Estimating Fusion Weights of a Multi-Camera Eye Tracking System by Leveraging User Calibration Data

Nuri Murat Arar and Jean-Philippe Thiran *
Signal Processing Laboratory (LTS5), École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

Cross-ratio (CR)-based eye tracking has been attracting much interest due to its simple setup, yet its accuracy is lower than that of the model-based approaches. In order to improve the estimation accuracy, a multi-camera setup can be exploited rather than the traditional single camera systems. The overall gaze point can be computed by fusion of available gaze information from all cameras. This paper presents a real-time multi-camera eye tracking system in which the estimation of gaze relies on simple CR geometry. A novel weighted fusion method is proposed, which leverages the user calibration data to learn the fusion weights. Experimental results conducted on real data show that the proposed method achieves a significant accuracy improvement over single camera systems. The real-time system achieves 0.82° of visual angle accuracy error with very few calibration data (5 points) under natural head movements, which is competitive with more complex model-based systems.

Keywords: eye gaze tracking, multi-camera, fusion, integration, weighting, estimation, learning, calibration, cross-ratio

Concepts: •Computing methodologies → Computer vision;

1 Introduction

Remote video-based eye tracking methods can be classified mainly into two groups, namely, interpolation-based and model-based methods [Hansen and Ji 2010]. Interpolation-based methods map image features to gaze points using machine learning while model-based methods mostly estimate three-dimensional (3D) gaze direction by modeling the eye in 3D. The intersection between scene geometry and gaze direction is computed as the point of regard (PoR). System requirements of interpolation-based methods tend to be smaller than model-based methods but they are suited to particular applications due to their limitations regarding accuracy and head movements. Model-based methods offer greater freedom of movement, however, they require more complex system setups such as camera and geometric calibration. Contrary to these methods, CR-based methods, e.g. [Yoo et al. 2002; Hansen et al. 2010; Huang et al. 2014; Arar et al. 2015b], share advantages from both interpolation and model-based methods. They do not only avoid camera calibration, but they also allow free head motion. Unfortunately, the performance of CR-based methods might be limited in accuracy and robustness due to the simplifications assumed which cause certain estimation bias [Kang et al. 2008].

In the literature, several efforts have been made in order to enhance the accuracy and robustness of CR-based gaze estimation systems by performing a subject-specific calibration to correct the estimation bias. For instance, homography-based bias correction by [Hansen et al. 2010] has been widely accepted as it provides a good calibration when there is not much head pose or location change when there is sufficient number of calibration points. Moreover, a linear regression-based calibration has recently been shown to model the bias more efficiently with fewer number of calibration points even though there is certain amount of head movement [Arar et al. 2015b]. Besides, a few other approaches were proposed to explicitly bring robustness against head movements ([Coutinho and Morimoto 2013],[Huang et al. 2014]).

Traditional eye tracking systems are commonly based on a single camera, and the ones using a multi-camera setup are mostly designed for the purpose of obtaining stereo vision. For instance, [Beymer and Flickner 2003] propose a four-camera system that estimates the 3D gaze direction based on a detailed 3D eye model. They use a wide field of view (FOV) stereo for face detection and a narrow FOV stereo for eye tracking.

Despite a few attempts ([Utsumi et al. 2012; Arar et al. 2015a]), the effectiveness of multi-camera setups, which jointly utilize several independent camera systems, has not adequately been investigated. [Utsumi et al. 2012] propose a multi-camera system to obtain a wide observation area. They use two cameras which are placed on the sides of a gaze-reactive signboard. However, their application scenario does not require precise gaze estimation as observed from the reported mean accuracy error which is $> 11°$. Their focus is to allow for a wide range of head motions and rotations. Alternatively, in our previous work, we exploit, for the first time, a multi-camera setup to enhance the estimation accuracy in the scope of precise gaze estimation [Arar et al. 2015a]. The setup outputs an overall PoR through an adaptive fusion of three independent camera systems. We first calculate an initial PoR using simple averaging of available gaze information, and then assign fusion weights according to the distances between the initial PoR and the cameras. This proof of concept study demonstrates that a multi-camera setup with a simple fusion approach results in a more accurate eye tracking.

In this paper, we investigate how to more efficiently combine the gaze data obtained from multiple camera systems. In order to enhance the estimation accuracy, we propose to leverage the user calibration data for estimating the fusion weights. Since user calibration is inevitable to compensate for the estimation bias, exploiting it for the purpose of fusion weights estimation does not cause any additional overhead. To this effect, we first define a few indicators such as the calibration accuracy, gaze availability and a histogram of the best performing sensors for each calibration point. We calculate several statistics with respect to these indicators to generate weight maps. Finally, we perform a weighted fusion of sensors during testing with the weights learned from the calibration data. The experiments conducted on real data show that the proposed system and weighting method produce significantly better results than a single camera system and a multi-camera system with simple fusion methods. Thus, the main contribution of this paper can be stated as the introduction of an offline fusion weights estimation method in order to achieve an improved accuracy for multi-camera eye tracking systems.

*e-mails:{murat.arar, jean-philippe.thiran}@epfl.ch

ETRA 2016, March 14 - 17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03
DOI: http://dx.doi.org/10.1145/2857491.2857510

The rest of the paper is organized as follows: Section 2 gives a detailed description of the proposed system. Experimental results are given in Section 3. Finally, Section 4 concludes the paper.

2 Overview of the Proposed System

The overview of the proposed multi-camera setup is illustrated in Figure 1. The overall system consists of single camera gaze estimation systems. Each single camera system is practically independent such that their feature detection and gaze estimation processes are independent. The details of the system are explained in the following sections.

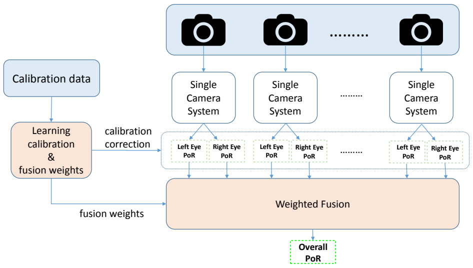

Figure 1: *Overview of the proposed multi-camera system.*

2.1 Hardware Setup

The system consists of three cameras, seven groups of LEDs for the illumination and a controller unit for the synchronization as shown in Figure 2. The cameras have a resolution of 1280×1024, and a 12 mm manual focus lens is used. LEDs are placed on the corners of the monitor to create glints and also placed as a ring around each camera to create the bright pupil effect. A micro-controller is programmed to synchronize the cameras and LEDs to obtain interlaced dark and bright pupil images at 30 frames per second (fps). In the current setup, the user sits approximately 70 cm away from a 24-inch monitor with a resolution of 1920×1200.

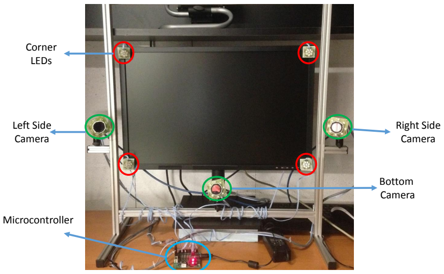

Figure 2: *Hardware setup.*

2.2 CR Gaze Estimation with User Calibration

We employ the original CR method [Yoo et al. 2002] for the estimation of the PoR. In CR method, a virtual tangent plane on the cornea surface, where four glints lie on, is assumed to exist. Therefore, the polygon formed by the glints is the projection of the monitor on the cornea. Another projection takes place from the corneal plane to the image plane, obtaining the glints and the projection of the pupil center. As the virtual tangent plane on the cornea has the same planar projective transformation of the monitor and image planes, the pupil center on image plane corresponds to the PoR on the monitor, which can be computed by equality of the cross-ratios.

CR-based gaze estimation has a limited performance due to the simplification assumptions such as non-coplanarity of the pupil and glints planes, and the angular offset between visual and optical axes of the eye [Kang et al. 2008]. Since the cornea curvature and the angular offset are subject-specific, a calibration needs to be performed to compensate for the estimation bias. The calibration is performed once, prior to the use of the system by asking the users to gaze at certain points on the monitor. Subject-specific bias correction is learned by minimizing the distances between the estimated gaze positions and the corresponding calibration points on the monitor.

In this paper, we use a linear regression-based calibration method ([Arar et al. 2015b]) to model the error vectors since it has been shown to have better modeling and generalization capabilities than the homography-based methods due to reduced model parameters and relaxed constraints. This method simply learns a linear transform between the estimated points and target points, and the learned transform is applied on the test samples for the bias correction.

2.3 Estimation of Fusion Weights

The hardware setup allows free head movement and captures both eyes simultaneously so that it enables to output two PoRs for the same frame for each camera system. This way our system generates multiple PoRs for a frame, one for each sensor (i.e., each eye of a camera system) and fusing them in an effective way would lead to an enhanced estimation accuracy compared to using a single camera system as in most of the previous work. Hence, we propose to combine PoRs in a weighted way in which we leverage the user calibration data statistics to estimate the weights. The proposed technique is, in fact, independent of the gaze estimation algorithm, therefore, the CR method used in this paper can practically be replaced with any other gaze estimation method (e.g., interpolation, 3D model). Once the weight maps are obtained, the proposed method performs a weighted averaging of individual PoRs as follows:

$$\mathbf{z}^* = \sum_c \sum_e \mathbf{z_c^e} * \mathbf{M_c^e}(\mathbf{z_c^e}.x, \mathbf{z_c^e}.y) \tag{1}$$

$$\sum_c \sum_e \mathbf{M_c^e}(x, y) = 1, \quad e \in \{L, R\}, c \in \{0, 1, 2\},$$

where $\mathbf{z}^*$ is the overall PoR, $\mathbf{z_c^e}$ are the individual PoRs, and $\mathbf{M_c^R}$ and $\mathbf{M_c^L}$ are the weight maps of the right and left eye of c-th camera, respectively. In case one of the PoRs can not be calculated for a given frame, the weight of the missing PoR is considered as zero. We do not report an overall PoR in case all sensors are unavailable.

For generating the weight maps ($\mathbf{M_c^e}$) we calculate different statistics on the calibration data. For instance, an effective weighting indicator would be the calibration accuracy per sensor on calibration points. The reason is that if the calibration accuracy on a point is consistently lower for a sensor than the others, that sensor's bias correction during testing is expected to be less reliable and accurate around the same point. Hence, the calibration accuracy based weighting assigns higher weights to the sensors whose bias corrections are more reliable, and so, a more accurate overall PoR can be computed. To calculate the calibration accuracy ($acc_{c,k}^e$) for each point, after learning the calibration model on the whole calibration data, we apply the learned model on the very same data. Then, we measure how close the calibrated samples are to their corresponding target points. We perform this process for each calibration point of each sensor separately. As we perform calibration for each eye of each camera independently, we obtain 6 values for each calibration point. We then normalize these accuracy values to compute the sensor weights ($w_{c,k}^e$) for each calibration point as shown in (2). Lastly, we interpolate and extrapolate the weight set ($\mathbf{W_c^e}$) over the whole monitor to generate the weight maps ($\mathbf{M_c^e}$). A set of generated weight maps is shown in Figure 3.

$$w_{c,k}^e = \frac{acc_{c,k}^e}{\sum_c \sum_e \sum_k acc_{c,k}^e} \tag{2}$$

$$\mathbf{W_c^e} = \{w_{c,k}^e | e \in \{L, R\}, c \in \{0, 1, 2\}, 1 \leq k \leq K$$

where K is the number of calibration points.

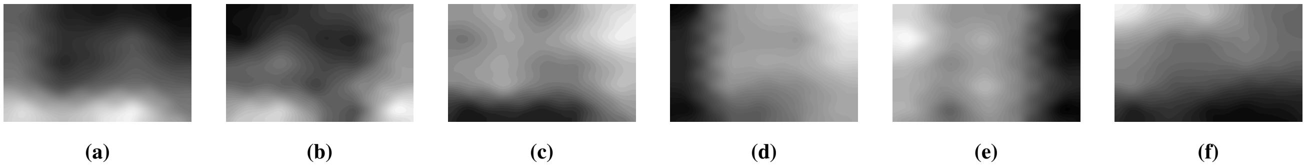

Figure 3: *Calibration accuracy and gaze availability based weight maps of the* **(a)** *right eye of the bottom camera;* **(b)** *left eye of the bottom camera;* **(c)** *right eye of the right camera;* **(d)** *left eye of the right camera;* **(e)** *right eye of the left camera;* **(f)** *left eye of the left camera.*

In addition, we use sensors' gaze availability statistics on each calibration point and a histogram of the best performing sensor on each calibration point as alternative weighting indicators. The gaze availability may indicate the reliability of feature detection. A low availability implies less consistent and less reliable features. Hence, a sensor with a higher availability is more likely to produce reliable PoRs. Similarly, the histogram of the best performing sensor stores the information about how often each sensor gets the best result for a given calibration point. If a sensor consistently gets the best result for a calibration point, it is more likely for the sensor to provide a more accurate PoR during testing. Maps generated by all the methods are supplied in the supplementary materials. Note that all these statistics can be calculated in a subject-specific and subject-independent manner. In order to investigate the subject influence on the fusion weights, we estimate the weights in both manners.

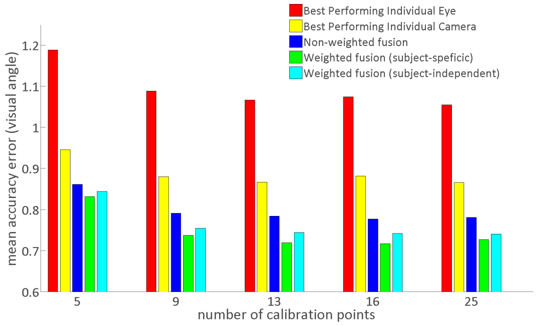

Figure 4: *Comparison of the proposed weighted multi-camera fusion with non-weighted fusion and without camera fusion.*

3 Experiments and Results

We conducted user experiments to evaluate and to compare the performances of the investigated fusion methods. Ten users participated in the experiment. Users were asked to look at the target stimuli points in a way that they feel comfortable. Therefore, we did not use a chin rest to keep user's head still and to keep user's eyes within the cameras' FOV to capture high-resolution eye data.

User experiments consist of acquiring the calibration and test data separately. In calibration data acquisition, users were asked to look at 25 uniformly distributed target stimuli points on the screen. In test data acquisition, users were asked to look at 18 target points in a 3×3 grid covering the whole screen. Test points are randomly generated inside the grids to avoid overfitting on the calibration points as well as creating a more realistic test condition. We report our eye tracker's performance as the gaze estimation accuracy error, which is defined as the average displacement between the real stimuli point and the estimated PoR. We report the estimation performance in degrees of visual angle as it is user distance invariant.

Our evaluation process starts with face tracking on the frames where we extract eye regions of size ∼130×70 pixels. We then detect features (pupil center and four glints) on the extracted regions using the methods in [Arar et al. 2015a]. Next, we apply CR based gaze estimation to calculate the raw gaze data. In the calibration process, we learn an estimation bias correction model on the raw gaze data. In the test process, we apply the learned model to correct the raw gaze data. The calibrated PoRs of all sensors are combined by the proposed weighted fusion scheme to output an overall PoR.

We obtain results using different weighting approaches based on the calibration accuracy, gaze availability, and histogram of the best performing sensor, and their combinations. Amongst all, the best performance is achieved by the method using the calibration accuracy and gaze availability combination to estimate the fusion weights. Table 1 and Figure 4 and 5 demonstrate the performance comparison of different camera setups and fusion methods. In Table 1, we list the results obtained from all camera setups with separate eye data as well as the multi-camera setup with the proposed fusion approach. Figure 4 shows firstly that multi-camera fusion with any kind of fusion method improves significantly the overall estimation performance compared to without camera fusion. Secondly, in all calibration configurations, the proposed fusion method achieves lower accuracy errors than the non-weighted multi-camera fusion. Note that the performance improvement becomes more significant when we learn the weights on the calibration data which consists of more than 5 points. This indicates that more calibration data leads to a more accurate estimation of fusion weights. Yet the performance seems to saturate after 9 points for the calibration. In addition, we observe that there is not a big performance difference between subject-specific and subject-independent estimation. Subject-specific estimation leads to a slightly better performance. However, subject-independent estimation can also accurately capture the overall tendency of weighting, and performs better than non-weighted fusion. This implies that the weighting is mostly based on hardware factors such as positions and viewing directions of the cameras and perhaps the monitor size. They are partly influenced by other subject-specific factors such as subjects' heights, gazing habits, eye dominances or vision problems. In order to take these factors into account, subject-specific estimation seems a better choice for a more accurate estimation. Since the calibration data needs to be acquired for each subject anyway, we propose to estimate the weights in a subject-specific manner.

In addition to accuracy enhancement, a multi-camera system provides a higher gaze estimation availability (see Table 1). It brings more flexibility for users' head and body movements. In single camera systems, the system can not output a PoR whenever the user is out of FOV of the only camera, whereas the user needs to be out of FOV of all the cameras in a multi-camera setup. Our user experiments do not contain large body movements, so there is a very high gaze availability even when only using the bottom camera, and the increase achieved by the multi-camera setup is not critical. To obtain a more concrete conclusion regarding the effectiveness of a multi-camera system on the gaze availability and robustness against head/body movements, another user study containing larger head/body movements must be performed. We leave such a study as our future work.

Figure 5 illustrates the comparison of the investigated methods with the previous work in the literature. The results indicates a significant improvement achieved by multi-camera setups over the best performing single camera system and the binocular fusion proposed

Table 1: *Comparison of different camera setups. Average gaze estimation accuracy errors are reported in degrees of visual angle.*

Camera Setup	Eye Data	Without Calibration	Number of Calibration Points					Gaze (%) Availability
			5	9	13	16	25	
Bottom Camera	Only Right Eye	7.14	1.22	1.16	1.15	1.16	1.14	95.8
	Only Left Eye	6.63	1.19	1.09	1.07	1.08	1.06	91.7
	Combined	5.12	0.95	0.88	0.87	0.88	0.87	96.4
Right Side Camera	Only Right Eye	4.69	1.61	1.48	1.46	1.43	1.45	89.2
	Only Left Eye	8.38	1.59	1.33	1.29	1.28	1.26	77.4
	Combined	4.79	1.32	1.23	1.21	1.19	1.20	91.9
Left Side Camera	Only Right Eye	7.53	1.56	1.43	1.58	1.49	1.47	76.7
	Only Left Eye	4.39	1.65	1.42	1.36	1.31	1.34	83.1
	Combined	4.68	1.35	1.19	1.22	1.17	1.18	90.0
Multi-camera	Overall	3.41	0.82	0.73	0.72	0.72	0.72	97.3

by [Zhang and Cai 2014]. The proposed fusion method further improves the estimation accuracy compared to [Arar et al. 2015a]. Besides, we plot another result showing the performance as if there is an oracle knowing the best performing camera for each frame in order to highlight the upper limits of the system through an optimal weighting system, and that this choice of the best camera is a critical factor for high accuracy. The oracle results imply a possible further enhancement.

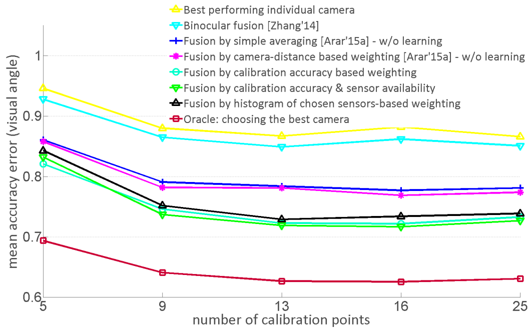

Figure 5: *Comparison of investigated methods with previous work.*

As shown in Figure 5, the error reduces with increasing amount of calibration data. However, a user-friendly system should involve as little effort as possible for the subject-specific calibration. Therefore, our results suggest a calibration with 9 points as the improvement with more points is not significant. Yet the system can reach a reasonable estimation accuracy of 0.82° with a calibration with only 5 points. Hence, it shows comparable performance to more complex 3D model-based systems whose reported accuracies are around 1° in [Hansen and Ji 2010].

Moreover, the proposed system brings another advantage, that is a lower computational complexity than 3D model-based methods. So the system is highly suitable for real-time gaze tracking. In fact, our system implemented in C++ can run at ∼30 fps, without performing any speed optimization, on a PC with Intel i7 3.2GHz processor.

4 Conclusion

In this paper, we investigate different fusion techniques to improve the overall gaze estimation accuracy of a multi-camera eye tracking system. We present a novel method which estimates the fusion weights by exploiting user calibration data statistics for efficiently combining multiple independent camera systems. The proposed method determines the weights using the statistics of certain weight indicators such as the calibration accuracy and available gaze data per sensor. The effectiveness of the proposed method has been validated with user experiments. The results show that the system's performance, even with very few calibration data (5 points), is competitive with more complex systems presented in the literature. Hence, the proposed system enables fast and user-friendly gaze tracking with minimum user effort without sacrificing too much accuracy. As the future work, we plan to investigate the robustness of the method against large head/body movements.

Acknowledgements

This project is supported by the Swiss Commission for Technology and Innovation (CTI) under grant number 13594.1 PFFLR-ES. Authors would like to thank Dr. Hua Gao for his valuable discussions.

References

ARAR, N. M., GAO, H., AND THIRAN, J.-P. 2015. Robust gaze estimation based on adaptive fusion of multiple cameras. In *FGR*.

ARAR, N. M., GAO, H., AND THIRAN, J.-P. 2015. Towards convenient calibration for cross-ratio based gaze estimation. In *WACV*, 642–648.

BEYMER, D., AND FLICKNER, M. 2003. Eye gaze tracking using an active stereo head. In *CVPR*, 451–458.

COUTINHO, F. L., AND MORIMOTO, C. H. 2013. Improving head movement tolerance of cross-atio based eye trackers. *IJCV 101*, 3, 459–481.

HANSEN, D. W., AND JI, Q. 2010. In the eye of the beholder: a survey of models for eyes and gaze. *PAMI 32*, 3, 478–500.

HANSEN, D. W., AGUSTIN, J. S., AND VILLANUEVA, A. 2010. Homography normalization for robust gaze estimation in uncalibrated setups. In *ETRA*.

HUANG, J.-B., CAI, Q., LIU, Z., AHUJA, N., AND ZHANG, Z. 2014. Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *ETRA*.

KANG, J. J., EIZENMAN, M., GUESTRIN, E. D., AND EIZENMAN, E. 2008. Investigation of the cross-ratios method for point-of-gaze estimation. *Transactions on Biomedical Engineering 55*, 9, 2293–302.

UTSUMI, A., OKAMOTO, K., HAGITA, N., AND TAKAHASHI, K. 2012. Gaze tracking in wide area using multiple camera observations. In *ETRA*.

YOO, D. H., KIM, J. H., LEE, B. R., AND CHUNG, M. J. 2002. Non-contact eye gaze tracking system by mapping of corneal reflections. In *FGR*.

ZHANG, Z., AND CAI, Q. 2014. Improving cross-ratio based eye tracking techniques by leveraging the binocular fixation constraint. In *ETRA*.