2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)
November 3 - 5, 2015, Seoul, Korea

Eye Gaze Tracking for a Humanoid Robot

Oskar Palinko, Francesco Rea, Giulio Sandini, Alessandra Sciutti

Abstract— **Humans use eye gaze in their daily interaction with other humans. Humanoid robots, on the other hand, have not yet taken full advantage of this form of implicit communication. In this paper we present a passive monocular gaze tracking system implemented on the iCub humanoid robot. The validation of the system proved that it is a viable low-cost, calibration-free gaze tracking solution for humanoid platforms, with a mean absolute error of about 5 degrees on horizontal angle estimates. We also demonstrated the applicability of our system to human-robot collaborative tasks, showing that the eye gaze reading ability can enable successful implicit communication between humans and the robot. Finally, in the conclusion we give generic guidelines on how to improve our system and discuss some potential applications of gaze estimation for humanoid robots.**

I. 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 [1]. This is particularly evident when eye gaze processing is atypical, as in autism spectrum disorders, where such impairment is linked to social and communicative deficits [2]. 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 facilitate turn taking.

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 conversant 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.

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 paper we introduce a geometric feature-based passive gaze estimation system for the iCub humanoid robot [3] 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.

In the next sections we will give a brief overview of the existing head/eye gaze tracking approaches, with a focus on application in robotics and human-robot interaction. 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 paper will be concluded with a discussion of future potential improvements and applications.

II. Background

A. 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.

An extensive comparison of possible eye tracking systems for a humanoid robot is out of the scope of this paper. Rather, we will give a short comparison of existing systems, highlighting their advantages and disadvantages for implementation on a humanoid robot.

*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).

978-1-4799-6885-5/15/$31.00 ©2015 IEEE

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 also affect the interaction, by forcing subjects to be aware of their own gaze. Remote eye trackers provide less precision than head mounted ones but with an 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 and exhibit a low mobility. Embedding a remote eye tracker into the visual system of a mobile robot would help overcome the limited working area of the tracking device.

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 them 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.

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 [4]. Their advantage is that they can operate even on noisy images, but they require training. Feature-based eye tracking algorithms [5] 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.

B. 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. There are a number of gaze behaviors which directly support communication, as mutual gaze [1], joint attention [6], gaze aversion [7], etc. Mutual gaze happens when the gaze of two agents interlock, a phenomenon that has a special prosocial value since very early in human development [8]. 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. Finally, gaze aversion is the process where turn taking is modulated by gazing away, thus holding the floor in a conversation. It has been studied how robots can generate these cues, but less so how they can "read" them [9].

Interestingly, due to technical limitations, many authors have substituted eye gaze with its closest proxy, head orientation (e.g., [10][11][12]), 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 interaction 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 [13] 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.

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 [14][15]. 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 [16] while Ido and colleagues implemented it in the HRP2 humanoid robot [17] 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. More recently Sciutti et al. used a mutual gaze detection system for facilitating turn-taking in a dictation scenario on the iCub platform [18], by making the humanoid responsive to the establishment of eye contact. 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 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 III.B and III.D for details).

III. APPROACH

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 paper we will implement the system on the iCub platform.

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 [19]. 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 (see Section IV) 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.

A. Image Acquisition and Rectification

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.

B. Face and Face Features Detection

To accomplish this step we rely on the *dlib* library [20] which implements Khazemi and Sullivan's algorithm for precisely detecting a 68-point face feature set using ensembles of regression trees [21]. This algorithm starts from the Viola-Jones [22] 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 1.

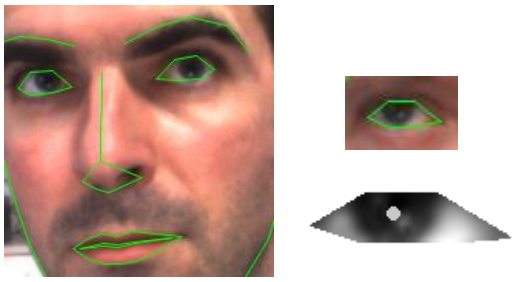

Figure 1. Left: face features detection as seen by the iCub. Right: eye area detection (top right) and extracted eye area with detected pupil center as seen by the iCub (bottom right).

C. Eye area and pupil center extraction

Both eyes are bound by a 6-point polygon as seen in Figure 1. 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.

D. 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 B) provides only a 2D representation of face points. Thus we adapted the constrained local models (CLM) approach by Baltrusaitis et al. [23]. 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.

E. Eye Model Geometry

In detecting the gaze we opted for a feature-based approach with using an eye model (cf. Section II). The method which we are adapting from Ishikawa et al. [5] 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 2). 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.

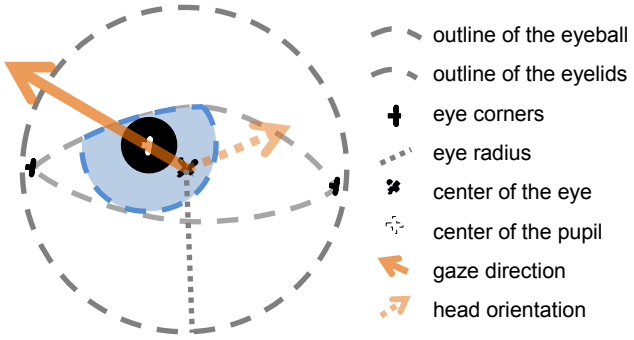

Figure 2. The model of the eye.

F. 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 [5]). 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

Figure 3. Gaze tracking estimates using our system.

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 Discussion).

Averaging is made possible by exploiting the fact that the distance between the eyes of people (interpupillary 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 [24]. Thus, we normalize each face by the distance between the two outer corners of the eyes, which is a very similar measure as the interpupillary 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 [25]. 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 is sufficiently reliable. The performance of the developed system is visualized in Figure 3.

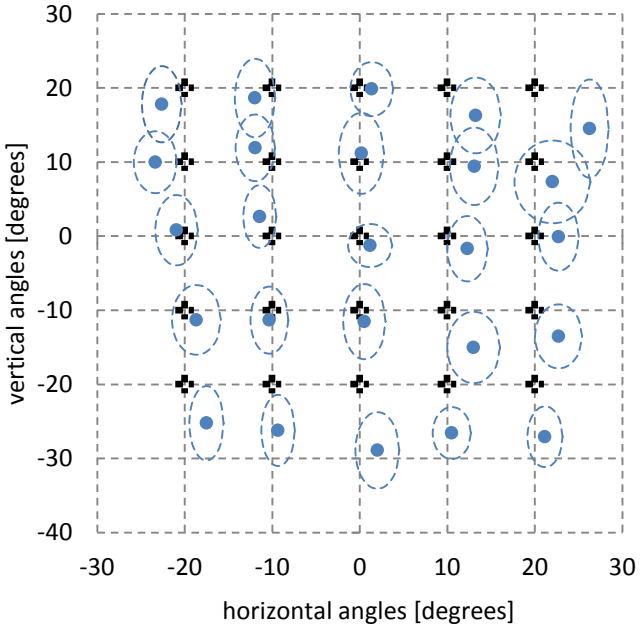

Figure 4. Nominal (cross) and estimated gaze locations (blue dots) with 1 standard error ellipses for the validation experiment at 60cm distance.

IV. 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 [26]. 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. We used a single board for both distances. The locations for near and far distances were marked with different shapes. The angle markings were created by assuming the same two distances for all subjects (60 and 100cm). 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. The robot's right eye, used for recording the images, was located at the origin of the board (0,0). Participants were positioned in front of the robot, with the eyes at the same height as the robot eye. 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 males and 2 females 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 4 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 5, where the estimated angles averaged across all head rotations and both viewing distances are plotted against the corresponding nominal angles separately for the horizontal and vertical components.

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 1).

Table 1. 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

It is interesting to note that the performance of our system in the horizontal plane seems to be in a similar range as that

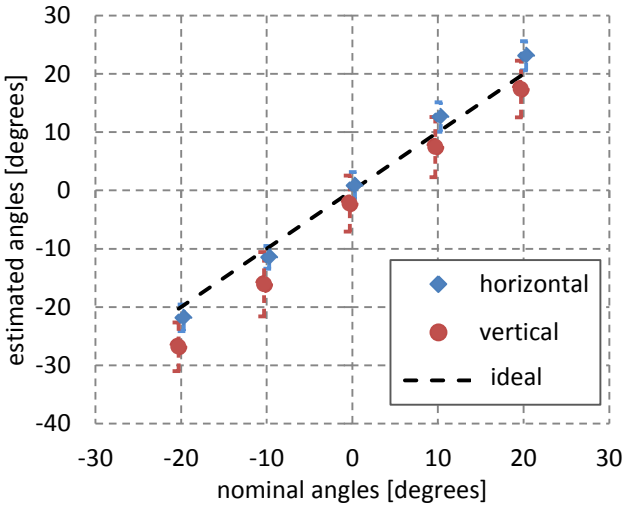

Figure 5. Average estimated gaze angles (+/- 1SE) along the horizontal (blue) and vertical (red) directions. Average across all subjects, distances and head orientations. Values for -20/+20 vertical are computed only on the near distance.

of a human observer, as measured in a slightly different task by S. Al Moubayed and G. Skantze [27].

The lower accuracy in the vertical plane was caused by the imperfect detection of eye corners, since the estimation of eye corner positions 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. One reason for this additional variability might be 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.

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^{o} and vertical 13.4^{o}). 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 him 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.

V. Human-Robot Interaction Scenario

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 shown in Fig. 6

Participants sat opposite the robot and the experimenter. At the beginning of the experiment the height of the chair was adjusted so that their eyes were approximately in line with the robot eyes. No further adjustments or constraints were applied to participants' position or movement. The participants' task was 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 6. 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 7. 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 [28]. To detect these sequences we defined activation rectangles around the three potential glance targets: the face of the robot, its left and its right hand. Whenever we detected two consequent gazes which switched from an arm-box to the face-box (or vice versa), the appropriate robot arm movement was triggered. The face and arm boxes were adjacent to each other vertically, with the border being at -10 degrees. As in the previous sections only the output of the right camera was used. 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 (see submitted video for an example of the experimental procedure). The experiment was completed by 7 subjects (5 males, 2 females, between 25 and 32 years of age). Three subjects wore eyeglasses.

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 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

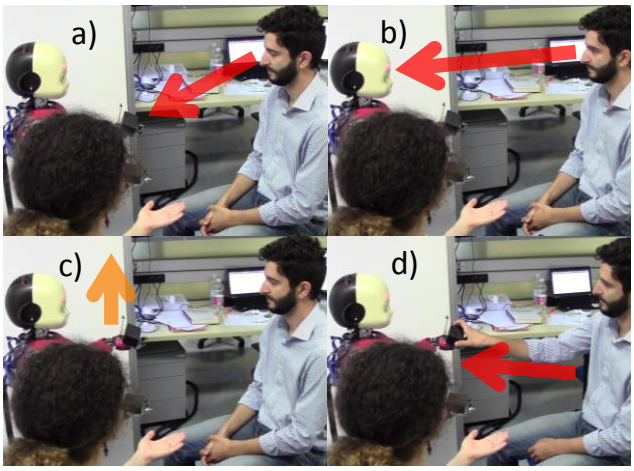

Figure 7. Interaction sequence: a) participant looks at robot's hand, then b) at the robot's face. c) The robot lifts its arm and finally d) the subject takes the block.

glasses, since different shapes of eyeglass lenses can have different effects on eye tracking systems [29]. If this subject is considered as an outlier and removed from the sample, then task completion rate changes to 100%, correct turn detection rate becomes 81.5%, and correct hand offering rates changes to 77.9%. It may be asked why this latter interaction error rate is still so high (more than 20%) when the system's horizontal average absolute error is only 5 degrees? First, most wrong hand selections occurred when participants needed to look down, where the gaze estimation is more imprecise due to eye corner shifts and iris occlusions (see section IV). However most of real life collaboration tasks include glances down (at tabletops for example) and therefore they need to be included in an ecologic scenario. Second, even if average gaze error is low, individual gazes can have much larger errors or even glitches when any element of the system (head orientation, face features detection, iris center estimation) has a momentary misrecognition. In the future we will work on eliminating these mis-estimations by filtering the output of the gaze tracking module over time.

VI. Discussion

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 propose 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 introduces more variance to our system, but it facilitates 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 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. We are currently looking for other solutions for more consistent eye corner detections. Since our software is designed to leverage on modularity, substituting certain modules with more effective alternatives will be facilitated.

The benefits of a built-in gaze tracker in a humanoid can be manifold: 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 V), 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 [30] 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 [31].

REFERENCES

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

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

[3] 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.

[4] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-Based Gaze Estimation in the Wild," *arXiv Prepr. arXiv1504.02863*, 2015.

[5] 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.

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

[7] 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.

[8] T. Farroni, G. Csibra, F. Simion, and M. H. Johnson, "Eye contact detection in humans from birth," vol. 2002, no. Track II, pp. 1–4, 2002.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] A. Sciutti, A. Bisio, and F. Nori, "Anticipatory gaze in human-robot interactions," *Gaze in HRI From Modeling to Communication* workshop at the 7th ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, 2012.

[16] 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.

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

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

[19] R. Beira, M. Lopes, M. Praça, J. Santos-Victor, A. Bernardino, G. Metta, F. Becchi, and R. Saltarén, "Design of the robot-cub (iCub) head," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2006, vol. 2006, pp. 94–100.

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

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

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

[23] 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.

[24] 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," 1989.

[25] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze Locking: Passive Eye Contact Detection for Human–Object Interaction," *ACM Symp. User Interface Softw. Technol.*, pp. 271–280, 2013.

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

[27] 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.

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

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

[30] 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 *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, 2014, pp. 1128–1133.

[31] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *J. Autism Dev. Disord.*, vol. 43, pp. 1038–1049, 2013.