Noname manuscript No.
(will be inserted by the editor)

Efficient analysis of gaze-behavior in 3D environments

Thies Pfeiffer · Patrick Renner · Nadine Pfeiffer-Lessmann

Published in 12th Biannual Conference of the German Cognitive Science Society (Gesellschaft fr Kognitionswissenschaft), Editors Anna Belardinelli and Martin V. Butz, Cognitive Processing, Volume 15, Issue 1 Supplement, September 2014.

Abstract We present an approach coined EyeSee3D to identify the 3D point of regard and the fixated object in real-time based on 2D gaze videos without the need for manual annotation. The approach does not require additional hardware except for the mobile eye tracker. It is currently applicable for scenarios with static target objects and requires fiducial markers to be placed in the target environment. The system has already been tested in two different studies. Possible applications are visual world paradigms in complex 3D environments, research on visual attention or human-human/human-agent interaction studies.

Keywords 3D eye tracking · natural environments

1 Introduction

Humans are evolved to live in a 3D spatial world. This affects our perception, our cognition and our action. If human behavior and in particular visual attention is analyzed in scientific studies, however, practical reasons often force us to reduce the three-dimensional world to two dimensions within a small field of view presented on a computer screen. In many situations, such as spatial perspective taking, situated language production, or understanding of spatial references, just to name a few, a restriction to 2D experimental stimuli can render it impossible to transfer findings to our natural everyday environments.

One of the reasons for this methodological compromise is the effort required to analyze gaze data in scenarios where the participant is allowed to move around and inspect the environment freely. Current mobile eye-tracking systems use a scene camera to record a video from the perspective of the user. Based on one or two other cameras directed at the participant's eyes, the gaze fixation of the participant is then mapped on the video of the scene camera. While binocular systems are already able to compensate for parallax by estimating the distance of the fixation from the observer, they have no representation of the 3D world but still only work on the 2D projection of the world visible in the scene camera video. The most important part then is identifying in the video stream that particular object the participant has been fixating. This currently requires manual annotations, which take several times as much as the recorded time. Depending on the complexity of the annotation (target object count and density), we had cases where the annotation of one minute recorded video required fifteen minutes of annotation or more.

With our EyeSee3D approach, we provide a software tool that is able to identify the fixated objects automatically if it can be allowed that the environment is covered with some visible markers that do not affect the visual behavior and if the target objects remain static.

Thies Pfeiffer
Center of Excellence Cognitive Interaction Technology, Bielefeld University, Germany
Tel.: +49-521-106-12373
E-mail: Thies.Pfeiffer@uni-bielefeld.de

Patrick Renner and Nadine Pfeiffer-Lessmann
SFB 673: Alignment in Communication, Bielefeld University, Germany
Tel.: +49-521-106-2918
E-mail: (prenner|nlessman)@techfak.uni-bielefeld.de

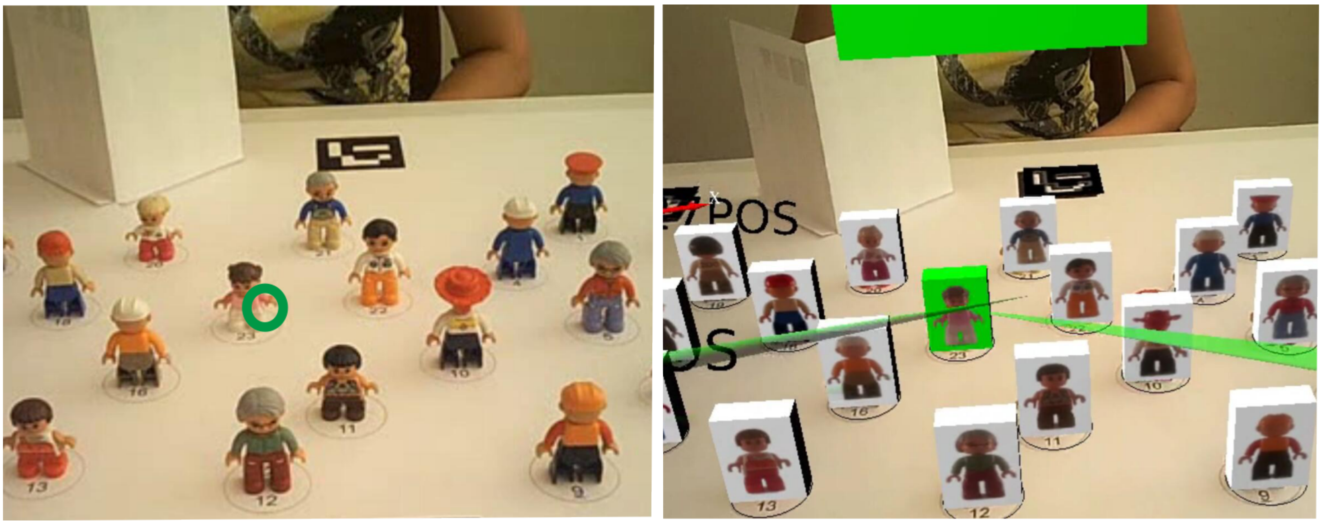

Fig. 1 The left snapshot is taken from a 2D mobile eye-tracking video taken from the egocentric perspective of the scene camera. The point of regard is visualized using a green circle and a human annotator would have to manually identify the fixated object, here the figure of a girl. With EyeSee3D, gaze-rays can be computed and cast into a 3D abstract model of the environment (simple white boxes around the figures), the intersection with the fixation target (box corresponding to figure of the girl) is computed automatically and in real-time.

2 Related Work

There are approaches for semi-automatic gaze annotation based on 2D computer vision, such as the SemantiCode approach by Pontillo et al. 2010 [4], which still requires manual annotation, but achieves a speed-up by incrementally learning the labelling of the targets using machine learning and computer vision techniques. Still, the experimenter has to at least validate every label. Approaches that also use 3D models are Toyama et al. 2012 [6], but they are targeting human-computer interactions, not scientific studies, and Paletta et al. 2013 [1], who use a 3D scan of the target environment to later identify the target position. Their approach requires much more effort during preparation but then does not require an instrumentation of the environments with markers.

3 Application Areas

The presented EyeSee3D approach can be applied as a method to accurately annotate fixations in 3D environments as required for scientific studies. We have already tested this approach in two studies. Both studies involve settings with two interacting interlocutors (no confederates) sitting face-to-face at a table.

In the first study, we were interested in gaze-patterns of joint attention [3]. We placed 23 figures of a LEGO Duplo set on a table, each of which facing either of the interlocutors. The experimenter then describes a certain figure and the interlocutors have to team up to identify the figure. The task, however, is not as simple as it sounds: the information given might only be helpful for one of the interlocutors, as it might refer to features of the figure only visible from a certain perspective. Even more, the interlocutors are instructed to neither speak nor gesture to communicate. This way we force the participants to use their gaze to guide their partner's attention towards the correct figure. The set-up used in this experiment will be used later in this paper to illustrate the EyeSee3D method.

In the second study, we were interested in creating computational models for predicting the targets of pointing gestures and more generally areas which in the near future will be occupied by a human interlocutor during interaction [5]. This research is motivated by human-robot interaction in which we want to enable robots to anticipate human movements in order to be more responsive, i.e., in collision-avoidance behavior.

Besides eye tracking, in this study we also combined the EyeSee3D approach with an external motion-tracking system to track the hands and the faces of the interlocutors. Using the same principles as presented in the next section, also the targets of pointing gestures as well as gazes towards the body of the interlocutor can be identified computationally without the need for manual annotations.

4 EyeSee3D

The EyeSee3D approach is easy to set-up. Figure 1 on the left shows a snapshot from one of our own studies on

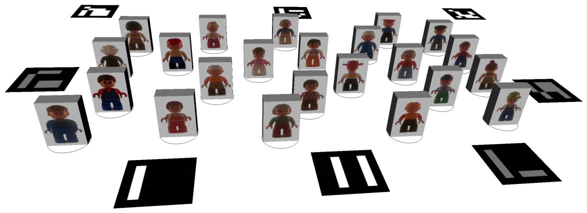

Fig. 3 The 3D proxy geometries that had to be created to determine the fixated objects. The different figures are textured with original pictures, which is not needed for the process but useful for controlling the orientation of the figures when setting up the experiment.

joint attention between two human interlocutors [3]. In this study we had 12 pairs of interaction partners and a total of about 160 minutes of gaze video recordings. It would have taken about 40 hours to manually annotate the gaze videos, excluding any additional second annotations to test for annotation reliability.

The process followed by EyeSee3D is presented in Figure 2. In a preparation phase, we covered the environment with so-called fiducial markers, highly visible printable structures that are easy to detect using computer-vision methods (see Figure 1, mid upper half). We verified that these markers did not attract significant attention by the participants. As a second step, we created proxy geometries for the relevant stimuli, in this example small toy figures (see Figure 3). For our set-up, a simple approximation using bounding boxes is sufficient, but any complex approximation of the target may be used. When aiming for maximum precision, it is possible to use 3D scans with exact replications of the hull of the target structure. The whole process for setting up such a table will take about 30 minutes. These preparations have to be made once, as the created model can be used for all study recordings.

Based on this preparations, we are now able to conduct the study and record the eye-tracking data (gaze videos and gaze data). EyeSee3D then automatically annotates the recorded gaze videos. For each frame of the video, the algorithms detect fiducial markers in the image and estimate the position and orientation of the scene camera in 3D space. For this process to succeed at least one fiducial marker has to be fully visible in each frame. The camera position and orientation are then used together with the gaze information provided by the eye tracker itself to cast a gaze ray into the 3D proxy geometries. This gaze ray intersects the 3D proxy geometries exactly at the point (see Figure 1, right) that is visualized by the gaze cursor in the scene camera video provided by the standard eye-tracking software (see Figure 1, left). As each of the proxy geometries is labeled, we can identify the target object automatically.

This annotation process can be either used online during the study, so that the annotation results are already available when the study session is completed. Or, alternatively, EyeSee3D can be used in offline-mode to analyse the previously recorded gaze videos and data files. This offline-mode has the advantage that it can be repeatedly applied to the same data. This is useful in cases where number and placement of the proxy geometries is not known beforehand and incrementally refined during the progress of understanding the problem domain. For example, at the moment we are only interested in locating the target figure. Later on we might be working together with psycholinguists on language processing following a visual-world paradigm. We might then be also interested in whether the participants have looked at the headdress, the head, the upper body or the lower body of the figures during sentence processing. After updating the 3D proxy models, we could use EyeSee3D to re-annotate all videos and have the more fine-grained annotation ready within minutes.

In our example study, we were able to cover about 130 minutes of the 160 minutes of total recordings using this technique. In the remaining 30 minutes, participants were either moving their head so quickly that the scene camera only provided a motion-blurred image or they turned towards the interaction partner or the experimenter for questions, so that no marker was visible in the image (but also no target stimuli). Thus, the remaining 30 minutes where not relevant for the evaluation of the study.

More technical details about the EyeSee3D approach have been presented at ETRA 2014 [2].

5 Discussion and Future Work

The presented initial version of our EyeSee3D approach can already significantly speed-up the annotation of mobile eye-tracking studies. There are no longer economic reasons to keep an eye on short sessions and low number of participants. The accuracy of the system depends on the one hand on the accuracy of the eye-tracking system. In this the accuracy of EyeSee3D does not differ from the normal 2D video-based analysis. On the other hand the accuracy depends on the quality with that the fiducial markers are detected. The larger the detected marker and the better the contrast, the higher the accuracy of the estimated camera position and orientation.

EyeSee3D is not only applicable for small setups, as the selected example of two interaction partners sitting at a table might suggest at first glance. The size

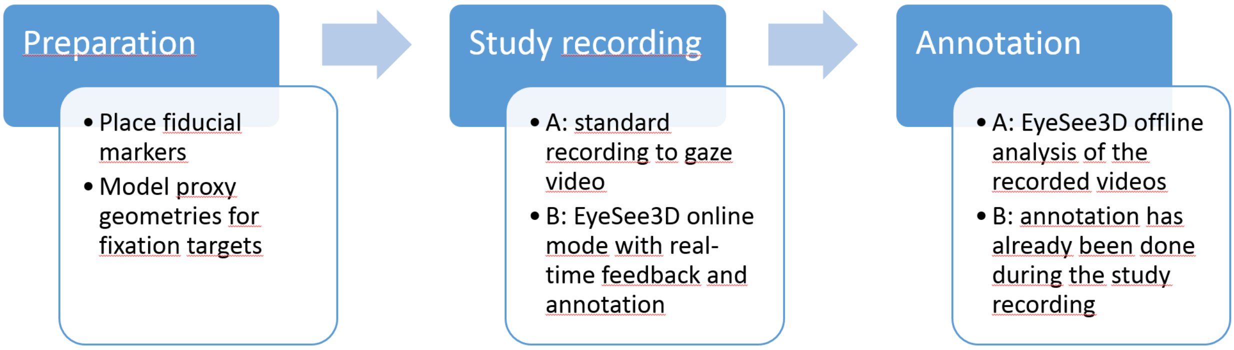
Preparation
• Place fiducial markers
• Model proxy geometries for fixation targets
Study recording
• A: standard recording to gaze video
• B: EyeSee3D online mode with real-time feedback and annotation
Annotation
• A: EyeSee3D offline analysis of the recorded videos
• B: annotation has already been done during the study recording

Fig. 2 The EyeSee3D method requires a one-time preparation phase. During study recording there are two alternatives, either (A) use the standard tools and run EyeSee3D offline to annotate the data or (B) use EyeSee3D online during the study.

of the environment is not restricted as long as at least one fiducial marker is in the field of view for every relevant target object. The markers might, for example, be sparsely distributed in a museum just around the relevant exhibits.

We are currently working on further improving the speed and the accuracy of the system. In addition to that, we are planning to integrate other methods for tracking the scene camera's position and orientation in 3D space based, e.g., on tracking arbitrary but significant images. In certain examples such as a museum or a shelf in a shopping center, this would allow for an automatic tracking without any dedicated markers.

In future work, we are planning to compare the results obtained by human annotators with those calculated by EyeSee3D. In a pilot evaluation we were able to identify situations of disagreement, i.e. situations in which EyeSee3D comes to slightly different results as a human annotator, when two target objects overlap in space (which is more likely to happen with a freely moving participant than in traditional screen-based experiments) and the fixation is somewhere in between. Such situations are likewise difficult to annotate consistently between human annotators, because of their ambiguity. Investigating the differences between the systematic and repeatable annotations provided by EyeSee3D and the interpretations of human annotators, which might depend on different aspects, such as personal preferences or the history of preceding fixations, could be very informative. Besides the described speed-up achieved by EyeSee3D, it might also provide more objective and consistent annotations.

In summary, using EyeSee3D the analysis of mobile gaze-tracking studies has become as easy as desktop-computer-based studies using remote eye-tracking systems.

Acknowledgements This work has been partly funded by the DFG in the SFB 673 Alignment in Communication.

References

1. Paletta, L., Santner, K., Fritz, G., Mayer, H., Schrammel, J.: 3D attention: measurement of visual saliency using eye tracking glasses, CHI 13 Extended Abstracts on Human Factors in Computing Systems, 199204, ACM, Paris, France (2013).
2. Pfeiffer, T., Renner, P.: EyeSee3D: A Low-Cost Approach for Analysing Mobile 3D Eye Tracking Data Using Augmented Reality Technology, Proceedings of the Symposium on Eye Tracking Research and Applications, ACM (2014)
3. Pfeiffer-Lessmann, N., Pfeiffer, T., Wachsmuth, I.: A Model of Joint Attention for Humans and Machines. Book of Abstracts of the 17th European Conference on Eye Movements (Bd. 6), 152, Lund, Sweden (2013)
4. Pontillo, D. F., Kinsman, T. B., Pelz, J. B.: SemantiCode: using content similarity and database-driven matching to code wearable eyetracker gaze data, ACM ETRA 2010, 267270, ACM (2010)
5. Renner, P., Pfeiffer, T., Wachsmuth, I.: Spatial references with gaze and pointing in shared space of humans and robots. Proceedings of the Spatial Cognition 2014, (2014, to appear)
6. Toyama, T., Kieninger, T., Shafait, F., Dengel, A.: Gaze guided object recognition using a head-mounted eye tracker, ACM ETRA 2012, 9198, ACM (2012)