but more importantly on how robots could understand people's gaze in order to improve their natural behavior.

For a general background please refer to Section 2.

Considering engaging social interaction scenarios with humanoid robots and multiple humans the research field is not very populated. Vazquez et al. used the "chest of drawers" robot, Chester,

to play the Mafia game, but only in a Wizard of Oz fashion [89]. Leite et al. used the Keepon mini-robot to explore individual versus group HRI [90]. They found that multiple robots could be beneficial for social interaction with children and that behaviors and social signals can be significantly different when interacting in groups rather than individually with a robot. Crick and Scassellati used a small robot-controlled toy truck to facilitate playing social games between human participants [91].

More specifically looking only at using humanoid robots, Sheiki employed the miniature humanoid Nao as an art exhibit guide who interacted with two humans at the same time [81]. The robot was able to manage a conversation by knowing the head pose of the humans. Mutlu researched humanoid robots (either Asimo or Robovie) communicating verbally and using gaze signals with two human participants in a storytelling and a tourist agent scenario [88]. In one of his scenarios the humans liked the robot more when Robovie paid more attention to them.

Looking at scenarios where robots played a social game with a human participant, Gori et al. implemented a gesture recognition game which was played between an iCub humanoid robot and one human participant [92]. Both of them needed to memorize sequences of hand gestures and repeat them. Bentivegna et al. taught a humanoid robot to play air hockey against a human, but they didn't report on a naïve human subject study using this setup [93]. Ito and Tani used a small Sony humanoid playing an imitation game to explore joint attention and turn taking [94].

To the best of our knowledge, there has not yet been a study which would employ more than one human subjects and more than one humanoid robots in a social game scenario, previous to our work.

## 7.3 Methodology

This section will describe the android platform, gaze tracking on the robots, the rules of the game as well as the details of the experimental setup [6].

### 7.3.1 The android robots

In our approach we used the Actroid-F humanoid robot platform. These robots are quite suitable for our research purpose because of their very human-like appearance and behavior, see Figure 41.

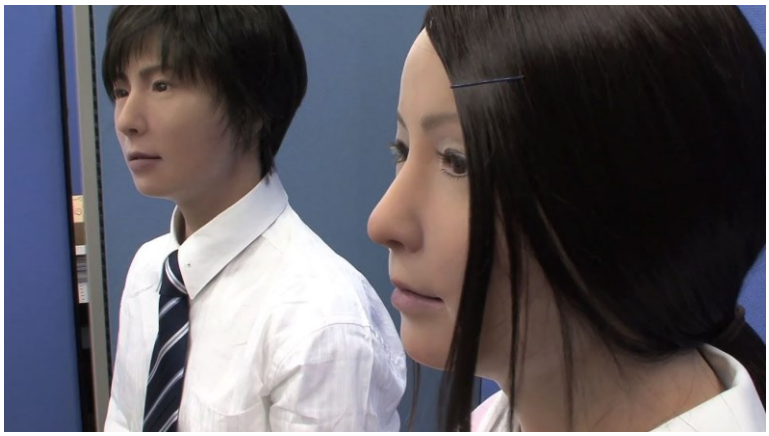

**Figure 41. The two Actroid-F humanoid robots.**

These robots' skin is made of silicon rubber. The faces are made as a verbatim copy of an unnamed person's face. The male and female versions both have the same facial features. They are made different by applying facial hair, wigs and cosmetics.

For more details on the androids please refer to Section 1.4.2.

### 7.3.2 Cameras

The robots have built-in cameras in both of their eyes. These are fixed focus CMOS devices (NCM13-J) with a horizontal field of view of about 45 degrees. They can provide images of VGA resolution (640x480 pixels) at 30 frames per second or at SXGA resolution (1280x1024) at 10 frames per second. The cameras fit right in the pupil opening of the robots' eyes.

For more details please refer to Section 1.4.2.

### 7.3.3 Gaze tracking

Gaze tracking of the human subjects was done using the approach we first described in Section 5. We introduced a couple of modifications to the original algorithm. Namely, for tracking the corners of the eyes instead the algorithm implemented in the dlib library [64] and described in [65], we opted for CLM [74]. This was done to speed up the execution as dlib is computationally more expensive and we needed to run the whole algorithm from a single computer. Since we already use CLM for head pose tracking we just used its eye corner detections too, even though they appear to be less precise than dlib's. Differently than in [4] but similarly as in [5] we performed the CLM algorithm only on a cutout of the whole image where the face was found in the previous frame. As we were tracking two faces this time, if we lost one of them, we reverted to search for the face on either the left or right half of the image. This windowing also sped up the approach but performing it twice (for two subjects) slowed it down compared to our previous solution [5]. In the end the whole algorithm was performing at around 10 frames per second,

which proved to be a usable speed.

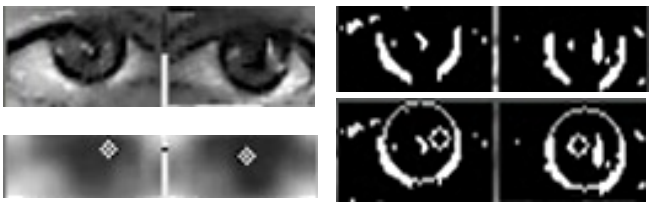

**Figure 42. Upper left: original image of eyes; lower left: blurred image; upper right: directional Sobel edge detection; lower right: circular Hough transformation.**

As an additional improvement of our gaze tracker, we improved our iris center detection approach compared to our previous studies, see Figure 42. We started off with a greyscale image of the eye area. Then we performed blurring of the image using a uniform filter with a width of the estimated iris radius. The iris radius was estimated by assuming an inter-pupillary distance of 65mm and an iris diameter of 12mm [75]. We selected the darkest point of the blurred image as the initial guess of the iris center. Then we performed directional Sobel edge detection to find the outline of the iris, knowing that the transition from the iris to the sclera needs to be a dark to light change. Once the edges were found we applied an own version of the circular Hough transformation with the previously estimated iris radius size to refine the center point recognition.

### 7.3.4 Rules of the game

The Wink Murder party game has several variations of gameplay. We opted for one which was the easiest for implementation and later data analysis. The practical minimum for playing this game is 4 participants. With 3 participants the game ends very quickly, so it does not provide much data to be analyzed, nor is it fun for the players. At the start of the game the players are dealt one card each. Only one of the cards contains an 'X' which designates the "murderer" (also called "villain"). The cards should not be revealed to anyone else other than the player who drew it. The other three players who get blank cards are the "innocents" who are also acting as "detectives" at the same time. In the original game the villain "murders" the innocents by winking at them, but since the Actroid-F is not able to wink (both eyelids are controlled via one actuator), we needed to find a different action. Blinking was considered but rejected, because human players cannot be expected to withhold their natural blinks for the purpose of the game. Thus we opted for a quick raising and lowering of both eyebrows as a "murder action". This can be performed by both the human and android participants easily without interfering with normal behavior.

The villain can "murder" an innocent only when there is mutual gaze between them, of course. Once an innocent is "murdered" he needs to continue playing the game for a few more seconds,

so that other players can't figure out the identity of the villain because of a sudden reaction. After this time the player needs to pitch his head downwards, signaling that he is out of the game. An innocent can win the game if she witnesses "an act of murder" in which case she is obliged to call out the villain. If she makes a wrong accusation, the game is ended and the accuser loses one point. The game practically ends when there is only the villain and one innocent left, as in this case the latter cannot perpetuate the game by not gazing at the villain.

## 7.4 Experimental design

In our implementation the game is played by two androids (R1 and R2) and two human participants (H1 and H2). One experimental session consists of 12 games played in a pseudo-random order, making sure that each player is the villain 3 times. The physical setup of the experiment can be seen in Figure 43. A photo of the setup can be seen in Figure 44. A video of the setup and experiments can be seen in the accompanying video[7].

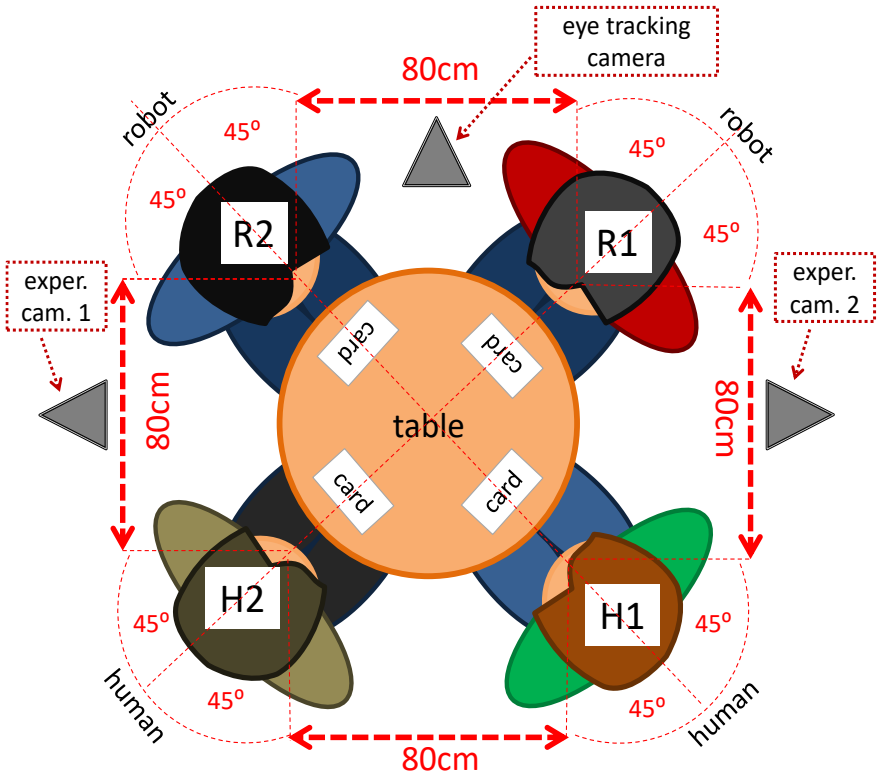


**Figure 43. Game setup.**

The physical location of the androids and humans stays the same throughout the experiment, thus R1 is always the female android, while R2 is the male and the humans don't switch places. As Figure 43 depicts, the four participants are seated at a round table 90 degrees apart. They are all

[7] https://youtu.be/CCKe61Z7Y_g

facing the center of the table. The approximate distance between the faces of the neighboring participants is around 80cm. A folded piece of paper is located in front of each participant at the start of each game. These are used to designate the villain. They are pseudo-shuffled and replaced after each game by the experimenter. The eye tracking camera is mounted on a tripod and placed between the two androids facing the two humans. Its position was chosen so that both humans appear in its field of view, as close as possible to the players. Since the C920 webcam is wide field of view, it can be placed close to the setup, thus providing better angles for eye tracking of the humans, see Figure 44. The TV screen seen in the same figure is only there for debugging purposes and is turned away for the actual experiments. The experimenter sits behind this screen, thus not interfering with the flow of the game. The experimenter sees the video feed from the webcam on the screen, so he can control the start and end of each game.

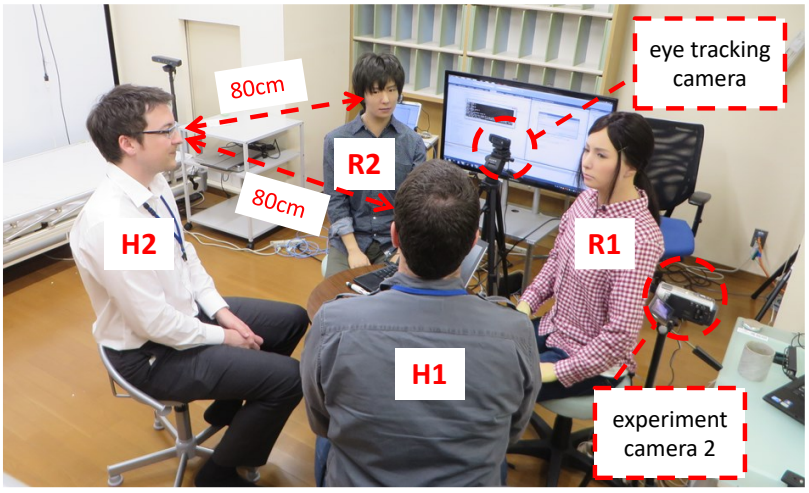


**Figure 44. Photo of the experimental setup.**

The whole experiment was recorded with two separate camcorders for analysis purposes ("experiment camera 1" and 2 in Figure 43).

### 7.4.1 Android head and eye movements

During each game the androids were randomly moving their gaze with independent timing. The fixations were uniformly distributed between 2 and 4 seconds. The only three focuses of the robots' attention were at the face of the three other participants. They changed their gaze either by rotating their eyes (-45, 0, 45 degrees) or in combination with head rotation. In this case the robots turned their heads to -22.5 or 22.5 degrees.

### 7.4.2 Android strategy as villain

When one of the androids was the villain, we provided it with some potential advantages over the human players. First, as an outside eye tracking camera was used, the robot knew the gaze direction of both humans at the same time, regardless of where it was looking. This was done

due to the fact that humans can also detect eyebrow movements even using their peripheral vision, thus we wanted to be at least as good as humans. Second, the non-villain robot did not call out the villain robot even if it saw it "murdering" someone. This was done due to the fact that our camera was not wide fielded enough to track all three other subjects, so we made the not-trackable subject (the other robot) passive. The villain robots were also not programmed to eliminate the non-villain robot.

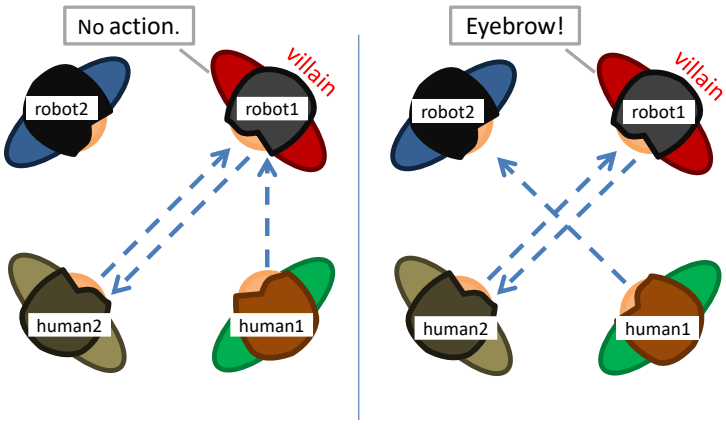


**Figure 45. Android strategy as villain.**

The main strategy of the villain android started with classifying the humans' gaze into three categories: looking at any of the other three participants. Once this was known based on the gaze tracking algorithm, the robot acted as follows: if the robot was looking at a human who was looking back at it, it would exert the eyebrow movement if the other human was looking somewhere else, see Figure 45. If both humans were looking at the villain-robot, then it did not perform the action, in order not to reveal itself. The androids were switching their gaze randomly and performed the "kill" action only when the above conditions were met. This was done so that the games would not end too quickly. All these robot actions were performed autonomously, i.e. without any Wizard of Oz influence by the experimenter.

### 7.4.3 Android strategy as innocent/detective

When one of the humans ended up being the villain, again, the robots were programmed to know that it was not the other robot, as we could not track three subjects. On the other hand, even though they were able to track both humans, each robot was programmed to detect eyebrow movements only on the human whom they were looking at. E.g. if one of the humans was trying to "kill" the other one and the eye tracking camera detected this, the action would go undetected if the robots were looking at each other at this time. The eyebrow movement itself was visually detected using the CLM algorithm. We compared any potential eyebrow movement with the movement of the mouth, as reference. If the mouth did not move much but the eyebrows did, it

was declared as a “kill action”. Unfortunately this was ambiguous with head pitch movements, so we needed to find a threshold which would minimize both false positives and false negatives. We did so by testing our system on a couple of pre-pilot subjects. Once all conditions were met for “kill” detection, the robot called out “Stop! You are the murderer.” while looking at the person who was detected to be the culprit.

### 7.4.4 Software component setup

The system’s software components and their connections are shown in Figure 46. The webcam image is fed to the CLM face feature and head pose detector. The algorithm expects to see two faces. Once it recognizes them, tracking is continued on small areas where the faces are detected in the previous frame. Gaze detection relies on the head pose and eye corner detection of CLM. Once the gaze of both humans is detected it is fed to the decision making algorithm, which acts according to the description under Sections 7.4.2 and 7.4.3 above. The robots are randomly looking at any of the three other participants as mentioned before, which is also an input to decision making. Robot actions (eyebrow movement if villain and calling out a villain when innocent) are generated as the output of the decision making block.

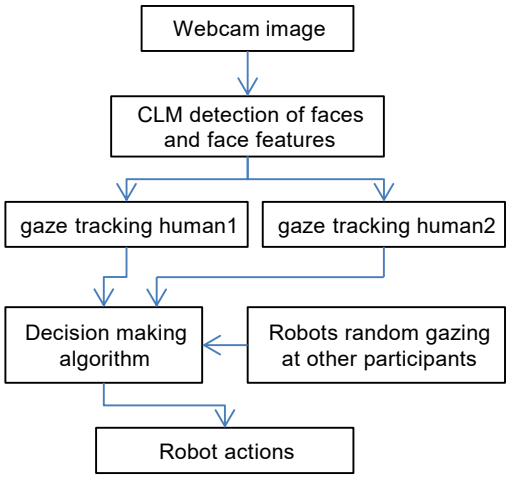


**Figure 46. Software components**

## 7.5 Experimental results

The pilot experiment was completed by people from the Information Technology and Human Factors department at AIST, Tsukuba: 12 subjects in 6 pairs. Four of them were women while 8 were men. Five pairs of subjects were Caucasians while one pair was Japanese.

Overall results of the pilot experiments can be found in Figure 47 while details are in Table 9.

The total results show that robots won a grand total of 31 games, while humans won 27 games. On average, robots won 5.2 while humans won 4.5 games. Looking at totals for each pair, it can be seen that three experiments were won by robots, one by humans and two experiments were tied. In Table 9 columns represent pairs of subjects, while rows show outcomes of games. The numerator in the results is the actual number of games won, while the denominator represents how many occasions the agent had to win the game. In the last row, humans had twice as many opportunities to win by detection, because the robots could detect only humans (by design), while humans could detect both robots and the other human as the "murderer".

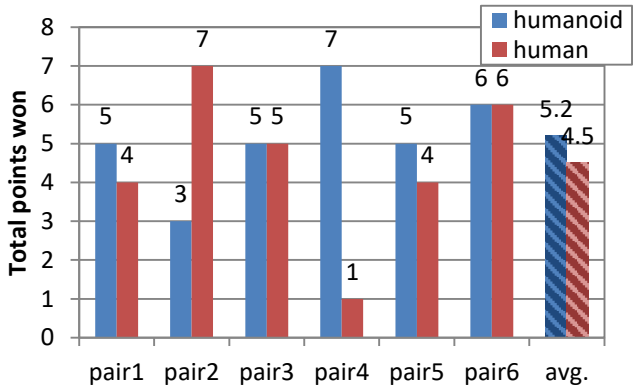


**Figure 47. Total points won in experiments by humans and humanoids.**

According to post-experimental discussion with participants, it was revealed that subjects felt quite interested in the experiment and thought it was fun. One pair reported that after the experiment they thought the robots were "much smarter" compared to what they thought before seeing the robots perform. Only Pair2 reported noticing that the robots "are working together" while others did not notice this even after being asked about it.

It is interesting to notice also the fact that only the pair which involved Japanese participants (Pair4) did not call out anyone (either human or robot) to be the villain. All other pairs did so at least 3 times.

**Table 9. Pilot experiment results: [number_of_wins]/[number_of_opportunities_to_win]**

| | Pair1 | Pair2 | Pair3 | Pair4 | Pair5 | Pair6 |
|---|---|---|---|---|---|---|
| *Robot won by "murder"* | 4/6 | 1/6 | 4/6 | 6/6 | 3/6 | 3/6 |
| *Human won by "murder"* | 0/6 | 1/6 | 1/6 | 1/6 | 0/6 | 0/6 |
| *Robot won by detection* | 1/6 | 2/6 | 1/6 | 1/6 | 2/6 | 3/6 |
| *Human won by detection* | 3/12 | 7/12 | 4/12 | 0/12 | 4/12 | 6/12 |

## 7.6 Discussion

The results of our pilot experiment show that the proposed gaze tracking system is appropriate to enable a natural interaction in the context of a social game as the Wink Murder. Indeed, the robots managed to win as villains the majority of the times (about 58%), suggesting that they could effectively monitor their human partners' gaze. One reason why humans were not able to win more games as villains can be the fact that the robots were not proficient at detecting the "kill actions" of humans (the raise of the eyebrows), see Section 7.4.3. This prevented humans from directly winning the games (row 2 of Table 9). On the other side, humans' results are boosted because in some situations humans called out the robots when the robots were trying to "kill" them. Humans thought that the robots were trying to eliminate the other player, i.e. they didn't detect mutual gaze with the robot and that's why they called out the villain (this was the case especially with Pair2 and Pair5). This might have been caused by the fact that in some situations the head movements of the robots were performed somewhat slower than the eye and eyebrow movements. The slower head movement was mostly unpredictable due to the nature of the pneumatic actuation system. In future implementations this technical issue needs to be addressed either by making sure the head turns are executed before the eyebrow action, or by eliminating head movements in favor of only eye movements altogether. The first solution might be problematic because there is no feedback mechanism in the robot which would inform us that the desired position is achieved. The latter solution might not be ideal because performing only eye movements with a still head looks quite unnatural.

As mentioned in the previous sections, the game was set up so that robots have the upper hand in two ways:

- The robots had foveal (sharp) vision of both human subjects at all times. Humans can only have foveal vision on one subject, while the other subject is seen peripherally (blurred).
- The two robots collaborated passively with each other by not calling out each other.

Both these benefits for the robots were caused mostly by technical limitations of our system. Namely, the fact that we had to use an outside camera for gaze tracking, because the in-eye cameras of the robots were not performing well enough for our purposes. In future improvements of our system we plan to fix these deficiencies to create an even playing field for both humans and robots. This might be difficult due to the fact that the space for installation of a different camera in the robots' eyeball is quite limited. On the other hand thanks to the recent advancements in miniature camera technology for mobile phone applications, it might be

possible to find replacements for the original sensors and optics.

Notwithstanding the current technical limitations, this framework allowed for an engaging interaction which might have evoked different behaviors and strategies in the participants. For example some people worked on eliminating the other human first and then focused on the robots. Other villains on the other hand tried to reassure their human co-player that they were also innocent by not performing the "kill action" right away. For the future, this type of multi-robot multi-human interactive scenario could be exploited to investigate natural coordination and collaboration mechanisms in naturalistic contexts. In particular it will be possible to extend the current understanding of phenomena supporting joint action, as for instance automatic imitation [95], going beyond the simplified, dyadic experimental settings usually adopted.

## 7.7 Study conclusions

Since eye gaze communication is very important in human-human interaction we think humanoid robots should also possess the ability to read others' gaze and communicate their own, to make themselves more natural, i.e. more human-like. In order to study this ability of gaze reading in human-humanoid interaction, "clever" scenarios need to be devised which would enable natural and diverse gaze behavior of the human participants. It is also needed to keep participants engaged and motivated in completing the experiment task. We think that the Wink Murder social game provides such a platform which could be even further refined. Our main contributions are the improvements of the gaze tracking algorithm and experiment design that allows this game to be played among humans and humanoid robots.

# 8. Conclusions

In the above chapters we argued for making human-robot interaction more natural than it is at the moment. For this we think implicit interaction signals need to be included in the robot's sensory repertoire. As argued, one of these very important signals is gaze, which humans use pervasively in their everyday communication. In the following we will give concluding arguments based on the contents of each presented study in this dissertation.

## 8.1 An affordable active gaze tracking head's advantages

We argue that having an affordable, simple gaze tracking robotic head made of off-the-shelf components could advance gaze research in human-robot interaction. For designing such a system a small number of readily available elements would be needed: webcams, servo motors, a PC computer and algorithms. Of course, the crucial element of such a system would be the webcam. As the availability of high definition webcams with hardware compression is becoming ubiquitous, there are less reasons why not to go for this option. Especially knowing that freely available software packages as OpenCV allow for the easy correction of optical deficiencies of these affordable cameras. In our work we presented a prototype system which fulfills all the requirements for easy dissemination of such a robot head system.

## 8.2 A gaze-contingent tutor robot provides a better teaching experience

In this study we enabled the iCub humanoid robot to detect eye contact (mutual gaze) with "its students", which allows for a better classroom experience, where the robot continues its teaching only when the student is ready to listen. This scenario can of course be generalized, where any kind of turn-taking interaction could be augmented by the detection of mutual gaze of the human participant. For instance a factory robot could only continue supplying parts when the human operator is ready to receive them, or in case of a home helper robot, it could add the chopped onions to the stew when the human supervisor gives "the sign". In our study we have pointed out both the benefits and shortfalls of such a contingent system: such a robot would allow for the adaptation to the work pace of the human operator, but at the same time this adaptation could allow for slacking with an unconscientious agent lacking training.

## 8.3 Our new gaze tracking algorithm could be the simple solution for robots

In Study III we report on a visual light, monocular, calibration-free gaze tracking algorithm,

which is designed with (humanoid) robots in mind. This algorithm works with almost any kind of camera system (e.g. webcams), which eliminates the need for high fidelity and high priced camera systems. Any robot could be retro-fitted with gaze tracking capabilities. The system has been designed with robots in mind: it is highly mobile, works in subpar lighting conditions and provides a robust estimation, even though it lacks high precision. As many times the robot's task is to categorize the human's gaze in wide bins, rather than acquire high precision gaze direction, this performance can often times be quite satisfactory for the task at hand. We designed, trained and tested our gaze tracker in this study. We achieved around 5 degrees average absolute error in the horizontal plane, and around 9 degrees error in the vertical plane. The vertical precision reflects the nature of eye trackers in general (due to looking down much of the sclera is covered by the eyelids, which makes tracking difficult), but is not a fully satisfactory result. This is why we suggest using soft calibration methods to offset the errors for individual subjects. We are also helped by advancements in facial feature detection in real time. This allowed us to train a generic geometric face model which will work fairly well for everyone, while ignoring person specific differences. This inevitably leads to some errors in estimating gaze, but eliminates the painful process of calibrating the gaze system before use. It allows us to use gaze tracking on our robots without the human participant even knowing about it. This approximates human-like capabilities, as we also do not need training to understand where our colleagues are looking at.

## 8.4 Arguing for eye gaze tracking instead of its first proxy head gaze

In this chapter we challenge the de facto accepted replacement of eye gaze with "head gaze" for human-robot interaction studies. In the past, due to limited camera spatial resolution many researchers have resorted to using head pose instead of eye gaze. Since high resolution, auto-focus, auto white balance cameras have lately become a more pervasive technology we argue that this approximation does not need to be made anymore. Of course, for gaze tracking to be applicable, the eyes need to be visible and discernable. In this study we explored if there is a significant amount of information that is lost if performing simple head pose tracking instead of proper gaze tracking. As it turned out in our collaborative tower building scenario, when our system relied on the eyes, the tasks were completed much quicker and with less errors than when tracking the head pose only and ignoring the direction of the eyes' gaze. Unexpectedly we found that women are performing great even when dealing with a head pose detection system. It turns out that this was in our case due to the increased amount of head movement of women compared to men. We also investigated if certain types of personalities are behaving aversely towards interacting with robots (e.g. introverts), but we did not find evidence for this. On the other hand

we did find different behavior of extreme extroverts as compared to extreme introverts: more head movement, longer time spent in eye contact with the robot, etc. This encourages us to continue designing gaze-based interaction systems, with a low chance of anyone not being able to interact with it due to their personality traits.

## 8.5 Competitive human-robot tasks provide a wealth of gaze behavior for engaging interactions

In our final study we turned to a competitive HRI task as opposed to the previous collaborative tower building task. In this study we also used a different platform, which is extremely human-looking (Actroid-F), while sacrificing the subtlety of body movements (iCub). We designed a gaze-based social game, in which two humanoids and two novice humans were competing in the so called “Wink Murder” game. This game allows for a very engaging and fun scenario, which is likely to keep the participants involved for a prolonged time. This time would allow detailed observation of gazing behavior and strategies of the humans in the game. Another benefit of this game is that as far as we know, this is the first interaction scenario in which multiple humanoids and multiple humans participate completing a competitive task. The robots were programmed to fulfill both the roles of the villain and an innocent, which provided for a more realistic game environment. In this study we have found that the two androids who passively collaborate during the game are able to beat the novice human players by a narrow margin. Better quality eye tracking and facial expression recognition mechanisms will be needed for the robots to be competitive even with experienced players.

## 8.6 Final conclusions and future work

As mentioned in the Introduction, the goal of this dissertation was to design and assess the usefulness of a gaze reading algorithm for HRI. As a conclusion to all of our studies we have found that a gaze tracking system built with HRI in mind will be very useful for providing robots more understanding of the intention and behavior of humans which will allow them to better attend to people’s needs.

Such systems (built specifically with HRI in mind) are not yet well disseminated and used in the field of HRI. We hope that by our design and experimental results the use of gaze reading will become more ubiquitous in human-robot interaction for the benefit of the whole field.

There is still a lot of space left for improving our gaze estimation algorithm. As mentioned previously we believe that to achieve a natural interaction it is important that it unfolds without

the need of ad hoc calibration procedures for eye tracking. This choice however limits the performance of the system. Future work will be dedicated to the development of a *soft calibration* mechanism, a system which could determine individual more exact parameters of the partner's face during the initial phases of the interaction without the participant's awareness. This could be achieved with the robot adjusting its human eye model parameters at certain times when it knows from context that the subject is looking at its eyes (e.g. during the greeting phase) or in a well-known direction (e.g., while showing an object to an engaged person).

Another important improvement to the proposed eye tracking system would consist of making it work with multiparty interactions. Currently we have demonstrated the possibility to exploit concurrent eye tracking of two people at the same time (see Section 7). However, we foresee that oftentimes robots will be faced with interactions involving groups of people, as when playing the role of a guide in a museum or of a teacher in a classroom. In these contexts, the possibility to monitor the eye gaze of multiple people could help maintain an accurate estimate of their engagement, enabling the robot to better cope with potential loss of interest. There are several aspects limiting this kind of group gaze tracking today. One is the lack of wide availability of very high resolution camera systems in robots. The other is the need for creating a distributed system of computers which would be able to perform multiple detections of face features and head pose, which is computationally very costly.

So far, our gaze tracking system is based on a single camera. Most humanoid robots are however endowed with two cameras, which could provide binocular vision. Another potential improvement of the system could be based on the exploitation of both camera streams. Using the second stream could help with refining the system's accuracy by averaging the left and right cameras' results. At the same time it would significantly increase processing time, for which an additional processing node (computer) would be needed. Distance estimation of observed people could also be done by adding the second camera stream by using stereoscopy, but this approach today is limited by the cameras' low resolution and the small parallax angle between the two eyes and the observed person.

Beyond the impact that the proposed system might have on HRI in general, it could provide particular benefits in the field of HRI for cognitive rehabilitation. Indeed, the sensitivity to the partner's gaze and to its potential anomalies could enable the robot to be used as a real-time diagnostic tool. For instance it could be exploited detect early behavioral problems associated with gaze processing as Autism Spectrum Disorders, by monitoring subjects' gaze during the interaction. This tool would simplify and shorten the time for gaze analysis, that although proved

promising [79], so far can be performed only with a lengthy a posteriori manual annotation of video recordings. Moreover, a functioning embedded eye tracking system would enable the robot to tailor its reactions to the individual's behavior during the interaction, with no need of continuous human intervention or Wizard of Oz scenarios.

We believe that the proposed eye tracker could have impact also beyond robotics. As the system works even with simple commercial webcams, it is possible to foresee its application in a variety of domains where the monitoring of eye gaze with no need of ad hoc specific hardware installation could be beneficial. For instance, it could be exploited in games for little children, as a low cost tool to make gaze-contingent interaction possible, e.g., providing the description of the object the infant is looking at, to facilitate the acquisition of language.

In summary in this dissertation we are shedding light on the possible role of gaze tracking in aiding and augmenting communication between humans and robots, by also providing new software and hardware tools for achieving this goal, which might have applications in various robotics domains and also beyond robotics. We hope this work could represent a contribution to the improvement of the HRI field, providing a more efficient and natural interaction between humans and robots, thanks to the understanding of an important implicit communication signal as gaze.

# 9. Publications of the candidate

[1] Sciutti A., Patanè L., Palinko O., Nori F. & Sandini G. 2014, 'Developmental changes in children understanding robotic actions: the case of lifting', *ICDL- Epirob 2014*, Genoa, Italy, October 13-16, 2014.

[2] Sciutti A., Palinko O., Patanè L., Rea F., Nori F., Noceti N., Odone F., Verri A. & G. Sandini, 'Bidirectional Human-robot action reading', *Human-Friendly Robotics Workshop (HFR'14),* Pontedera, Pisa, Italy, October, 23-24, 2014.

[3] Palinko O., Sciutti A., Patanè L., Rea F., Nori F. & Sandini G. 2014, 'Communicative Lifting Actions in Human-Humanoid Interaction', *IEEE/RAS International Conference of Humanoids Robotics (Humanoids'14)*, Madrid, Spain, November 18-20, 2014.

[4] Palinko O., Sciutti A., Rea F., Sandini G. 2014, 'Towards Better Eye Tracking in Human Robot Interaction Using an Affordable Active Vision System', *2nd International Conference of Human-Agent Interaction*, Tsukuba, Japan, October, 29-31, 2014.

[5] Palinko O., Sciutti A., Rea F., Sandini G. 2014, 'Weight-Aware Robot Motion Planning for Lift-to-Pass Action', *2nd International Conference of Human-Agent Interaction*, Tsukuba, Japan, October, 29-31, 2014

[6] Palinko O. & Sciutti A. 2014, 'Exploring the Estimation of Cognitive Load in Human Robot Interaction', *Workshop HRI: a bridge between Robotics and Neuroscience at 9th ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld, Germany, March 03-06, 2014.

[7] Palinko O., Rea F., Sandini G. & Sciutti A. 2015, 'Eye Gaze Tracking for a Humanoid Robot', *Proceedings of the International Conference on Humanoids Robots (Humanoids'15)*, Seoul, Korea, November 3-5, 2015.

[8] Palinko O., Rea F., Sandini G. & Sciutti A. 2015, 'Gaze Tracking for Human Robot Interaction', *International Workshop on Vision and Eye Tracking in Natural Environments and Solutions & Algorithms for Gaze Analysis (SAGA'15),* Bielefeld, Germany, September 29-30, 2015.

[9] Sciutti A., Schillingmann L., Palinko O., Nagai Y. & Sandini G. 2015, 'A Gaze-contingent Dictating Robot to Study Turn-taking', *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp.137-138, Portland, Oregon, USA, March 2-5, 2015.

[10] Palinko O., Sciutti A., Schillingmann L., Rea F., Nagai Y. & Sandini G. 2015, 'Gaze Contingency in Turn-Taking for Human Robot Interaction: Advantages and Drawbacks', *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'15)*, Kobe, Japan, August 31 - September 4, 2015.

[11] Palinko O., Sciutti A, Wakita Y., Matsumoto Y., & Sandini G., 2016. 'If Looks Could Kill: Humanoid Robots Play a Gaze-Based Social Game with Humans', *Proceedings of the IEEE/RAS International Conference of Humanoids Robotics (Humanoids'16)*, 2016.

[12] Palinko O., Rea F., Sandini G. & Sciutti A. 2016. 'A Robot Reading Human Gaze: Why Eye Tracking Is Better Than Head Tracking for Human-Robot Collaboration', *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16),* 2016.

[13] Palinko O., Rea F., Sandini G. & Sciutti A. 2016. *'Eye Tracking for Human Robot Interaction'*, *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA'16),* 2016.

# 10. References

[1] O. Palinko, A. Sciutti, F. Rea, and G. Sandini, “Towards better eye tracking in human robot interaction using an affordable active vision system,” in *Proceedings of the second international conference on Human-agent interaction - HAI ’14*, 2014, pp. 217–220.

[2] A. Sciutti, L. Schillingmann, O. Palinko, Y. Nagai and G. Sandini, “A Gaze-contingent Dictating Robot to Study Turn-taking,” in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[3] O. Palinko, A. Sciutti, F. Rea, Y. Nagai, and G. Sandini, “Gaze Contingency in Turn-Taking for Human Robot Interaction: Advantages and Drawbacks,” in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication*, 2015.

[4] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, “Eye Gaze Tracking for a Humanoid Robot,” in *Proceedings of the 2015 IEEE-RAS International Conference on Humanoid Robots*, 2015, pp. 318–324.

[5] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, “A Robot Reading Human Gaze: Why Eye Tracking Is Better Than Head Tracking for Human-Robot Collaboration,” in *Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[6] O. Palinko, A. Sciutti, Y. Wakita, Y. Matsumoto, and G. Sandini, “If Looks Could Kill: Humanoid Robots Play a Gaze-based Social Game with Humans,” in *Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots*, 2016.

[7] N. George and L. Conty, “Facing the gaze of others.,” *Neurophysiol. Clin.*, vol. 38, no. 3, pp. 197–207, Jun. 2008.

[8] S. Baron-Cohen, "Mindblindness: An essay on autism and theory of mind." *MIT Press,* 1997.

[9] A. Borji, D. Parks, and L. Itti, “Complementary effects of gaze direction and early saliency in guiding fixations during free viewing.,” *J. Vis.*, vol. 14, no. 13, p. 3-, Jan. 2014.

[10] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, “The iCub humanoid robot : an

open platform for research in embodied cognition.," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, 2008, pp. 50–56.

[11] U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini, "An experimental evaluation of a novel minimum-jerk Cartesian controller for humanoid robots," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2010, pp. 1668–1674.

[12] S. Ivaldi, M. Fumagalli, M. Randazzo, F. Nori, G. Metta, and G. Sandini, "Computing robot internal/external wrenches by means of inertial, tactile and F/T sensors: Theory and implementation on the iCub," in *Proceedings of the 2011 IEEE-RAS International Conference on Humanoid Robots*, 2011, pp. 521–528.

[13] R. Beira *et al.*, "Design of the robot-cub (iCub) head," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2006, vol. 2006, pp. 94–100.

[14] U. Pattacini, "Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub," *PhD dissertation, Istituto Italiano di Tecnologia*, 2011.

[15] M. Yoshikawa, Y. Matsumoto, M. Sumitani, and H. Ishiguro, "Development of an android robot for psychological support in medical and welfare fields," in *Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics, ROBIO*, 2011, pp. 2378–2383.

[16] G. Bradski, "The OpenCV Library," *Dr Dobbs J. Softw. Tools*, vol. 25, pp. 120–125, 2000.

[17] N. Mavridis, "A review of verbal and non-verbal human--robot interactive communication," *Rob. Auton. Syst.*, vol. 63, pp. 22–35, 2015.

[18] S. Lackey, D. Barber, L. Reinerman, N. I. Badler, and I. Hudson, "Defining Next-Generation Multi-Modal Communication in Human Robot Interaction," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 55, no. 1, pp. 461–464, 2011.

[19] J. Adams, P. Rani, and N. Sarkar, "Mixed initiative interaction and robotic systems," *AAAI Workshop on Supervisory Control of Learning and Adaptative Systems,* 2004

[20] E. L. Blickensderfer, R. Reynolds, E. Salas, and J. A. Cannon-bowers, "Shared Expectations and Implicit Coordination in Tennis Doubles Teams," *J. Appl. Sport Psychol.*, vol. 22, no. April 2013, pp. 486–499, 2010.

[21] J. Greenstein and M. Revesman, “Two Simulation Studies Investigating Means of Human-Computer Communication for Dynamic Task Allocation,” *IEEE Trans. Syst. Man. Cybern.*, vol. 16, no. 5, pp. 726–730, 1986.

[22] E. Pagello, A. D’Angelo, F. Montesello, F. Garelli, and C. Ferrari, “Cooperative behaviors in multi-robot systems through implicit communication,” *Rob. Auton. Syst.*, vol. 29, no. 1, pp. 65–77, 1999.

[23] A. Mehrabian, "Silent messages," vol. 8, *Wadsworth,* 1971.

[24] P. N. Wilson, “Training to Foster Implicit Communications,” *Mar. Corps Gaz.*, vol. 91, no. 4, pp. 29–32, 2007.

[25] M. J. Matarić, “Issues and approaches in the design of collective autonomous agents,” *Rob. Auton. Syst.*, vol. 16, no. 2–4, pp. 321–331, 1995.

[26] M. F. Martins and Y. Demiris, “Impact of Human Communication in a Multi-teacher, Multi-robot Learning by Demonstration System.,” *Proceedings of the Workshop on Agents Learning Interactively from Human Teachers* 2010.

[27] D. Vogel and R. Balakrishnan, “Interactive Public Ambient Displays: Transitioning from Implicit to Explicit, Public to Personal, Interaction with Multiple Users,” *UIST ’04 Proc. 17th Annu. ACM Symp. User interface Softw. Technol.*, vol. 6, no. 2, pp. 137–146, 2004.

[28] N. J. Emery, “The eyes have it: The neuroethology, function and evolution of social gaze,” *Neuroscience and Biobehavioral Reviews*, vol. 24, no. 6. pp. 581–604, 2000.

[29] A. Frischen, A. P. Bayliss, and S. P. Tipper, “Gaze cueing of attention: Visual attention, social cognition, and individual differences.,” *Psychol. Bull.*, vol. 133, no. 4, pp. 694–724, 2007.

[30] M. Von Cranach and J. H. Ellgring, “The perception of looking behaviour,” *Soc. Commun. Mov.*, 1973.

[31] Y. Nagai, M. Asada, and K. Hosoda, “Learning for joint attention helped by functional development,” *Adv. Robot.*, vol. 20, no. 10, pp. 1165–1181, 2006.

[32] T. Farroni, E. M. Mansfield, C. Lai, and M. H. Johnson, “Infants perceiving and acting on the eyes: Tests of an evolutionary hypothesis,” *J. Exp. Child Psychol.*, vol. 85, no. 3, pp. 199–212, 2003.

[33] J. S. Marzillier, “Gaze and mutual gaze,” *Behaviour Research and Therapy*, vol. 14. p. 486, 1976.

[34] T. Farroni, G. Csibra, F. Simion, and M. H. Johnson, “Eye contact detection in humans from birth,” *Proceedings of the National Academy of Sciences*, 99(14), 9602-9605, 2002.

[35] M. Argyle and J. Dean, “Eye-contact, distance and affiliation,” *Sociometry*, pp. 289–304, 1965.

[36] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, “Equilibrium theory revisited: Mutual gaze and personal space in virtual environments,” *Presence*, vol. 10, no. 6, pp. 583–598, 2001.

[37] M. Argyle, "Bodily communication." *Routledge*, 2013.

[38] S. Ivaldi, S. M. Anzalone, W. Rousseau, O. Sigaud, and M. Chetouani, “Robot initiative in a team learning task increases the rhythm of interaction but not the perceived engagement,” *Front. Neurorobot.*, vol. 8, 2014.

[39] D. N. Saito *et al.*, “‘Stay tuned’: inter-individual neural synchronization during mutual gaze and joint attention,” *Front. Integr. Neurosci.*, vol. 4, p. 127, 2010.

[40] F. Kaplan and V. V Hafner, “The Challenges of Joint Attention,” *Interact. Stud.*, vol. 7, pp. 135–169, 2006.

[41] H. Admoni, A. Dragan, S. Srinivasa, and B. Scassellati, “Deliberate Delays During Robot-to-Human Handovers Improve Compliance With Gaze Communication,” *Int. Conf. Human-Robot Interact.*, pp. 49–56, 2014.

[42] M. Argyle and M. Cook, “Gaze and mutual gaze”, *Cambridge University Press,* 1976.

[43] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, “Conversational gaze aversion for humanlike robots,” *Proc. 2014 ACM/IEEE Int. Conf. Human-robot Interact. - HRI ’14*, pp. 25–32, 2014.

[44] S. Andrist, B. Mutlu, and M. Gleicher, “Conversational gaze aversion for virtual agents,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8108 LNAI, pp. 249–262.

[45] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-Based Gaze Estimation in the Wild,” *Proceedings of the IEEE Conference on Computer Vision and Pattern*

*Recognition.* 2015.

[46] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive Driver Gaze Tracking with Active Appearance Models," in *Proc. World Congress on Intelligent Transportation Systems*, 2004, pp. 1–12.

[47] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interact. Stud.*, vol. 7, no. 2, pp. 135–169, 2006.

[48] L.-P. Morency, C. M. Christoudias, and T. Darrell, "Recognizing gaze aversion gestures in embodied conversational discourse," *Proc. 8th Int. Conf. Multimodal interfaces - ICMI '06*, p. 287, 2006.

[49] M. W. Doniec, G. Sun, and B. Scassellati, "Active Learning of Joint Attention," in *Proceedings of the 2006 IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 34–39.

[50] H. Kim, H. Jasso, G. Deák, and J. Triesch, "A robotic model of the development of gaze following," in *Proceedings of the 2008 IEEE 7th International Conference on Development and Learning, ICDL*, 2008, pp. 238–243.

[51] F. Broz and H. Lehmann, "Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation," *RO-MAN,* 2012.

[52] A. Sciutti, A. Bisio, F. Nori, G. Metta, L. Fadiga, and G. Sandini, "Anticipatory gaze in human-robot interactions," in *Gaze in HRI from modeling to communication workshop at the 7th ACM/IEEE international conference on human-robot interaction, Boston, Massachusetts, USA*, 2012.

[53] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000., pp. 499–504.

[54] J. Ido, Y. Matsumoto, T. Ogasawara, and R. Nisimura, "Humanoid with interaction ability using vision and speech information," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2006, pp. 1316–1321.

[55] R. J. K. Jacob and K. S. Karn, "Eye Tracking in Human-Computer Interaction and Usability Research. Ready to Deliver the Promises.," in *The Mind's Eye: Cognitive and*

*Applied Aspects of Eye Movement Research*, 2003, pp. 531–553.

[56] Y. Matsumoto, N. Sasao, T. Suenaga, and T. Ogasawara, "3D Model-based 6-DOF head tracking by a single camera for human-robot interaction," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009, pp. 3194–3199.

[57] R. Atienza and A. Zelinsky, "Active gaze tracking for human-robot interaction," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, ICMI 2002*, 2002, pp. 261–266.

[58] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Comput. Vis. Pattern Recognit.*, vol. 1, p. I--511--I--518, 2001.

[59] G. Sandini, P. Questa, D. Scheffer, B. Diericks, and A. Mannucci, "A retina-like CMOS sensor and its applications," in *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2000, vol. 2000–January, pp. 514–519.

[60] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing In Human-Robot Conversations : How Robots Might Shape Participant Roles Using Gaze Cues," *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.* ACM, 2009.

[61] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating Vision and Audition within a Cognitive Architecture to Track Conversations," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2008.

[62] K. S. Lohan *et al.*, "Tutor Spotter: Proposing a Feature Set and Evaluating It in a Robotic System," *Int. J. Soc. Robot.*, vol. 4, no. 2, pp. 131–146, Dec. 2011.

[63] D. G. Novick, B. Hansen, and K. Ward, "Coordinating turn-taking with gaze," *Proceeding Fourth Int. Conf. Spok. Lang. Process. ICSLP '96*, vol. 3, 1996.

[64] D. E. King, "Dlib-ml : A Machine Learning Toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[65] V. Kazemi and S. Josephine, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Computer Vision and Pattern Recognition (CVPR), 2014*, 2014.

[66] F. Timm and E. Barth, "Accurate eye centre localisation by means of gradients," in *International Conference on Computer Theory and Applications (VISAPP)*, 2011, pp.

125–130.

[67] B. A. Smith, S. K. Feiner, and S. K. Nayar, “Gaze Locking : Passive Eye Contact Detection for,” *UIST*, pp. 271–280, 2013.

[68] M. Schröder and J. Trouvain, “The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching,” in *International Journal of Speech Technology*, 2003, vol. 6, pp. 365–377.

[69] C. M. Brown, “Human-computer interface design guidelines,” *Intellect Books,* 1988.

[70] H. Brugman and A. Russel, “Annotating multi-media/multi-modal resources with ELAN,” *Int. Conf. Lang. Resour. Eval.*, pp. 2065–2068, 2004.

[71] M. Hanheide, M. Lohse, and A. Dierker, “SALEM - Statistical AnaLysis of Elan files in Matlab,” in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010, pp. 121–123.

[72] K. Johnson and E. Street, “Response to intervention and precision teaching: creating synergy in the classroom,” *Guilford Press,* 2011.

[73] P. Viola and M. Jones, “Robust real-time face detection,” *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[74] T. Baltrusaitis, P. Robinson, and L. P. Morency, “3D Constrained Local Model for rigid and non-rigid facial tracking,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2610–2617.

[75] C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, and J. T. McConville, “Anthropometric survey of US army personnel: methods and summary statistics 1988,” *Anthropology Research Project Inc,* 1989.

[76] Y. Kim and B. Mutlu, “How social distance shapes human-robot interaction,” *Int. J. Hum. Comput. Stud.*, vol. 72, pp. 783–795, 2014.

[77] S. Al Moubayed and G. Skantze, “Perception of gaze direction for situated interaction,” *Proc. 4th Work. Eye Gaze Intell. Hum. Mach. Interact. - Gaze-In ’12*, pp. 1–6, 2012.

[78] A. Poole and L. J. Ball, “Eye tracking in HCI and usability research,” *Encycl. Hum. Comput. Interact.*, vol. 1, pp. 211–219, 2006.

[79] S. M. Mavadati, H. Feng, A. Gutierrez, and M. H. Mahoor, “Comparing the gaze responses of children with autism and typically developed individuals in human-robot interaction,” in*, Proceedings of the 2014 14th IEEE-RAS International Conference on Humanoid Robots,* 2014, pp. 1128–1133.

[80] E. S. Kim *et al.*, “Social robots as embedded reinforcers of social behavior in children with autism,” *J. Autism Dev. Disord.*, vol. 43, pp. 1038–1049, 2013.

[81] S. Sheikhi and J.-M. Odobez, “Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions,” *Pattern Recognit. Lett.*, 2014.

[82] Y. Iizuka, “Extraversion, introversion, and visual interaction.,” *Percept. Mot. Skills*, vol. 74, no. 1, pp. 43–50, 1992.

[83] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, “Towards Engagement Models that Consider Individual Factors in HRI: On the Relation of Extroversion and Negative Attitude Towards Robots to Gaze and Speech During a Human--Robot Assembly Task,” *Int. J. Soc. Robot.*, vol. 9, no. 1, pp. 63–86, 2017.

[84] M. Beatty, J. McCroskey, and K. Valencic, "The biology of communication: A communibiological perspective." *Hampton Press,* 2001.

[85] B. Mutlu, J. Forlizzi, and J. Hodgins, “A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior,” *Proceedings of the 6th IEEE-RAS Int. Conf. Humanoid Robot.*, pp. 518–523, Dec. 2006.

[86] C.-M. Huang and B. Mutlu, “The Repertoire of Robot Behavior: Designing Social Behaviors to Support Human-Robot Joint Activity,” *J. Human-Robot Interact.*, vol. 2, no. 2, pp. 80–102, Jun. 2013.

[87] A. Gelman and J. Hill, "Data analysis using regression and multilevel/hierarchical models." *Cambridge university press*, 2006.

[88] B. Mutlu, “Designing gaze behavior for humanlike robots,” *Doctoral dissertation, Ford Motor Company*, 2009.

[89] M. Vazquez, A. Steinfeld, and S. E. Hudson, “Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation,” in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2015, vol. 2015–

Decem, pp. 3010–3017.

[90] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, “Comparing Models of Disengagement in Individual and Group Interactions,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 99–105.

[91] C. Crick and B. Scassellati, “Intention-based robot control in social games,” in *Proceedings of the Cognitive Society Annual Meeting*, 2009.

[92] I. Gori, S. R. Fanello, G. Metta, and F. Odone, “All gestures you can: A memory game against a humanoid robot,” in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2012, pp. 330–336.

[93] D. Bentivegna, A. Ude, C. Atkeson, and G. Cheng, “Humanoid robot learning and game playing using PC-based vision,” *Int. Conf. Intell. Robot. Syst.*, vol. 3, pp. 2449–2454, 2002.

[94] I. Masato and J. Tani, “Joint attention between a humanoid robot and users in imitation game,” *Proc. Thrid Int. Conf. Dev. Learn.*, 2004.

[95] A. Bisio *et al.*, “Motor contagion during human-human and human-robot interaction,” *PLoS One*, vol. 9, no. 8, 2014.

[96] P. T. Costa and R. R. McCrae, “Normal personality assessment in clinical practice: The NEO Personality Inventory.,” *Psychol. Assess.*, vol. 4, no. October, pp. 5–13, 1992.