A Robot Reading Human Gaze: Why Eye Tracking Is Better Than Head Tracking for Human-Robot Collaboration

Oskar Palinko, Francesco Rea, Giulio Sandini, Alessandra Sciutti

Abstract — **Robots are at the position to become our everyday companions in the near future. Still, many hurdles need to be cleared to achieve this goal. One of them is the fact that robots are still not able to perceive some important communication cues naturally used by humans, e.g. gaze. In the recent past, eye gaze in robot perception was substituted by its proxy, head orientation. Such an approach is still adopted in many applications today. In this paper we introduce performance improvements to an eye tracking system we previously developed and use it to explore if this approximation is appropriate. More precisely, we compare the impact of the use of eye- or head-based gaze estimation in a human robot interaction experiment with the iCub robot and naïve subjects. We find that the possibility to exploit the richer information carried by eye gaze has a significant impact on the interaction. As a result, our eye tracking system allows for a more efficient human-robot collaboration than a comparable head tracking approach, according to both quantitative measures and subjective evaluation by the human participants.**

I. 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 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.

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 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 and for example 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 gaze tracking devices are usually designed to be static, observing just one spot in space, while robots often need a solution that can deal with agents moving in space. The common solution adopted in experimental settings is then 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 II) 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. 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 or 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 [1]. 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.

In this work we add performance improvements to our calibration-free, visual light, monocular eye gaze tracking algorithm designed to work on humanoid robots. 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

*Research supported by the European Project CODEFROR (PIRSES-2013-612555)

All authors (members of IEEE) are with the Robotics, Brain and Cognitive Sciences Department of the Fondazione Istituto Italiano di Tecnologia, Genova, 16163, Italy (corresponding author’s e-mail: oskar.palinko@iit.it).

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. We measure the fluidity 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.

In this paper we will use the terms "head gaze", "head pose" and "head orientation" interchangeably to describe the 3D orientation of the head in space. To be precise, gaze by definition refers only to the direction in which the eyes are pointing in, but here we will use head gaze as an approximation of eye gaze. We will also alternate in using the terms "eye gaze tracking", "eye tracking" and "eye gaze estimation" as synonyms.

The next chapter will give an overview of previous work on head and eye gaze tracking in robotics, and position our study in this field. Chapter III will describe our gaze tracking system. Chapter IV will be dedicated to the experiment we designed to evaluate the efficiency of eye gaze tracking versus head gaze tracking alone. Chapter V will summarize our findings concerning the accuracy of the employed system, while subsequent chapters will discuss these results and propose conclusions.

II. Background

There are a number of different approaches to implementing gaze tracking which can be divided along different lines. First, they can be divided into *active and passive systems*. Active trackers use infrared illumination and cameras to cast light to gain a better view of the eye. They are less effective in daylight and at greater distances. Passive systems work in visual light, thus being more natural, although more sensitive to lighting conditions. Another important distinction is between *head mounted and remote solutions*. Head mounted systems require the human partner to wear cameras mounted either on helmets or spectacle frames. Helmet systems can provide high quality eye tracking results, but their weight renders them impractical for extensive use. Glass mounted systems are gaining popularity as they are lightweight and accurate. Remote systems, which track gaze from a location not connected to the human, can suffer from lower quality eye images, but work with no effort from the human side. Eye tracking solutions might also be differentiated on the basis of whether they require a *calibration* or not. Although the former option provides higher precision in gaze detection, the latter avoids any tedious calibration procedure and it makes it possible to interact with people "on the fly", with no prior preparation.

The approach that we believe more 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 chapter 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 [2]. 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 [3]. 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 [4]. 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.

Sheiki and Odobez explore attention recognition in HRI [5]. 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 [6]. 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 [2,3,4,5] authors either used direct substitution of gaze with head pose or utilized contextual information for training their machine learning methods. Sheiki and Odobez [5] 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 [7]. 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. [4]). 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.

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 [8] which was implemented on the HRP2 humanoid robot [9]. 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 [10]. 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 [11], 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.

III. Approach – Eye Gaze Tracking

In the following we will shortly describe our original eye gaze tracking system first introduced in [11] 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 1). Head pose was calculated using the Constrained Local Models (CLM) approach implemented by Baltrusaitis [12]. 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 [13] and implemented by King in [14]. 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. 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 [15]. 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 can reliably understand which objects are gazed upon by its interacting partner [11].

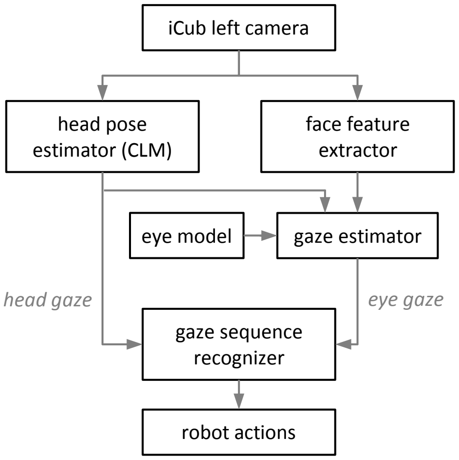

Figure 1. Eye tracking system diagram.

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 2), which was adequate for capturing even faster eye movements. To deal with possible momentary glitches triggering gaze events in

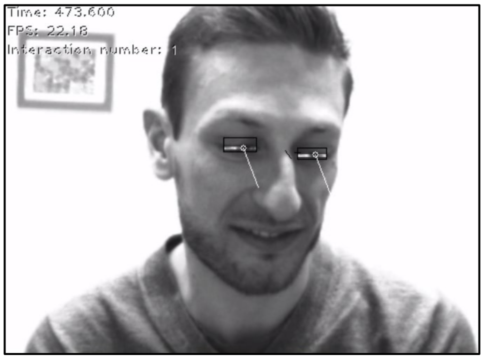

Figure 2. Example output of our eye tacking system. White lines from eyes – eye gaze, black line from top of the nose – head gaze.

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.

IV. The Experiment

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 3. 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 3, the separation between the robot's two arms was around 50° 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. 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.

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 4a) or first at one of the hands and then at the face (Figure 4b). 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.

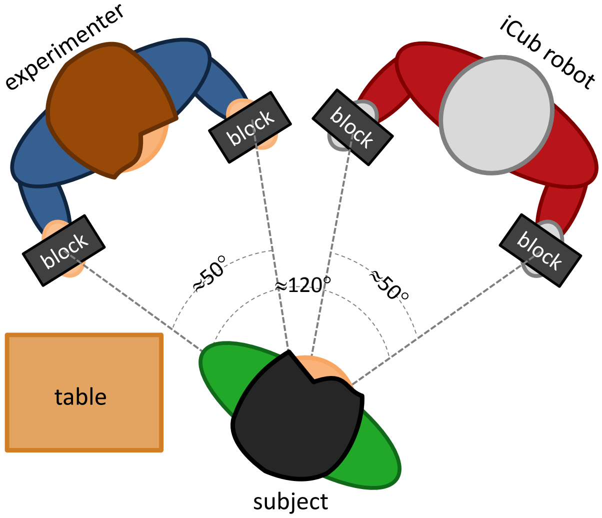

Figure 3. Experimental setup.

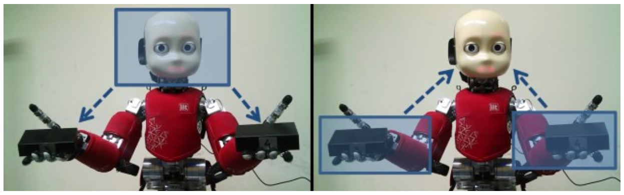

Figure 4. Gaze actions triggering robot reaction: a) gazing at face then one of the hands and b) gazing at a hand then at face.

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 accompanying video to understand how the robot behaved).

V. RESULTS

The experiment was completed by 10 subjects (3 females and 7 males) 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 1). 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 1. 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 5. Applying a paired t-test we found that the difference between the two conditions is statistically significant with t(9)=3.171, p=0.011.

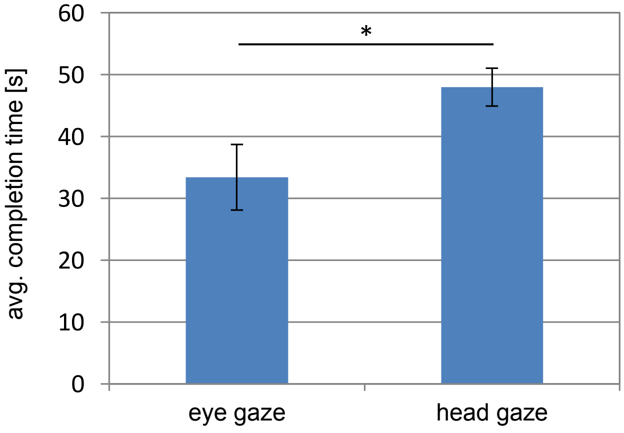

Figure 5. Task completion time averages over all subjects. Error bars represent +/- 1standard error. “*” indicates significant difference in a paired t-test.

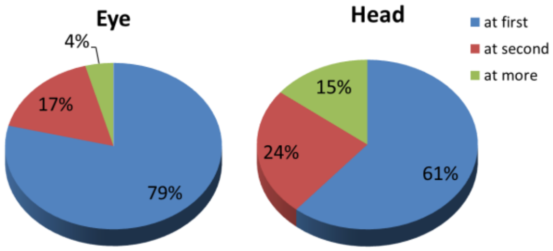

Figure 6. How many offering motions it took before the robot handed over the proper block. “At first”: the robot lifter the proper arm at once. “At second”: the robot lifted its wrong arm before lifting the proper one. “At more”: there were more than one wrong arm lifts before the proper one.

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 2.

TABLE 2. AVERAGE TURN DETECTION AND HANDOVER ACCURACY RATE WITH STANDARD ERROR 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 first author by reviewing and annotating the experiment videos.

From Table 2 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 Figure 6). 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.

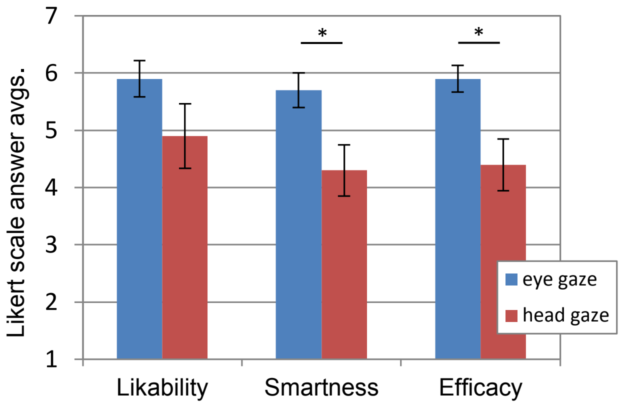

Figure 7. Subjective ratings of the two interaction types. Error bars represent +/- 1standard error. "*" indicates statistically significant difference.

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 7 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.

VI. Discussion

The results of our collaborative 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 subjects managed to complete all trials of the interaction, on average in less than 40s per trial. When relying on head gaze only, the interaction was slower and less effective: in five cases the subjects were not even able to complete their task (10% failure of all interactions) with the *head gaze* option.

Moreover, from the pie charts in Figure 6 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.

An interesting finding is that while head-based gaze estimation seems good enough for turn change detection, it is not as good for object recognition. 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-based 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.

Also the subjective results are quite informative of the fact that participants thought more positively of the better functioning option. Although there is no significant difference in likability, the *eye gaze* option was evaluated to be much more efficient and an indication of a smarter robot. It was interesting to see that in the *eye gaze* condition all 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.

VII. Conclusion

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 on the interaction of head-based and eye-based gaze tracking. 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 43% 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 Chapter II) 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 far away objects 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 from 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 as pointing and speech could lead to a very robust interactive robot.

To conclude, our open source 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.

References

[1] 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.

[2] M. W. Doniec, G. Sun, and B. Scassellati, "Active Learning of Joint Attention," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, 2006, pp. 34–39.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] J. Ido, Y. Matsumoto, T. Ogasawara, and R. Nisimura, "Humanoid with interaction ability using vision and speech information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, no. November 2008, pp. 1316–1321.

[10] 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.

[11] O. Palinko, A. Sciutti, F. Rea, and G. Sandini, "Eye Gaze Tracking for a Humanoid Robot," in *IEEE/RAS International Conference on Humanoids Robotics*, 2015.

[12] 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.

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

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

[15] 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.