Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras

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

***Abstract*— Gaze movements play a crucial role in human-computer interaction (HCI) applications. Recently, gaze tracking systems with a wide variety of applications have attracted much interest by the industry as well as the scientific community. The state-of-the-art gaze trackers are mostly non-intrusive and report high estimation accuracies. However, they require complex setups such as camera and geometric calibration in addition to subject-specific calibration. In this paper, we introduce a multi-camera gaze estimation system which requires less effort for the users in terms of the system setup and calibration. The system is based on an adaptive fusion of multiple independent camera systems in which the gaze estimation relies on simple cross-ratio (CR) geometry. Experimental results conducted on real data show that the proposed system achieves a significant accuracy improvement, by around 25%, over the traditional CR-based single camera systems through the novel adaptive multi-camera fusion scheme. The real-time system achieves $<0.9^\circ$ accuracy error with very few calibration data (5 points) under natural head movements, which is competitive with more complex systems. Hence, the proposed system enables fast and user-friendly gaze tracking with minimum user effort without sacrificing too much accuracy.**

I. INTRODUCTION

As eye movements are natural and fast, they are suitable to interact with a computer vision system as a modality for user interfaces. Therefore, robust estimation and tracking of gaze, that is to accurately determine user's point of regard (PoR) on the screen, is of great interest for the development of HCI applications.

Remote video-based gaze tracking is preferred for interactive applications as they are non-intrusive. Video-based gaze estimation methods can be mainly categorized into two groups [13]: model-based methods [1], [2], [3] and interpolation-based methods [4]. Model-based methods estimate the line of sight by modeling the eye in 3-dimensional (3D) space. They require complex system setups such as camera calibration and geometric system calibration, however, they provide large head motion tolerance. On the other hand, interpolation-based methods estimate the PoR by mapping the image features to the gaze points. However, they are only suited to particular applications due to their limitations regarding accuracy and head movements. As alternative to these methods, CR-based methods [10], [9], [8], [5], [6], [12] share advantages of both interpolation and model-based methods. They allow free head motion without requiring any camera or geometric system calibration. The main drawback of CR-based methods is that they may be limited in accuracy and robustness due to the simplifications assumed. There are two major sources of estimation bias in CR-based methods [7]. Firstly, the model assumes that the pupil center and corneal reflections, i.e. glints, are coplanar. This assumption is not valid as the cornea has a spherical surface. Secondly, the model computes the PoR by considering the eye ball's optical axis rather than the visual axis, i.e. the real line of sight. The resulting estimation bias can be compensated for through subject-specific calibration in order to achieve high accuracy and head movement robustness.

In the original CR method [10], there was no estimation bias correction. Later, many extensions have been proposed to compensate for the estimation bias. For instance, homography-based bias correction as in [8], [11] has been widely accepted. In case of generic HCI scenarios in which users gaze at their monitor, most of the time no abrupt change is observed in head pose or head location. For such scenarios homography-based bias correction work well when there is sufficient number of calibration points. Moreover, in a recently published work [12], we have shown that regularized least-squares regression (LSR) can be utilized for robustly modeling the bias, and that LSR-based bias correction achieves higher accuracy than homography-based methods when there are fewer calibration points. Additional approaches have also been proposed ([5], [6]) to bring robustness against extreme head movements.

The multi-camera setup systems are mostly designed for the purpose of obtaining stereo vision since it allows for 3D eye modeling for model-based gaze estimation systems as in [1], [2], [3]. For instance, Beymer et al. [1] propose a four-camera system that can estimate the 3D gaze direction based on a complicated 3D eye ball model with several parameters. They use a wide field of view (FOV) stereo for face detection and a narrow FOV stereo for eye tracking. On the other hand, there are only a few interpolation-based and CR-based methods using multi-camera setup. In [4], two cameras are used to form a stereo vision system where the gaze is estimated through a nonlinear mapping through support vector regression. A pan-tilt unit oriented setups, e.g. [9], form another usage of multi-camera systems. In such systems, a narrow FOV camera, which is used to capture high-resolution eye data, is mechanically oriented by a wide FOV head camera using the pan-tilt unit.

Despite some attempts since many years, we believe that the effectiveness of multi-camera setups has not adequately been investigated. Regarding gaze tracking, there are only a few multi-camera based works which jointly utilize several independent camera systems. Utsumi et al. [15] propose such a system to obtain a wide observation area. They use two cameras which are placed on the left and right sides of a

*e-mail: murat.arar@epfl.ch

gaze-reactive signboard. However, their application scenario does not require precise gaze estimation as it is observed from the reported mean accuracy error which is $>11^{\circ}$. Their focus is to allow for a wide range of head motions and rotations. Furthermore, [16] presents a three-camera setup in which they use multiple cameras to robustly estimate the head pose and to increase working volume. Similar to [15], their system does not focus on the precise gaze estimation. They detect the eyes from different cameras, and then they jointly estimate the head pose from multi-camera eye data. Lastly, they perform head pose-based gazed region estimation in an indoor setting.

In this paper, a multi-camera gaze estimation system is presented which requires minimum effort in terms of the system setup and calibration. The system is based on adaptive fusion of multiple independent camera systems. The main contribution of this paper is two-fold:

1) We introduce a multi-camera setup for the purpose of precise gaze tracking. Since the single camera systems are independent, and the estimation of the gaze relies on simple CR geometry in each camera system, the setup requires neither camera calibration nor geometric system calibration unlike most of the previous work in the literature. In addition, the system does not need high-resolution eye data to reach high estimation accuracy. Instead we capture video frames of the whole face with visible but lower resolution of the eye pair. This enables each independent camera system to output PoRs for both eyes.
2) We propose a novel adaptive multi-camera fusion scheme in order to achieve improved estimation accuracy and coverage. The proposed scheme is independent of the chosen gaze estimation algorithm. It performs distance-based camera weighting to assign weights to PoRs obtained from independent camera systems, and outputs an overall PoR for each frame. We demonstrate that the proposed system achieves significant performance improvement over single camera systems, and the system's performance is comparable to other more complex systems under natural head movements.

Furthermore, this study targets a generic HCI environment. We collected ground truth data separately for subject-specific calibration for the bias correction and testing. We capture the user data in a natural manner (no use of a chin rest) in which the users were not particularly asked to move or keep their heads still with respect to the monitor. In addition, we use a new evaluation scheme where the test points are not chosen among the calibration points but are generated randomly covering the whole monitor. This does not only prevent overfitting on the points, but also creates a more natural and realistic test condition.

The rest of the paper is organized as follows: Section II explains a detailed description of the proposed system. Experimental results and discussions are given in Section III. Finally, Section IV concludes the paper.

II. PROPOSED SYSTEM

As the main contributions of this paper, we present a multi-camera setup and propose a novel adaptive fusion of multiple single camera systems to achieve increased working volume as well as improved estimation accuracy (Fig. 1). Each single camera gaze estimation system consists of gaze features detection and precise gaze estimation processes (Fig. 2). The details of the system are explained in the following sections.

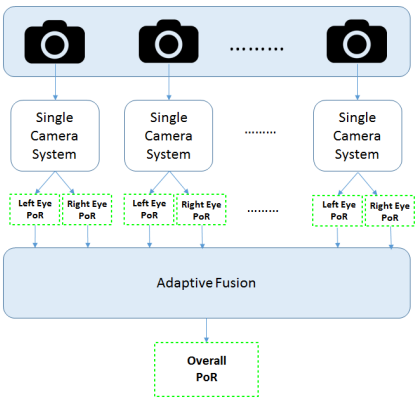

Fig. 1. Multi-camera system overview.

A. Hardware Setup

Our system consists of three PointGrey Flea3 monochrome cameras for the video capturing, seven groups of near-infrared (NIR) LEDs for the illumination and a controller unit for the synchronization. The cameras have a resolution of 1280×1024, and a 12 mm manual focus lens is used. They are installed on a frame around the monitor as shown in Fig. 3. One of the cameras is located below the monitor while the other two are placed symmetrically on the left and right sides of the monitor. In order to create the glints, 4 groups of NIR LEDs with 850 nm wavelength are placed on the corners of the monitor. Band-pass filters around 850 nm are used to get rid of the ambient light. A group of LEDs is placed as a ring around the lens of each camera to create the bright pupil effect. A micro-controller is programmed to synchronize the cameras and LEDs in order to obtain interlaced dark and bright pupil images at 30 frames per second. In addition, we synchronize the LEDs with cameras' shutters to minimize the emitting duration regarding the user eye safety. In the current setup, the user sits approximately 70 cm away from a 24-inch monitor with a resolution of 1920×1200.

B. Gaze Features Detection

We employ a robust non-rigid face tracker based on supervised decent method (SDM) [17] in order to localize the facial features. We extract the eye regions without performing any registration or scaling to ensure any particular resolution. Then, we determine whether there is an eye blink by the positioning of the landmarks around the eyes. If there is no eye blink, we perform gaze features detection, namely,

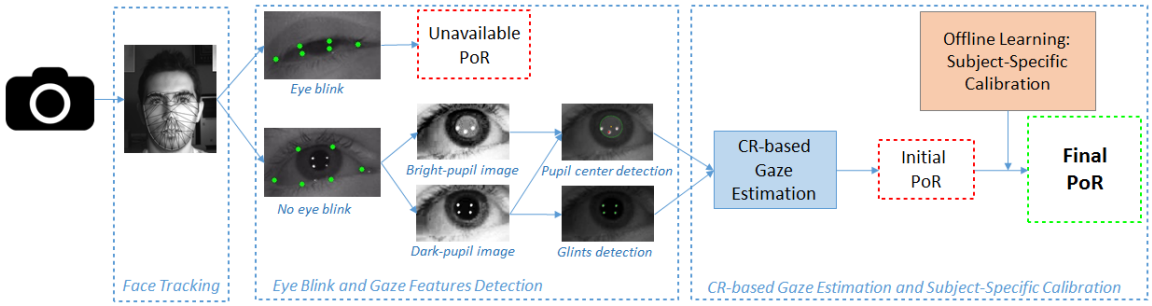

Fig. 2. Single camera system overview.

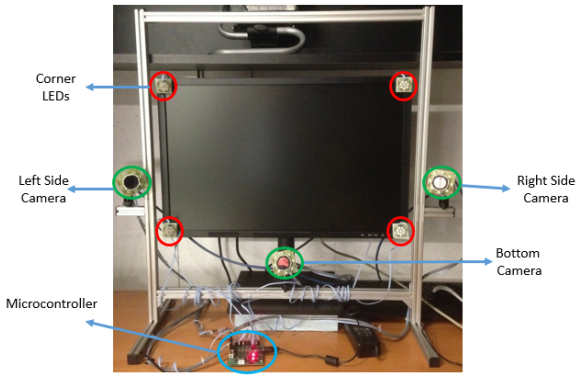

Fig. 3. Hardware setup.

glints and pupil center detection. For the glints detection, we make use of basic image processing methods while we exploit a robust pupil detection method based on the bright pupil effect for the pupil center detection. Fig. 4 illustrates the feature detection processes and outputs of the system. Further details of the eye blink detection and gaze features detection processes can be found in our previous work [12].

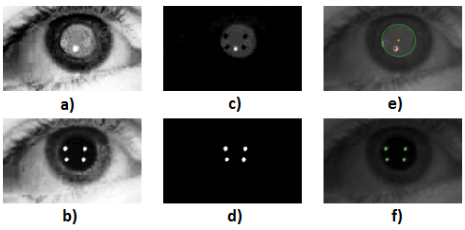

Fig. 4. Input and preprocessed images for feature detection: (a) pupil reflection and bright-eye effect, (b) corneal reflection and dark-eye effect, (c) difference image, (d) thresholded dark pupil image, (e,f) output images, detected pupil and glints.

C. Cross-Ratio Gaze Estimation with Subject-Specific Bias Correction

We employ the original CR method [10] for the estimation of the PoR. Fig. 5 shows a schematic diagram of the CR method, which is based on the only invariant of projective space, i.e. cross-ratio. In CR method, a virtual tangent plane on the cornea surface, where four glints (v_1, v_2, v_3, v_4) lie on, is assumed to exist. Hence, the polygon formed by the glints is the projection of the monitor. Another projection takes place from the corneal plane to the image plane, obtaining the glints (g_1, g_2, g_3, g_4) and the projection of the pupil center, p. 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, that can be computed by equality of the cross-ratios.

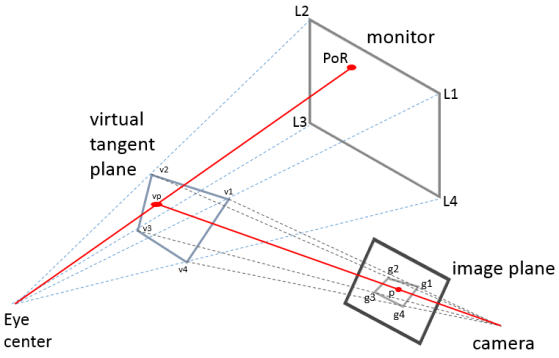

Fig. 5. The four light sources are projected onto a reflection plane. The corneal reflections are then projected onto the image plane.

CR-based gaze estimation algorithms have some assumptions that limit their performance. There are two major sources of error: i) non-coplanarity of the pupil and glints planes, and ii) the angular offset between visual and optical axes of the eye. 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. The users are asked to look at N calibration points on the monitor for K frames long. Subject-specific bias correction, $\mathcal{F}$, can be learnt by minimizing the distances between the estimated gaze positions and the corresponding calibration points on the monitor as

$$min \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P_{i,j}} - \mathcal{F}(\mathbf{Z_{i,j}})\|, \quad (1)$$

where $\mathbf{P_{i,j}}$ and $\mathbf{Z_{i,j}}$ are the target calibration points and estimated PoRs on the monitor, respectively.

In this paper, we use two different methods for modeling the estimation bias. The first method is called normalized Homography mapping (N-HOM). It estimates the bias correction by learning Homography transformations followed by mapping the glints into a unit square [8]. The second method models the error vectors using linear regression, namely regularized least-squares regression (LSR). LSR-based method has been shown to have better modeling and generalization capabilities than the homography-based methods due to reduced model parameters and relaxed constraints [12].

D. Adaptive Multi-Camera Fusion Scheme

Our hardware setup allows free head movement as well as capturing the data for both eyes although the resulting eye resolution is low, i.e. 90×60 pixels. Despite low resolution data, the system enables to get two PoRs for the same frame for each camera system. In order to output an overall PoR per frame, we propose an adaptive multi-camera fusion scheme which improves the overall estimation accuracy compared to the performance achieved by using single camera single eye data as in most of the traditional methods. The adaptive fusion scheme is independent of the gaze estimation algorithm, therefore, the current CR-based method can be replaced with any other method (e.g. interpolation-based, 3D model-based). It ideally performs a weighted averaging of individual PoRs obtained from individual camera systems as follows:

$$\mathbf{z}^* = \sum_c \sum_i \mathbf{z}_\mathbf{c}^\mathbf{e} * w_c^e \tag{2}$$

$$\sum_c \sum_i w_c^e = 1, \quad e \in \{L, R\}, c \in \{0, 1, 2\},$$

where $\mathbf{z}^*$ is the overall PoR and, w_c^R and w_c^L are the weights for the right and left eye's PoRs for c^{th} camera respectively. In case one of the PoRs can not be calculated for a given frame, then the weight of the missing PoR is set to zero. We do not report an overall PoR in case both PoRs of all the cameras are unavailable for a given frame.

In this initial work, we perform two simple weighting approaches such as adaptive and non-adaptive weighting. Firstly, we perform a non-adaptive weighting based on simple averaging, that is, assigning equal weights to any available PoR, and we achieve improved overall estimation accuracy. Secondly, we apply distance-based adaptive camera re-weighting. The idea of this method is to assign smaller weights to the cameras in which the relative head pose in the captured frame is higher. The reason is that when the user gazes at the upper left corner of the monitor, the left side camera has assumably the most frontal head pose, and therefore, it ideally outputs a more reliable estimation. In more detail, the method first roughly estimate an initial PoR, $\mathbf{z}'$, on the monitor with simple averaging. Then we calculate the distance of this rough estimation from each camera. According to the simple assumption above, the relative head pose angles increase directly proportional to the distances. Therefore, we assign weights inversely proportional to the distances from the cameras as follows:

$$\lambda_c^e = \frac{1}{\|\mathbf{z} - \ell_\mathbf{c}\|} \tag{3}$$

$$w_c^e = \frac{\lambda_c^e}{\sum_i \sum_j \lambda_i^j}, \tag{4}$$

where $\mathbf{z}$ is the estimated PoR, and $\ell_\mathbf{c}$ is the location of the c^{th} camera. We then iteratively update the overall PoR, $\mathbf{z}^*$, using the assigned weights until convergence, which often takes 2 iterations. Algorithm 1 summarizes the proposed adaptive multi-camera fusion scheme.

Algorithm 1 Adaptive Multi-Camera Fusion

Input: $\mathbf{z}_\mathbf{c}^\mathbf{e}, \ell_\mathbf{c}$

if $\mathbf{z}_\mathbf{c}^\mathbf{e} \neq null$ **then** ▷ For any available $\mathbf{z}_\mathbf{c}^\mathbf{e}$
  $\lambda_c^e \leftarrow 1$ ▷ Initialize weights equally
else
  $\lambda_c^e \leftarrow 0$
end if
$w_c^e \leftarrow \frac{\lambda_c^e}{\sum_i \sum_j \lambda_i^j}$ ▷ Normalize weights using (4)
$\mathbf{z}' \leftarrow \sum_c \sum_i \mathbf{z}_\mathbf{c}^\mathbf{e} * w_c^e$ ▷ Get initial PoR using (2)
$\mathbf{z}^* \leftarrow \mathbf{z}'$
repeat
  $\mathbf{z}_\mathbf{old}^* \leftarrow \mathbf{z}^*$
  $\lambda_c^e \leftarrow \frac{1}{\|\mathbf{z}^* - \ell_\mathbf{c}\|}$ ▷ Reweight using (3)
  $w_c^e \leftarrow \frac{\lambda_c^e}{\sum_i \sum_j \lambda_i^j}$ ▷ Normalize weights using (4)
  $\mathbf{z}^* \leftarrow \sum_c \sum_i \mathbf{z}_\mathbf{c}^\mathbf{e} * w_c^e$ ▷ Update the PoR using (2)
until $\|\mathbf{z}^* - \mathbf{z}_\mathbf{old}^*\| < \tau$
return $\mathbf{z}^*$ ▷ Return the overall PoR

Besides the weighting methods described above, the scheme allows for more sophisticated weighting methods. For instance, the weights of the PoRs can be assigned by the feature detection module considering the reliability of the detected features and the eye dominance of the user, or the head pose angles obtained by the face tracker can be used to determine the weights, or an offline learning of weights can be performed on the calibration data. The advantage of this scheme is its independence of such additional information and that it provides a significant increase in accuracy despite its simplicity.

III. EXPERIMENTS AND RESULTS

A. Evaluation Data and Protocol

We have performed a user study to evaluate the performance of the proposed system. Ten users, nine of whom had no previous experience with any gaze tracking system, participated in our experiments. We collect the ground truth data for a generic and natural HCI environment. The users were asked to look at the target stimulus points naturally the way they feel comfortable. We did not require the use of a chin rest to keep the user's head still and to keep user's

TABLE I

AVERAGE GAZE ESTIMATION ACCURACY (IN DEGREE) WITH LSR-BASED [12] AND N-HOM [8] BIAS CORRECTION METHODS.

Camera	Eye Data	No Calibration	Calibration (LSR)		Calibration (N-HOM)		Coverage (%)
			5 Points	25 Points	5 Points	25 Points	
Bottom	Left	7.09 ± 1.6	1.38 ± 0.29	1.29 ± 0.26	1.61 ± 0.44	1.26 ± 0.26	91.5
	Right	9.03 ± 2.3	1.44 ± 0.35	1.29 ± 0.26	1.74 ± 0.71	1.28 ± 0.35	94.4
	Combined	5.88 ± 2.2	1.15 ± 0.33	1.03 ± 0.23	1.47 ± 0.6	1.07 ± 0.27	96.8
Right Side	Left	9.44 ± 2.3	1.59 ± 0.51	1.49 ± 0.51	1.94 ± 0.82	1.44 ± 0.50	72.4
	Right	5.45 ± 2.2	1.52 ± 0.23	1.41 ± 0.31	2.00 ± 0.69	1.35 ± 0.29	84.1
	Combined	5.35 ± 1.6	1.31 ± 0.24	1.19 ± 0.25	1.75 ± 0.66	1.12 ± 0.24	87.9
Left Side	Left	5.18 ± 2.2	1.67 ± 0.45	1.45 ± 0.34	1.77 ± 0.43	1.38 ± 0.24	80.4
	Right	11.8 ± 2.7	1.70 ± 0.44	1.56 ± 0.36	1.84 ± 0.62	1.55 ± 0.39	73.6
	Combined	7.44 ± 2.1	1.52 ± 0.27	1.21 ± 0.25	1.88 ± 0.88	1.28 ± 0.28	87.6
Multi-Camera	**Overall**	3.97 ± 1.6	$\mathbf{0.86 \pm 0.18}$	$\mathbf{0.79 \pm 0.16}$	1.02 ± 0.32	0.83 ± 0.17	**97.3**

one of the eyes within the field of view of the camera in order to capture high resolution eye data. Head pose variation statistics obtained from the bottom camera recordings using the pose estimation method described in [18] are illustrated in Table II.

TABLE II

HEAD POSE VARIATION STATISTICS (IN DEGREE) ON THE COLLECTED EXPERIMENTAL DATA OF THE BOTTOM CAMERA.

	Yaw Angle		Pitch Angle	
	Cal	Test	Cal	Test
Min	-19.11	-11.18	-18.51	-19.5
Max	23.06	16.52	7.95	3.88
Mean	2.37	2.09	-6.92	-7.23
Std	4.28	3.22	2.78	1.79

We capture calibration and test data in two separate sessions for each camera simultaneously. For the capture of the calibration data, we ask users to gaze at 25 uniformly distributed target points on the monitor. The target stimulus points were displayed in a left to right and top to bottom sequence in a 5×5 grid. For the capture of the test data, we aim to prevent overfitting on the points, and to create a natural and realistic test condition. For this purpose, we introduce a new evaluation scheme where the test points are independent from calibration points. We ask users to gaze at 18 target stimulus points in a 3×3 grid covering the whole monitor. The positions of the target stimulus points in a region were randomly determined. We ensure that two stimulus points are shown in each region in order to cover the whole monitor. The display order of the regions and the points is also randomly determined.

We display each target point for 100 frames (3.33 seconds), and capture the data of both eyes during this period. To keep the attention of the user on the target stimulus points, we varied the size of the circular target from an initial radius of 30 pixels to a final radius of 20 pixels. For testing, we discard the first 20 frames of each target point and keep the latter 80 frames for the evaluation in order to avoid saccadic gaze movement at the beginning of each point display. We report our eye tracker's performance as 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 since it is invariant to the distance between the user and the monitor.

B. Results

For our evaluation, separately for each camera, we first run the face tracker on the captured data to extract eye regions. Due to the limited resolution of the eye region, the size of the extracted eye region is around 90×60 pixels and the size of the polygon formed by the glints is around 12×7 pixels. On the detected features we apply CR-based gaze estimation to calculate the initial PoR. In the calibration process, which is performed separately for each eye and camera, we model the subject-specific error vectors according to (1). In the test process, we apply the learnt models to correct the initial PoRs estimated on the test data.

The results achieved by the proposed system on the test data are shown in Table I. We report the mean and the standard deviation of the average estimation accuracy error in degrees of visual angle over all subjects with respect to different calibration methods and number of calibration points. We list the results obtained from single camera setups with separate eye data as well as the overall data from multiple cameras with the proposed adaptive fusion scheme. The rightmost column, **Coverage**, shows the percentage of frames in which we are able to output a PoR for the given eye data.

The results demonstrate that the estimation error reduces with increasing number of calibration points used. In addition, it validates the effectiveness of the multi-camera fusion scheme. Firstly, it improves the estimation accuracy by about 38% and 25% over the best performing single eye setup (bottom camera left eye) and the best performing single camera setup (bottom camera combined eyes) respectively. Secondly, it increases the coverage, the working volume, of the system compared to single camera systems. As shown in Table I, the system outputs a PoR for 97.3% of all frames while eye blink is detected for 2.41% of all the frames. Hence, the system could not output a PoR only 0.29% of all the frames due to missing features. On the other hand, the coverage drops when single camera setup is used (i.e., the bottom camera) because the data obtained from a single camera system may not be sufficient to reliably calculate a PoR for some of the test points, especially those positioned

close to the right or left borders of the monitor. For those points, it is highly likely that the additional left and right side cameras provide more accurate PoRs.

As there is no previous multi-camera fusion work for the purpose of precise gaze estimation on the monitor, we can not directly compare our results with the others. However, we conduct experiments with different camera setups to illustrate the effectiveness of the proposed multi-camera setup over single camera setups under the same conditions. Besides, we apply two competent calibration techniques to demonstrate that the efficacy of our system is independent of the calibration techniques used. Fig. 6 shows the mean and standard deviation of the estimation accuracy error with respect to different number of calibration points.

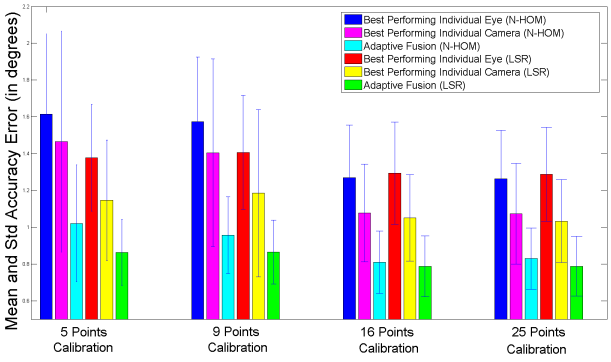

Fig. 6. Comparison of the multi-camera system with the best performing single camera system.

As depicted from Fig. 6, the multi-camera system achieve significantly better performance in any calibration method-number of calibration points configuration. In addition, we can state that the LSR-based calibration method is superior to the homography-based method, especially when fewer number of calibration points are used.

1) Effectiveness of Adaptive Multi-Camera Fusion: In this work, we investigate two simple weighting approaches, namely, simple averaging and distance-based camera re-weighting. Fig. 7 and Fig. 8 show the comparison of different weighting methods in different calibration configurations.

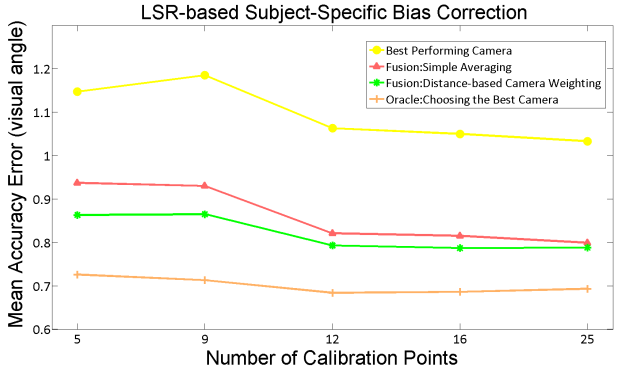

Fig. 7. Comparison of different weighting approaches in case LSR is used for the calibration.

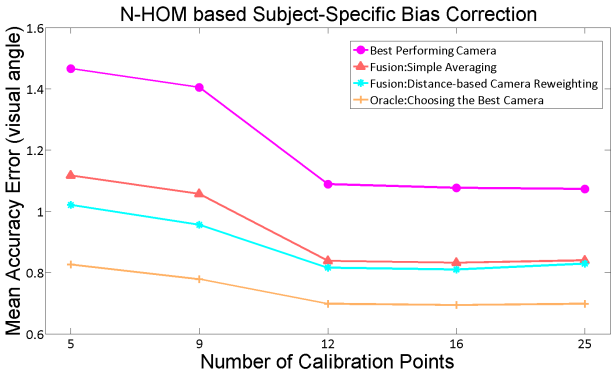

Fig. 8. Comparison of different weighting approaches in case N-HOM is used for the calibration

The results illustrated in the two figures show that a remarkable improvement can be obtained even using a non-adaptive weighting in comparison to the best performing single camera system. When the proposed distance-based adaptive camera re-weighting is applied, the estimation error is decreased further. The results support the simple assumption of the gaze-pose correlation, and demonstrate the effectiveness of the proposed weighting scheme. In order to notice the upper limits of the multi-camera system through an optimal weighting system, we also display the results as if there is an oracle knowing the best performing camera for each frame. The oracle results imply a further performance enhancement can be achieved using more complicated weighting methods as mentioned in Section II-D.

As shown in Fig. 7 and 8, the estimation error reduces with increasing number of points used for calibration. However, as targeted in this study, a less tedious and user-friendly system should involve as little effort as possible for the subject-specific calibration. The simplest way to achieve this is to minimize the number of calibration points, without sacrificing too much the estimation accuracy. Our proposed system can reach a reasonable estimation accuracy of 0.86 ± 0.18 degree using LSR-based calibration with only 5 calibration points. Hence, the system shows comparable performance to the 3D model-based systems [13], [14] whose reported accuracies are $<1^\circ$ but they require more complex system setups such as camera and geometric calibration.

Moreover, the proposed methodology brings another advantage in addition to the enhanced estimation accuracy and coverage. The computational complexity of the proposed system is less than the 3D model-based methods and is suitable for real-time gaze tracking. In fact, the most computationally expensive process of the proposed system is the face detection/tracking. The PoR estimation using CR, the subject-specific bias correction and the adaptive multi-camera fusion processes require negligible computational effort. For instance, it takes <8 ms on a PC with Intel i7 3.2GHz processor. Therefore, the real-time gaze tracking system can easily be obtained with a real-time face tracker.

IV. CONCLUSIONS

In this paper, we present a multi-camera gaze estimation system which accurately works under natural head movements. Our real-time system requires no effort in terms of the camera and geometric calibration as it is based on multiple independent camera systems in which the gaze estimation in each camera system relies on simple CR geometry. In addition, the system does not require high-resolution eye data as opposed to most of previous work. Operating with low-resolution data enables the system to output PoRs from each eye simultaneously. In order to jointly estimate an overall PoR, a novel adaptive multi-camera fusion scheme is suggested. The effectiveness of the proposed system has been validated with a new evaluation scheme, where the test points are not chosen among the calibration points. The results indicate that the system's performance, even with very few calibration data (5 points), is competitive with more complex systems presented in the literature. Therefore, 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 develop more sophisticated weighting approaches for the adaptive fusion scheme, and we also plan to investigate the system's robustness against extreme head/body movements for non-generic HCI applications.

V. ACKNOWLEDGMENTS

This project is supported by the Swiss Commission for Technology and Innovation (CTI) under grant number 13594.1 PFFLR-ES. The authors would like to thank Yves Moser from Logitech for his valuable contributions in the user experiments.

REFERENCES

[1] D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. In *Conf. in Computer Vision Pattern Recognition (CVPR)*, 2003.

[2] S.-W. Shih and J. Liu. A novel approach to 3D gaze tracking using stereo cameras. In *Trans. Systems, Man, and Cybernetics*, vol. 34, no. 1, pp. 234-245, Feb. 2004.

[3] T. Ohno and N. Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *Symp. Eye Tracking Research and Applications (ETRA)*, 2004.

[4] Z. Zhu, Q. Ji and K.P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression In *Conf. on Pattern Recognition (ICPR)*, 2006.

[5] F. Coutinho and C. Morimoto. Improving head movement tolerance of cross-ratio based eye trackers. In *Int. Journal of Computer Vision (IJCV)*, 101(3):459–481, 2013.

[6] J.-B. Huang, Q. Cai, Z. Liu, N. Ahuja, and Z. Zhang. Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *Symp. Eye Tracking Research and Applications (ETRA)*, 2014.

[7] J. J. Kang, M. Eizenman, E. D. Guestrin, and E. Eizenman. Investigation of the cross-ratios method for point-of-gaze estimation. In *Trans. on Biomedical Engineering*, 55(9):2293–302, 2008.

[8] D. W. Hansen, J. S. Agustin, and A. Villanueva. Homography normalization for robust gaze estimation in uncalibrated setups. In *Symp. Eye Tracking Research and Applications (ETRA)*, 2010.

[9] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. In *Computer Vision and Image Understanding (CVIU)*, 98(1):25–51, 2005.

[10] D. H. Yoo, J. H. Kim, B. R. Lee, and M. J. Chung. Non-contact eye gaze tracking system by mapping of corneal reflections. In *Conf. Automatic Face & Gesture Recognition (AFGR)*, 2002.

[11] Z. Zhang and Q. Cai. Improving cross-ratio based eye tracking techniques by leveraging the binocular fixation constraint. In *Symp. Eye Tracking Research and Applications (ETRA)*, 2014.

[12] N. M. Arar, H. Gao and J. P. Thiran. Towards Convenient Calibration for Cross-Ratio based Gaze Estimation. In *Conf. on Applications of Computer Vision (WACV)*, 2015.

[13] D. W. Hansen and Q. Ji. In the eye of the beholder: a survey of models for eyes and gaze. In *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(3):478–500, 2010.

[14] C.H. Morimoto and M.R.M. Mimica. Eye gaze tracking techniques for interactive applications. In *Computer Vision and Image Understanding (CVIU)*, vol. 98, no.1, pp.4-24, 2005.

[15] A. Utsumi, K. Okamoto, N. Hagita, and K. Takahashi. Gaze tracking in wide area using multiple camera observations. In *Symp. Eye Tracking Research and Applications (ETRA)*, 2012.

[16] R. Ruddaraju, A. Haro, K. Nagel, Q. Tran, I. Essa, G. Abowd, and E. Mynatt. Perceptual user interfaces using vision-based eye tracking. In *Conf. on Multimodal Interfaces (ICMI)*, 2003.

[17] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Conf. in Computer Vision Pattern Recognition (CVPR)*, 2013.

[18] S.C. Chen, C.H. Wu, S.Y. Lin and Y.P. Hung. 2D face alignment and pose estimation based on 3D facial models. In *Conf. on Multimedia & Expo (ICME)*, 2012.