Towards Better Eye Tracking in Human Robot Interaction Using an Affordable Active Vision System

Oskar Palinko
Istituto Italiano di Tecnologia
via Morego 30
16163 Genova, Italy
oskar.palinko@iit.it

Francesco Rea
Istituto Italiano di Tecnologia
via Morego 30
16163 Genova, Italy
francesco.rea@iit.it

Alessandra Sciutti
Istituto Italiano di Tecnologia
via Morego 30
16163 Genova, Italy
alessandra.sciutti@iit.it

Giulio Sandini
Istituto Italiano di Tecnologia
via Morego 30
16163 Genova, Italy
giulio.sandini@iit.it

HAI '14, Oct 29-31 2014, Tsukuba, Japan
ACM 978-1-4503-3035-0/14/10.
http://dx.doi.org/10.1145/2658861.2658926

Abstract
Knowing where a person is looking is an important parameter of every human-human interaction. Detecting a person's gaze could significantly improve the interaction capabilities of today's robotic agents. But many robots' visual systems are limited by data bandwidth and optical hardware. We propose a low-cost high-def pan/tilt/zoom active vision system that could significantly improve the robot's eye tracking capabilities. We tested the proposed system for improving mutual gaze detection in a human-robot interaction scenario and found significant results compared to systems without zoom capability.

Author Keywords
Eye tracking; human robot interaction; active vision.

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

Figure 1. Human-agent interaction.

Figure 2. Camera and motor setup.

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 robot camera system are 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 Fig. 1) which in its currently limited repertoire performs only pan and tilt movements. 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.

Related Work

Since their appearance, eye tracking systems have found many applications in human-machine interaction [2]. 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 [3]. Commercially available eye tracking systems, however, are in general very expensive and require ad hoc hardware, making the system less affordable and not customizable enough for the use in a robotic device. Our work proposes instead an active vision system developed by using affordable webcams. Drawing inspiration from the work of Atienza and Zelinsky [1], 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 [5], the exploration of which is also one of the goals of our research.

System Implementation

The robot head system consists of the visual and actuator system, see Figure 2.

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.

Because of the limited data bandwidth of the communication networks on modern robots we decided

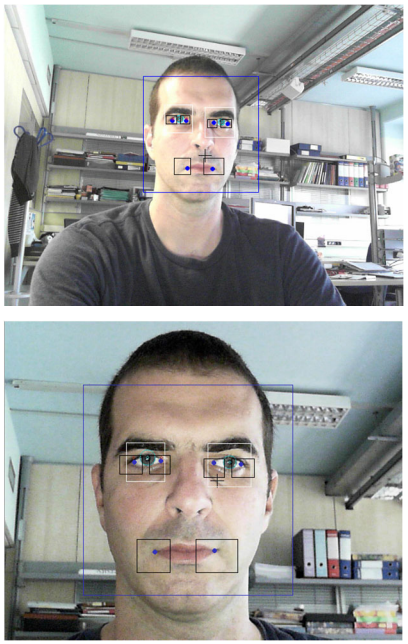

Figure 3. Zoomed out (above) and maximum zoomed in (below) images.

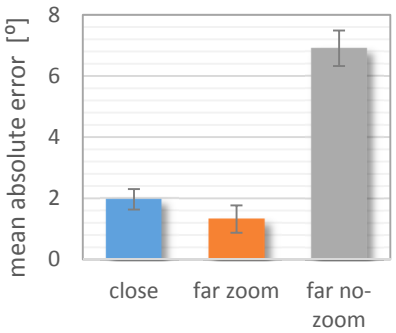

Figure 4. Mean absolute error.

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 [6]). 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 [4].

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 to track the face of the human conversant and keep their face in the middle of the image.

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 3. 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. 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.

Validation Experiment

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. 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 3. below) while the latter condition did not use any zoom (Figure 3. above). 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 4. shows the absolute error between real and detected gaze direction, 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

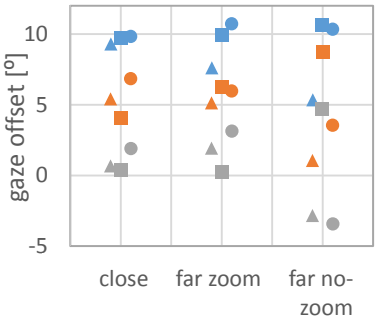

Figure 5. Average gaze points for each subject, each gaze offset and each condition. Triangular markers denote Subject01, squares Subject02 and circles Subject03. Gray markers denote straight ahead gazes, orange ones 5 degree offset, while blue stands for 10 degree offsets.

Future Work

We plan to further develop the eye tracking capabilities of our system in order to put the active vision hardware to full use. First steps will include ellipse matching of the iris and compensation for head movement/rotation. The long term goal is to make a robotic platform contingently react to the gaze of its human partner.

Acknowledgements

This work has been conducted in the framework of the European project CODEFROR (PIRSES-2013-612555).

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° 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 5. 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.

Conclusion and Discussion

In this paper 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 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 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 OpenCV library, which makes it simple to calibrate these low-cost cameras for manufacturing imperfections.

References

[1] Atienza, R., and Zelinsky, A., Active Gaze Tracking for Human-Robot Interaction, Proceedings of Intl. Conf. on Multimodal Interfaces, 2002.

[2] Jacob, R. J. K. and Karn, K. S., Eye tracking in human–Computer interaction and usability research, in The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, 2003.

[3] Matsumoto, Y., Sasao, N., Suenaga, T, and Ogasawara, T., 3D Model-based 6-DOF Head Tracking by a Single Camera for Human-Robot Interaction, Proceedings of ICRA 2009.

[4] Sandini, G., Questa, P., Scheffer, D., Diericks, B., & Mannucci, A., A retina-like CMOS sensor and its applications. Proc. of Sensor Array and Multichannel Signal Processing Workshop. *2000.*

[5] Scassellati, B., Imitation and mechanisms of joint attention: A developmental structure for building social skills in a humanoid robot. In Comput. for Metaphors, Analogy and Agents (Nehaniv, C., ed.), Vol. 1562,1998.

[6] Viola, P., and Jones, M., Rapid object detection using a boosted cascade of simple features. Proceedings CVPR, 2001.