UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Visual Gaze Estimation by Joint Head and Eye Information

Valenti, R.; Lablack, A.; Sebe, N.; Djeraba, C.; Gevers, T.

DOI
10.1109/ICPR.2010.1160

Publication date
2010

Document Version
Submitted manuscript

Published in
2010 20th International Conference on Pattern Recognition: ICPR 2010: Istanbul, Turkey, 23-26 August 2010

Link to publication

Citation for published version (APA):
Valenti, R., Lablack, A., Sebe, N., Djeraba, C., & Gevers, T. (2010). Visual Gaze Estimation by Joint Head and Eye Information. In *2010 20th International Conference on Pattern Recognition: ICPR 2010: Istanbul, Turkey, 23-26 August 2010* (pp. 3870-3873). IEEE. https://doi.org/10.1109/ICPR.2010.1160

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

Visual Gaze Estimation by Joint Head and Eye Information

Roberto Valenti[1], Adel Lablack[2], Nicu Sebe[3], Chabane Djeraba[2], and Theo Gevers[1]

[1]Intelligent Systems Lab Amsterdam, University of Amsterdam, The Netherlands
[2]Laboratoire d'Informatique Fondamentale de Lille, University of Lille, France
[3]Department of Information Engineering and Computer Science, University of Trento, Italy

Abstract

In this paper, we present an unconstrained visual gaze estimation system. The proposed method extracts the visual field of view of a person looking at a target scene in order to estimate the approximate location of interest (visual gaze). The novelty of the system is the joint use of head pose and eye location information to fine tune the visual gaze estimated by the head pose only, so that the system can be used in multiple scenarios. The improvements obtained by the proposed approach are validated using the Boston University head pose dataset, on which the standard deviation of the joint visual gaze estimation improved by 61.06% *horizontally and* 52.23% *vertically with respect to the gaze estimation obtained by the head pose only. A user study shows the potential of the proposed system.*

1 Introduction

Visual gaze estimation is important in many applications, spanning from human computer interaction (HCI) to human behavior analysis. In applications where human activity is under observation from a static camera, the estimation of the visual gaze provides important information about the interest of the subject, which is commonly used as control devices for disabled people [2], to analyze the user attention while driving [4], and other applications.

It is known that the visual gaze is a product of two contributing factors [10]: the head pose and the eye locations. The estimation of these factors is often achieved using expensive, bulky or limiting hardware [1]. Therefore, the problem is often simplified by either considering the head pose or the eye center locations as the only feature to understand the interest of a user [11, 15]. While this assumption could be considered valid in certain scenarios [16, 20], it is not reliable in every situation. For instance, if we consider an environment composed of a target scene (a specific scene under analysis, such as a computer monitor, an advertising poster, a shelf, etc.) and a monitored area (the place from which the user looks at the target scene), an eye gaze tracker would fail when trying to understand which product on the shelf is being observed, while an head pose gaze estimator would fail in finely control the cursor on a computer screen.

There is an abundance of literature concerning these two distinct topics: recent surveys on head pose and eye center location estimation can be found in [12] and [5]. Within the appearance-based methods for eye location proposed in literature, [13, 6, 7, 18] reported results which support the evidence that accurate appearance based eye center localization is becoming feasible and that it could be used as an enabling technology for a various set of applications. However, no study was performed on the feasibility of an accurate appearance based visual gaze estimator which considers both the head pose and the eye location factors.

Therefore, our goal is to build a system capable of analyzing the visual gaze of a person starting from monocular video images. This allows to study the movement of the user's head and eyes in a more natural manner than traditional methods, as there are no additional requirements needed to use the system.

The rest of this paper is structured as follows: In the next section the methods used for the estimation of the head pose and the eye center location are described. We then provide the methodology used for the extraction of the gaze information in Section 3. In Section 4, the proposed method is evaluated using the Boston University datasets and by a user study on the final system. Finally, the conclusions are discussed in Section 5.

2 Head Pose and Eye Location Estimation

To be able to correctly estimate the visual gaze of a user, joint head pose and eye location information should be taken into account. The needed information are retrieved by our system proposed in [19], which is able to accurately extract head pose and eye center location information from a monocular video sequence.

Figure 1. Example of eye reference points (yellow dots) and detected eye center locations (green dots).

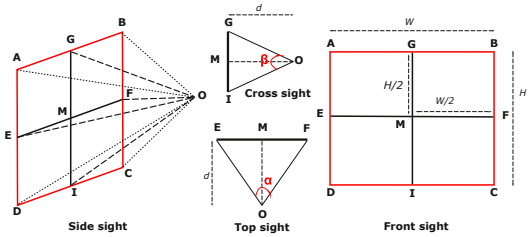

Figure 2. Representation of the visual field of view at a distance d.

The method combines a robust cylindrical head model (CHM) pose tracker [21] and an isophote based eye center locator [18], in order to obviate drawbacks of both systems when taken independently.

The system integrates the eye locator with the cylindrical head model by using the transformation matrices obtained by both systems in an interleaved way. In this way the eye center locations are detected given the pose, and the pose is adjusted given the eye center locations. The 2D eye center locations detected in the first frame are used as reference points (the yellow markers in Figure 1). These reference points are projected onto the cylindrical head model and are then used to estimate the successive eye center locations.

Instead of projecting the found eye center location back to the image plane to obtain a more accurate eye center estimation, here the detected eye center locations are projected to a normalized model view. Thus, the displacement between the reference eye point and the current eye center location will be independent of the head pose. In the next section, these displacement vectors will be used as the eye component to obtain the sought after joint visual gaze estimation.

3 Visual Gaze Estimation

Studies on the human visual field of view [14] show that it has a vertical span of $130°$ ($60°$ above and $70°$ below) and approximately $90°$ on each side, which corresponds to a photographic objective angle of $180°$. The common field of view of the two eyes is called binocular field of view and spans $120°$. It is surrounded by two monocular field of view of approximately $30°$.

The visual field of view of a human can be approximated by a pyramid $OABCD$ where the starting point O represents the point between the two eyes and the rectangle $ABCD$ represents the boundary of the visual field of view at a distance d. Besides the distance, the calculation of the width W and the height H of the rectangle $ABCD$ requires the angles α and β, which represent the horizontal and vertical angles of the visual field of view in binocular human vision, respectively [8] .

The point M is the middle point of the visual field and, when looking straight ahead, the vector OM will represent the visual gaze vector (Figure 2).

The width (W) and the height (H) of the visual field at a distance d are computed by:

$$W = 2MF = 2d.\tan\frac{\alpha}{2} \quad , \quad H = 2MG = 2d.\tan\frac{\beta}{2}$$

The projection of the visual field of view on a target scene (the scene in front of the user) is a quadrilateral $A'B'C'D'$ with a central point M' , and it is constructed from the intersection between the target scene plane P : $ax + by + cz + d = 0$ and the lines (OA), (OB), (OC), (OD), and (OM). After a calibration step (using the method that we suggested in [9]), the head pose parameters obtained using the system described in Section 2 are used to extract the region of interest in the target scene.

The visual gaze of a person indicates the point of interest (*i.e*. the foveated point) and corresponds to the middle of the visual field of view. Therefore, the displacement of the eyes from their resting position plays an important role in the properties of the visual field of view. Instead of focusing on modeling the shape of the eyes, and how eye movements alter the visual gaze vector in relation to the head position, we assume that extreme eye positions are rare (*e.g*. the visual gaze vector does not fall outside the visual field of view defined by the head pose only). This assumption allows us to approximate the eye movement as horizontal and vertical shifts on the head surface, which is in turn modeled as a cylinder. As stated in Section 2, the current eye locations and the reference eye locations are used to compute the eye displacement vectors in a normalized (*i.e*. pose-free) model view. The displacement vectors of the two eyes are averaged and used to modify the basic head-only visual gaze estimate.

Note that this new gaze estimate will always be bounded by the computed visual field, as we assumed that the maximum allowed correction for the visual gaze vector never crosses the boundary of the computed visual field. The result is a visual gaze point M_{disp} which represents the displacement of the normal visual gaze

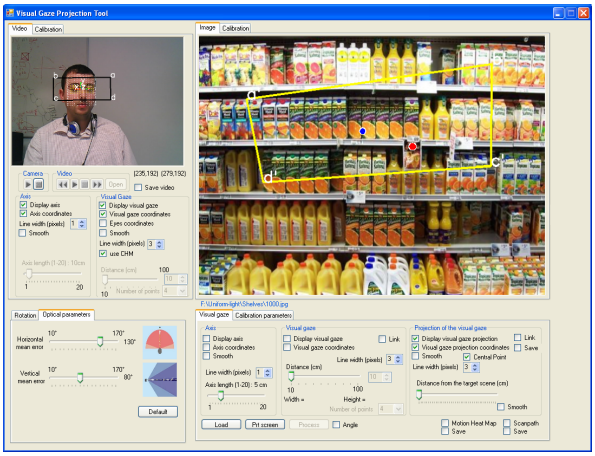

Figure 3. A Screenshot of the final system, working in real time using a simple webcam. The yellow quadrilateral indicates the user's region of interest (defined by head pose only), while the red dot is the combined head and eyes visual gaze estimation.

point M. The projection of this point M'_{disp} (represented by the red dot in Figure 3) is obtained by the intersection between the plane P and the line (OM_{disp}) by $M_{disp}(M_x + W * disp_x * w, M_y + H * disp_y * w)$, where w is a weighting factor which represents how much the eye displacement vectors will affect the final gaze vector. In our experiments, we empirically found that a weighting factor w of 0.2 achieves the best results, as a weighting factor of 0.1 has a little effect on the final estimation, while a weighting factor superior to 0.3 generates many outliers. These findings agree with the study in [17], where it is shown that head pose contributes to about 70% of the visual gaze.

4 Experiments & Results

In order to evaluate the described method, we tested it on a publicly available head pose dataset. Furthermore, a user study is performed to asses the potential of the final system.

4.1 Boston University Dataset

To have better insight of the accuracy of the system, we tested it on a subset of the Boston University head pose dataset [3]. The dataset consists of 45 video sequences, where 5 subjects were asked to perform 9 different head motions under uniform illumination in a standard office setting. The head is always visible and there is no occlusion except for some minor self-occlusions. The videos are recorded in low resolution and a Flock of Birds tracker records the pose information coming from the magnetic sensor on the person's head. This system claims a nominal accuracy of 1.8 mm in translation and 0.5 degrees in rotation. Since the subjects in this dataset were not requested to perform any gazing task, the ground truth of the visual gaze is not available. In most of the videos, however, the subject constantly gazes at the camera as his head pose changes (as displayed in Figure 4). As this eye movement is supposed to compensate for the head pose, we expect the coordinates of the projected head and eyes gaze vector to have a smaller standard deviation than the one obtained solely by the head pose.

Figure 4. Example of the Boston University head pose dataset: some subjects keep gazing at the camera, even if they were not instructed to do so.

Therefore, the deviation of the combined head and eye visual gaze with respect to the head pose alone is analyzed on the subjects of the BU dataset which are gazing at the camera during the experiment.

Figure 5 shows one example (for subject jam8) in which the use of eye displacements (dots) significantly helps in reducing the standard deviation of the visual gaze given by the head pose only (crosses). Note how the gaze is localized around the same location (*i.e.* the camera location) when including the eye location information. The full experimentation on the subset of the BU database shows that the standard deviation of the gaze points was reduced by 61.06% on the x dimension and 52.23% in the y dimension, with respective stds of 20.48% and 19.05%.

4.2 User Study

For the user study, we aimed to investigate how different subjects will achieve tasks with their visual gaze in two different scenarios. Twenty-two undergraduate students participated in the study. The mean age of the participants was 22.8. As target scenes, we used a 19" LCD monitor placed 50 cm in front the participant and a shelf displaying numbered markers (*i.e.* products) placed at 1.5 to 2 meters away from the subject. The subjects were filmed by a webcam with a resolution of 320x240 pixels. The participants were presented with specific tasks to achieve by looking at the target scene, using either using the head and the eyes or the head alone. The progress and the completion of the tasks were notified by a visual or an audio signal (for the

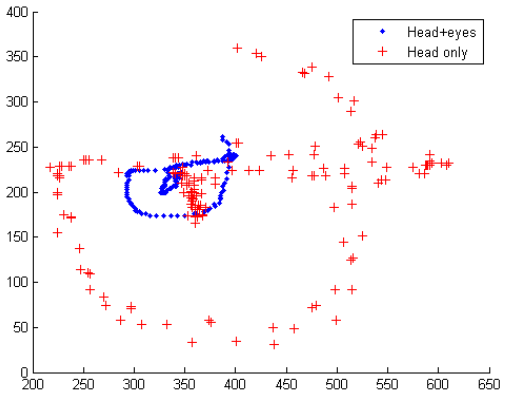

Figure 5. An example of differences between a joint head and eye gaze estimation (.) and head only gaze estimation (+), on the movie jam8.avi.

monitor and for the shelf, respectively). The subjects were asked to perform the tasks as quickly and as accurately as possible and to fill in a questionnaire about the feeling of the system. By analyzing the questionnaires, we made the following observations: (1) every subject found the combined eye and head visual gaze estimation more natural than the head only; (2) even though the results were similar, tasks were completed faster when using the combined head and eye visual gaze estimation; (3) moving from a point A to a point B which are far away was more difficult than expected, as the system might have diverted to wrong paths; (4) on the monitor, the users preferred smooth cursor movements instead of direct jumps to the gazed locations. Overall, all the subjects were able to successfully use the system in both scenarios.

5 Conclusions

In this paper, we have presented an unconstrained system that estimates the visual gaze of a person looking at a target scene. A non sequential system was used to estimate the head pose and the eye center locations. This multimodal information was combined in order to project the visual gaze of a person on the target scene. The eye displacements information was used to fine tune the gaze vector obtained by the head pose.

Evaluation using the Boston University head pose dataset proves that joint eye and head information results in a better visual gaze estimation, reducing the standard deviation on the target scene of about 61.06% horizontally and 52.23% vertically. The user study proves that the system is able to usefully combine head and eye information to better estimate visual gaze in a more natural manner.

References

[1] Cogain: communication by gaze interaction, gazing into the future. http://www.cogain.org, September 2006.

[2] J. S. Agustin, J. P. Hansen, and J. Mateo. Gaze beats mouse: hands-free selection by combining gaze and emg. In *CHI*, 2008.

[3] M. L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *PAMI*, 22(4), 2000.

[4] A. Doshi and M. M. Trivedi. On the roles of eye gaze and head pose in predicting driver's intent to change lanes. *ITS*, 10(3), September 2009.

[5] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *PAMI*, 32(3), 2010.

[6] S. Kim, S.-T. Chung, S. Jung, D. Oh, J. Kim, and S. Cho. Multi-scale gabor feature based eye localization. In *World Academy of Science, Engineering and Technology*, 2007.

[7] B. Kroon, A. Hanjalic, and S. M. Maas. Eye localization for face matching: is it always useful and under what conditions? In *CIVR*, 2008.

[8] A. Lablack, F. Maquet, and C. Djeraba. Determination of the visual field of persons in a scene. In *VISAPP*, January 2008.

[9] A. Lablack, F. Maquet, N. Ihaddadene, and C. Djeraba. Visual gaze projection in front of a target scene. In *ICME*, June 2009.

[10] S. R. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5), 2004.

[11] X. Liu, N. Krahnstoever, T. Yu, and P. Tu. What are customers looking at? In *Advanced Video and Signal Based Surveillance (AVSS)*, 2007.

[12] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, 31(4), 2009.

[13] Z. Niu, S. Shan, S. Yan, X. Chen, and W. Gao. 2d cascaded adaboost for eye localization. In *ICPR*, 2006.

[14] J. Panero and M. Zelnik. *Human Dimension and Interior Space: A Source Book of Design Reference Standards*. Watson-Guptill, 1979.

[15] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006.

[16] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *PAMI*, 30(7), July 2008.

[17] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems*, 2002.

[18] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *CVPR*, 2008.

[19] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *CVPR*, 2009.

[20] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *ICMI*, 2008.

[21] J. Xiao, T. Kanade, and J. Cohn. Robust full motion recovery of head by dynamic templates and re-registration techniques. In *FG*, 2002.