A Regression-based User Calibration Framework for Real-time Gaze Estimation

Nuri Murat Arar, *Student Member, IEEE,* Hua Gao and Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*—Eye movements play a very significant role in human computer interaction (HCI) as they are natural and fast, and contain important cues for human cognitive state and visual attention. Over the last two decades, many techniques have been proposed to accurately estimate the gaze. Among these, video-based remote eye trackers have attracted much interest since they enable non-intrusive gaze estimation. To achieve high estimation accuracies for remote systems, user calibration is inevitable in order to compensate for the estimation bias caused by person-specific eye parameters. Although several explicit and implicit user calibration methods have been proposed to ease the calibration burden, the procedure is still cumbersome and needs further improvement. In this paper, we present a comprehensive analysis of regression-based user calibration techniques. We propose a novel weighted least squares regression-based user calibration method together with a real-time cross-ratio based gaze estimation framework. The proposed system enables to obtain high estimation accuracy with minimum user effort which leads to user-friendly HCI applications. Experimental results conducted on both simulations and user experiments show that our framework achieves a significant performance improvement over the state-of-the-art user calibration methods when only a few points are available for the calibration.**

I. Introduction

Gaze is considered as an essential modality for HCI because they contain crucial cues indicating visual attention, cognitive processes, emotional states and interpersonal interactions [1]. Besides, they are natural and fast which make them highly suitable to interact with computer vision systems. Consequently, robust estimation of gaze with high precision and accuracy is of great interest for the development of many interesting diagnostic and HCI applications such as human attention and cognitive state analysis, usability testing, market research, disabled aids, gaze-based interactive user interfaces, mouse cursor positioning, page scrolling, gaze-based map navigation, gaze-based gaming and many other gaze-controlled computer functionalities. Recently, gaze estimation systems with a variety of applications have been introduced, and promising advancements have been made by both industry and scientific community [2]–[6]. However, there is still room for further research so as to improve the robustness and convenience of the systems.

Gaze-based interfaces aim to accurately map user gaze to the screen coordinates. For interactive applications, mostly remote gaze trackers are preferred even though their accuracies are lower compared to head/eye mounted gaze trackers. The main reason is that head mounted gaze trackers provide an unnatural and invasive experience for users due to their intrusiveness. Therefore, our focus is on remote video based gaze tracking where users' eye are non-intrusively captured by a single or multiple cameras and the gaze is estimated through image processing and computer vision methods. Remote video based gaze tracking methods can be classified mainly into two groups as described in a recent survey [7]: appearance-based methods [8]–[12] and model-based methods [13]–[19]. 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). On the other hand, appearance-based methods simply map image features to gaze points. Their system and hardware requirements tend to be simpler than model-based methods. They simply require an ordinary camera. Neither camera nor geometric calibration is necessary. However, they are restricted to particular applications due to their limitations regarding the estimation accuracy and head movements. Although there are a few recent works (e.g., [11], [12]) that put an effort on improving the accuracy and, head pose and movement robustness, further advancements are necessary for them to be utilized for the precise eye tracking applications. On the contrary, model-based methods offer greater freedom of movement and high estimation accuracy ($\leq 1^\circ$). However, their biggest disadvantage is that they require more complex system setups such that camera and geometric calibrations are required to obtain 3D information. As they are based on accurate 3D modeling of user eye, user calibration is very crucial to estimate individual-specific eye parameters. Recently, there have been interesting calibration efforts for the purpose of more convenient and natural HCI. For instance, Sun et al. [17] propose a real-time gaze estimation system with online calibration. Instead of displaying a fixed number of calibration points, they update the eye parameters after each new point. The calibration process is completed as soon as the updates of eye parameters reach convergence. They reported that the system adapts to a new user by online calibration within 3 minutes and achieves an accuracy error $\sim 2^\circ$. Chen and Ji [19] suggests a user-friendly implicit calibration in which they estimate the probability distributions of eye parameters and gaze. They display several images with salient objects to a user and the method adapts to the user over time. They report an estimation accuracy error of $<3^\circ$.

In addition to appearance-based and 3D model-based methods, there also exists another group of methods which are

N. M. Arar, H. Gao and J.-P. Thiran are with the Signal Processing Lab (LTS5), École Polytechnique Fédérale de Lausanne, Switzerland, e-mail:(see http://people.epfl.ch/name.surname).

called as cross-ratio (CR)-based methods. They only span a very small portion of studies in gaze estimation research. Contrarily, CR-based methods [20]–[30] share advantages from both appearance and model-based methods. For instance, they do not require any camera or geometry calibration and they allow free head motion. Unfortunately, the performance of CR-based methods might be limited in accuracy and robustness due to the simplifications assumed. There are two major sources of estimation bias in CR-based methods as described in [31]. First, the model assumes that the pupil center and the corneal reflections (glints) lie on the same plane. They are, in fact, not coplanar because the cornea has a spherical surface. Second, the model computes the PoR on the basis of eye ball's optical axis rather than the visual axis, the real line of sight. In order to compensate for the estimation bias, a subject-specific bias correction, in other words, a user calibration is required. In the literature, several efforts have been made in order to enhance the accuracy and robustness of CR-based gaze estimation systems. In the original CR-based gaze estimation method introduced by Yoo et al. [20], there was not any subject-specific calibration. Later, they refined their method by several enhancements in feature detection and they introduced a calibration to compensate for cornea's non-coplanarity using an additional LED illuminator in their hardware setup [21]. Even though the calibration did not consider the correction for the axes difference, it significantly improved the estimation accuracy. In a similar approach, Coutinho and Morimoto [22] proposed a method to compensate for the axes difference for the first time. Yet, their system required a fifth LED in the hardware setup similar to [21]. Later, homography-based correction was introduced by Kang et al. [23]. They simplified the error correction using a similar calibration procedure but eliminated the need for the fifth LED. It outperformed all previous methods with a simpler hardware setup. Similarly, Hansen et al. [24] proposed a normalized homography mapping to further improve the robustness against perspective distortions. They also proposed to apply a non-linear regression, i.e. Gaussian process regression (GPR), followed by the normalized homography mapping so as to further improve the estimation accuracy.

Homography-based calibration is widely accepted by the community as the state-of-the-art method. When users gaze their monitor under normal conditions, most of the time no abrupt change is observed in their head pose or location. For such HCI scenarios homography-based calibration methods work successfully when there is a sufficient number of calibration points. On the other hand, a few modified homography-based approaches (e.g., [25], [27]) have been recently proposed in order to explicitly bring robustness against large head movements for non-generic HCI scenarios. Moreover, Zhang and Cai [26] introduced a binocular fixation constraint to jointly estimate the CR homography matrix. Contrary to all previous work, they utilized information from both eyes to improve the calibration. One drawback of their system is that the features from both eyes must be available to compute a gaze output, which constrains the estimation coverage (availability) of the system due to the limited head pose. In addition, we note that most of the previous work required high resolution eye data to operate. They used either a mechanical pan-tilt unit (e.g., [21]) or a chin rest (e.g., [16], [18], [23], [25]–[27]) to keep users' eyes within the field of view (FOV) of the camera, and they captured images with zoomed lenses (large focal length) to obtain high-resolution eye data. Use of a chin rest also enabled to keep users' heads fixed during the calibration and testing. However, such restrictions would not be convenient and practical for users in real world HCI applications. In our previous work [28], towards a more convenient and natural calibration, we introduced a linear regression-based bias correction which has a better generalization capability with fewer number of calibration points under certain head movements compared to the state-of-the-art methods.

In this paper, as an extension of our previous work, we further perform an extensive investigation of different regression techniques to compensate for the estimation bias in CR-based gaze estimation. To this extent we introduce several weighted and iterative regression-based user calibration methods in addition to the classical regression methods so as to achieve enhanced estimation accuracy and robustness. We conduct simulations and user experiments to analyze the pros and cons of the investigated methods under different experimental configurations. Consequently, we propose a novel weighted regression-based calibration framework, which outperforms the state-of-the-art approaches as well as other investigated methods, especially when few points are used for the calibration.

In addition, we present a complete real-time gaze estimation system using the proposed calibration framework. Our proposed system consists of a rather simple setup while still obtaining high estimation accuracy. Unlike most of the previous efforts in the literature, the system does not require high resolution eye data to reach high estimation accuracy. Instead, we capture video frames of the whole face with visible but lower resolution eye pair. This way the system enables to output PoRs from each eye simultaneously. The handicap of low resolution data is compensated for with a novel adaptive fusion scheme that allows outputting an overall PoR using the data of either or both eyes, rather than using a single chosen eye as often performed in the previous literature. The proposed scheme improves not only the mean accuracy, but also the system's estimation coverage (availability) compared to operating with a single eye.

Furthermore, our proof-of-concept effort targets a generic real-world HCI environment and focuses on a user-friendly experience. To this extent, we collected ground truth data under natural and realistic conditions. Contrary to most of the previous work, the data was captured in two independent sessions for user calibration and testing. A new evaluation scheme is introduced such that test points were randomly generated over the whole screen rather than using the same stimulus points for both calibration and testing in order to avoid reporting false test results due to overfitting on the calibration point locations. In addition, no chin rest was used during the captures and users were not particularly asked to move or standstill their heads with respect to the monitor.

The main contributions of this paper can be summarized as follows:

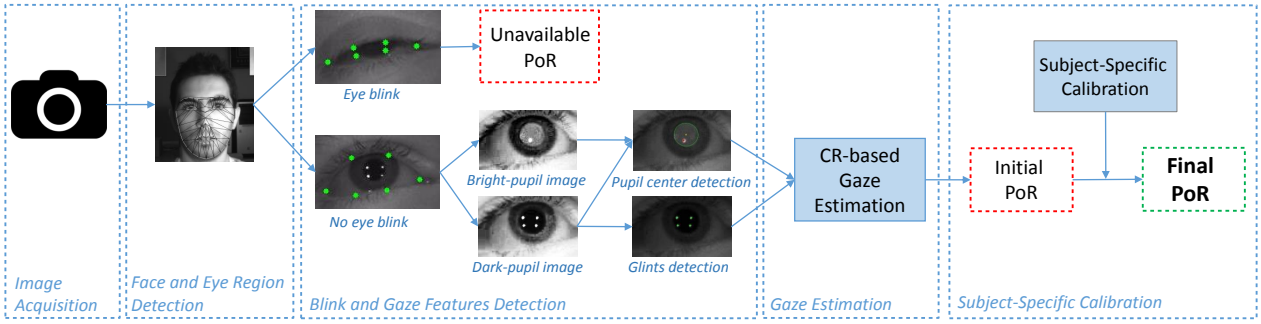


Fig. 1. System overview (single eye). This figure is best viewed in color.

- We propose a novel subject-specific gaze calibration method based on a weighted linear regression technique.
- We introduce a complete gaze estimation system which operates in real-time and enables user-friendly HCI applications through a simple setup and convenient calibration.
- We propose an adaptive fusion scheme which makes use of both eyes to output an improved overall PoR as well as an enhanced estimation availability over the monitor.
- We introduce a more natural and realistic data capture setting together with a more reliable evaluation scheme compared to the previous work.

The rest of the paper is organized as follows: Section II describes the hardware setup and details of the proposed method and system. Section III explains the procedure and scenarios used for the simulations and user experiments as well as the real-time implementation. Section IV contains a discussion on robustness and a comparison of eye tracking techniques. Lastly, conclusions are given in Section V.

II. Proposed System

The proposed gaze estimation system consists of six main processes: 1) image acquisition, 2) face and eye detection/tracking, 3) blink and gaze features detection, 4) gaze estimation, 5) subject-specific calibration correction and 6) adaptive fusion. An overview of the proposed gaze estimation scheme for a single eye, therefore, the first five processes, is illustrated in Fig. 1. Gaze output of each eye is then fed into the adaptive fusion process to output an overall PoR. The details of each process are explained in the following sections.

A. Hardware Setup

Our real-time eye tracking system consists of one PointGrey Flea3 monochrome camera for video capturing, 5 groups of near-infrared (NIR) LEDs for the illumination and a controller unit for the synchronization. The camera has a resolution of 1280×1024, and a 12 mm lens is used. The camera is located below the monitor and slightly closer to the user. In order to create the glints, 4 groups of NIR LEDs with 850 nm wavelength are placed on the corners of the monitor. A band-pass filter around 850 nm is mounted in front of the lens in order to get rid of the ambient light in other wavelengths. The fifth group of LEDs is placed as ring around the lens of the camera to create the bright pupil effect. A micro-controller is

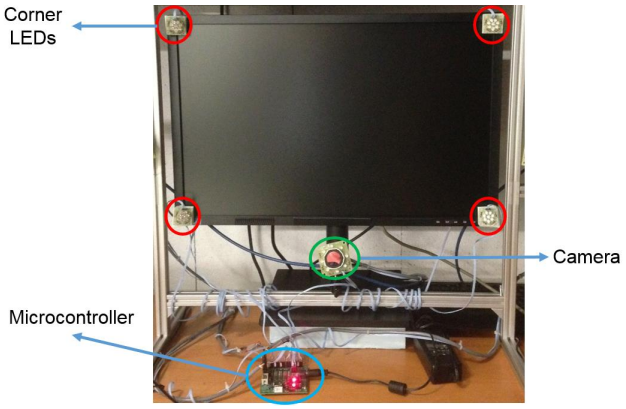


Fig. 2. Hardware setup.

programmed in a way 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 considering the user eye safety. In the current setup, the user is located approximately 70 cm away from a 24-inch monitor with a resolution of 1920×1200. The head is not fixed, therefore users are allowed to perform natural head movements. Fig. 2 shows the equipment employed.

B. Gaze Features Detection

Our system starts with eye localization where existence of eyes is determined. In order to localize and track the eyes we utilize a robust non-rigid face tracker based on supervised decent method (SDM) [32]. SDM method assumes that an accurate final face shape with 66 landmarks can be estimated with a cascade of regression models given an initial shape. Viola & Jones face detector [33] is used to initialize the shape. The tracker first fits the mean shape in the initial frame and continues the fitting in the succeeding frames. Once the shape is fitted accurately, we extract eye regions by considering the landmarks representing eyes. We do not perform any registration or scaling on the extracted eye regions to ensure any particular eye region resolution. On the extracted eye regions, first we detect whether there is any eye blink or not. If there is no eye blink, we then detect image features for the gaze estimation. The image features include the pupil center and corneal reflections of NIR LEDs, i.e. glints.

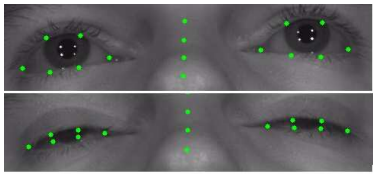


Fig. 3. The positioning of facial landmarks in case of **(top)** no eye blink; **(bottom)** an eye closure during a blink.

1) Eye Blink Detection: For the eye blink detection, we check the positioning of the landmarks around the eyes. We measure vertical opening (height) of both eyes relative to the eye width. As illustrated in Fig. 3, if the average of the ratio of eye height to eye width for both eyes is significantly lower (< 0.15) than the open eye form (~ 0.5), we determine that a natural eye blink occurs. Once an eye blink is detected, the system skips the following processes and no output is generated. Since the average eye blink is completed 100 to 200 milliseconds after the peak closure of eyelids, we do not output any PoR for the corresponding number of frames once an eye blink is detected. On the other hand, if the system misses an eye blink, the system follows with the feature detection, and undoubtedly no features are detected as the pupil area is not visible due to the blink. Hence, the performance of the system does not heavily depend on the blink detection process.

2) Glints and Pupil Center Detection: We employ simple image processing algorithms to precisely localize the glints. Firstly, histogram equalization is performed followed by thresholding on the input image which results in a binary image. We use spatial adaptive thresholding in order to take into account spatial variations in illumination. Instead of tuning a global threshold value, adaptive thresholding calculates the threshold for small regions of the image. So, different thresholds are applied for different regions of the same image and it gives more stable results under varying illumination. We use OpenCV's adaptive thresholding function. The parameters *block size* and *C* (i.e., a constant subtracted from the mean) are set to 10% of the original image width and -100, respectively. The actual threshold value, *T(x,y)*, is a mean of the *block size* $\times$ *block size* neighborhood of (x, y) minus *C*. Following the adaptive thresholding, the resulting binary image is processed by morphology operations to get rid of the small blobs caused by noise. In the resulting binary image, we expect to find four blobs which should form a trapezium since they emerge by the reflections of four NIR LEDs located on the corners of the computer monitor (Fig. 4.d). Hence, we get the candidate glints by performing connected component analysis. If there are four or more candidate glints remaining, we consider the shapes formed by any four-glints combination. The set of candidates whose convex hull has the highest match with a template shape representing the screen are considered as the final glints.

For the pupil center detection, a more sophisticated technique is required since the intensity of the pupil is more similar to its surrounding pixels. For this purpose, we use the robust pupil detection method suggested by [34]. The method is based on bright-pupil effect which is obtained when an NIR LED is located in the optical axis of a camera as shown in Fig. 4.a. In a similar approach, to robustly detect the pupil, we use two images: one is taken when the corner LEDs on the monitor are turned on and the LEDs on the camera axis are turned off, the other is taken when the monitor LEDs are off and the camera LEDs are on. If these images are obtained from the camera in a very short interval, e.g. consecutive frames, then the intensity difference of the pupil region in two images is large and that of the region outside of the pupil is very small. Therefore, the difference image has high intensities in the pupil region. The pupil region can be extracted by a segmentation method that is very similar to glint detection, and the center of gravity is considered as the pupil center. Fig. 4 illustrates the feature detection processes and outputs of the system.

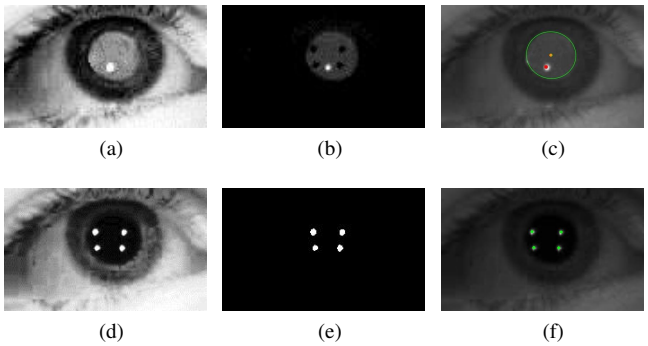


Fig. 4. Sample images from feature detection process: **(a)** bright-pupil effect; **(d)** corneal reflections (glints); **(b)** difference image; **(e)** thresholded dark-pupil image; **(c,f)**) output images of the detected pupil and glints.

C. Cross-Ratio Gaze Estimation

We employ the original CR method [20] for the estimation of the PoR. The CR method is a geometry based gaze estimation technique for uncalibrated gaze estimation setups. It is based on the cross-ratio property, the only invariant of projective space. Fig. 5 illustrates the geometric setup used by the CR method.

In CR method, a virtual tangent plane on the cornea surface, where the 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

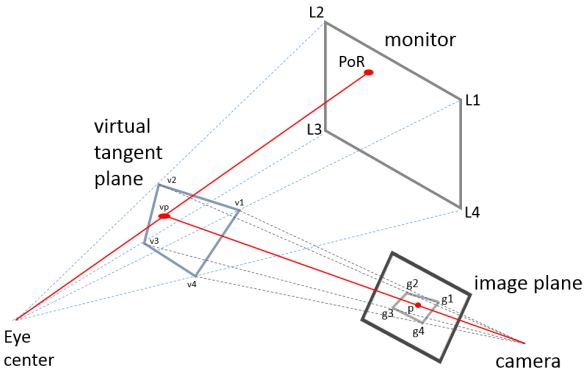


Fig. 5. Geometric setup of CR-based gaze estimation systems.

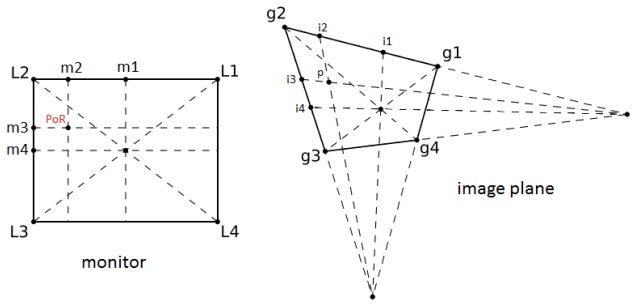


Fig. 6. Cross-ratio of image and screen points.

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.

The PoR on the monitor can be computed by the equality of the cross-ratios on the monitor plane, $CR_{monitor}$ and the camera image plane, CR_{image} (Fig. 6). The CR is defined for four collinear points as:

$$CR(p_1, p_2, p_3, p_4) = \frac{|p_1p_2||p_3p_4|}{|p_1p_3||p_2p_4|} \tag{1}$$

where

$$|p_ip_j| = det \begin{bmatrix} p_i^x & p_j^x \\ p_i^y & p_j^y \end{bmatrix} \tag{2}$$

The cross-ratio on the x axis of the monitor plane can be computed as follows:

$$CR^x_{monitor}(L_1, m_1, m_2, L_2) = \frac{(w - \frac{w}{2})\ \ \hat{p}_x}{(w - \hat{p}_x)\ \ \frac{w}{2}} = \frac{\hat{p}_x}{w - \hat{p}_x} \tag{3}$$

where w is the width of the monitor and $\hat{p}_x$ is the x coordinate of the estimated gaze point p.

The corresponding cross-ratio of the image plane is:

$$CR^x_{image}(g_1, i_1, i_2, g_2) = \frac{|g_1i_1||i_2g_2|}{|g_1i_2||i_1g_2|} \tag{4}$$

Since the cross-ratios of both configurations are equal, the estimated x coordinate of the PoR, $\hat{p}_x$, can be calculated as follows:

$$\hat{p}_x = \frac{w}{1 + CR^x_{image}} \tag{5}$$

A similar derivation on the y axis gives the estimated y coordinate of the PoR, $\hat{p}_y$, as follows:

$$\hat{p}_y = \frac{h \cdot CR^y_{image}}{1 + CR^y_{image}} \tag{6}$$

where h is the height of the monitor.

D. Subject-Specific Calibration

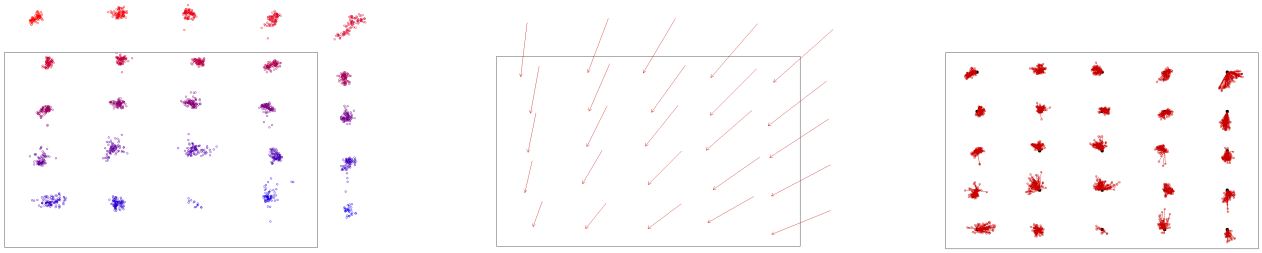
Subject-specific calibration is crucial for remote eye tracking systems to compansate for the estimation bias caused by subject-specific eye parameters. In CR-based gaze estimation, the estimation bias is largely introduced by the simplifying assumptions which are not valid in practice. Kang et al. [31] identified two major sources of estimation bias: i) non-coplanarity of the pupil and glints planes, and ii) the angular offset between visual and optical axes of the eye. Firstly, CR methods assume that the glints and pupil center lie on the same plane. However, there is no guarantee that they will be coplanar since the cornea has a curved surface. Secondly, it computes the PoR by canceling out the effect of the angular offset difference between the optical and visual axis of the eye ball. The real line of sight is based on the visual axis. However, the algorithm bases the optical axis for the PoR estimation. Since the cornea curvature and the angular offset are subject-specific parameters, a calibration needs to be performed to compensate for the estimation bias. The calibration procedure 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 learned by minimizing the distances between the estimated gaze positions and the corresponding calibration points on the monitor as follows:

$$min \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P}_{i,j} - \mathcal{F}(\mathbf{Z}_{i,j})\|, \tag{7}$$

where $\mathbf{Z}_{i,j}$ and $\mathbf{P}_{i,j}$ are the estimated PoRs on the monitor and the corresponding target calibration points, respectively.

As mentioned in Section I, many techniques have been proposed to compensate for the estimation bias. There is no doubt that the calibration performance increases when the amount of calibration data increases. However, augmenting the amount of data by increasing the number of calibration points could be tedious and thus harms the user experience. Moreover, CR-based methods are highly sensitive to feature detection by their nature. For this reason, previous methods preferred to use high resolution eye data [21], [22], [25] and fixed head position [25]–[27]. As opposed to many of the previous work, our system prefers to operate with low resolution eye data in order to capture and process both eyes simultaneously due to practical reasons such as higher operability under real-world head movement conditions and lower computational burden to achieve real-time gaze tracking. Hence, our system is even more vulnerable to feature detection noise. These factors motivated us to further investigate different methods to model the estimation bias more robustly against outliers and noise when using few number of calibration points.

In this paper, we emphasize regression-based bias correction methods as they implicitly model the estimation bias with high accuracy. These methods include regularized least squares regressions (LSR) such as Ridge and Lasso, partial least squares regression (PLSR) and Gaussian process regression (GPR). Moreover, we propose weighted LSR methods (WLSR) to improve the calibration quality. We introduce two different weighting schemes of the calibration data: 1) a weighting of calibration data points with respect to the point cluster variance and 2) a weighting of individual samples with respect to their within point cluster variance. We perform a comprehensive and detailed analysis of all the methods to see the advantages and disadvantages of each, and we compare with the traditional



(a) (b) (c)

Fig. 7. Impact of user calibration **(a)** raw gaze data; **(b)** vector fields indicating the effect of calibration; **(c)** calibrated gaze data.

homography-based calibration methods. The details of the investigated methods are presented in the following subsections.

1) L2-Regularized LSR (Ridge): Traditional homography-based calibration methods rely on a perspective homography transformation with 8 degrees of freedom (DOF). In total 9 parameters need to be recovered and no regularization is considered. When the image region in which the homography is computed is small or the image has been acquired with a large focal length, an affine homography is a more appropriate model of displacements with lower DOF [35]. Thus, we consider an affine transform for the subject-specific estimation bias modeling. Since an affine transform has less model parameters (i.e., 6 parameters) than a homography transform, the problem is more determined when fewer points are used for the calibration. Consequently, a better generalization (less overfit) is expected on unseen test points due to less relaxed constraints.

To this effect, we firstly employ a L2-regularized least squares regression (also known as Ridge regression) [36] to find an affine transform with 6 DOF. The transform $\boldsymbol{\beta}$ is defined with a 3×2 matrix, where the first column corresponds to the offset parameters. The input data $\mathbf{X}$ is a stack of the estimated PoR coordinates:

$$\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}$$

The corresponding output data $\mathbf{Y}$ stores the target coordinates for calibration. The cost function $E(\boldsymbol{\beta})$ for the regularized least squares problem is defined as:

$$E(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^T\mathbf{X} - \mathbf{Y}\|^2 + \lambda\|\boldsymbol{\beta}\|_F^2 \tag{8}$$

where λ is the regularization shrinkage and $\|\|_F$ stands for the Frobenius norm. A closed form solution can be found by setting the first order derivative of the cost function $E(\boldsymbol{\beta})$ to zero, and we obtain:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T \tag{9}$$

Using the learned model $\widehat{\boldsymbol{\beta}}$, we can predict a calibrated coordinate giving an input PoR, $\mathbf{x}$:

$$\widehat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}^T\mathbf{x} \tag{10}$$

In addition, we apply a kernelized Ridge regression for calibration, based on the assumption that the error we need to compensate can be nonlinear due to perspective projection. The prediction using a kernel Ridge regression becomes:

$$\|\boldsymbol{\beta}^T\Phi(\mathbf{X}) - \mathbf{Y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{11}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}\mathbf{Y}^T \tag{12}$$

$$\begin{aligned} \widehat{\boldsymbol{y}} &= \widehat{\boldsymbol{\beta}}^T\Phi(\mathbf{x}) \\ &= \mathbf{Y}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}^T\Phi(\mathbf{x}) \\ &= \mathbf{Y}(\mathbf{K} + \lambda\mathbf{I})^{-1}\boldsymbol{\kappa}(\mathbf{x}) \end{aligned} \tag{13}$$

where $\boldsymbol{\kappa}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x})]^T$, and $\mathbf{K}$ denotes the kernel matrix. In this study, we use a second order polynomial kernel.

2) L1-Regularized LSR (Lasso): Least Absolute Shrinkage and Selection Operator (Lasso) regression is another form of regularized linear regression where the regularization is based on L1 norm [37]. Therefore, it involves penalizing the absolute size of the regression coefficients. The cost function $E(\beta)$ for Lasso is defined as:

$$E(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^T\mathbf{X} - \mathbf{Y}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{14}$$

where λ is the regularization shrinkage, and a large enough λ may set some coefficients to zero.

The regularization can also be interpreted as prior in a maximum a posteriori estimation method. Under this interpretation, the Ridge and the Lasso make different assumptions to relate input and output data on the class of linear transformation. In the Ridge, the coefficients of the linear transformation are normal distributed whereas in the Lasso they are Laplace distributed. Hence, in the Lasso, it is easier for the coefficients to be zero and therefore, it is easier to eliminate some of the input variables which do not contribute to the output.

3) Partial Least Squares Regression (PLSR): PLSR is a method that bears some relation to principal components regression (PCR) in which the regression analysis is based on principal component analysis (PCA) by finding hyperplanes of minimum variance between the response and independent variables. PLSR instead finds a linear regression model by projecting the predicted variables and the observable variables to a new latent space in such a way that covariance between projected input and output vectors is maximized [38]. It is based on partial least squares which is used to find the fundamental relations between two matrices ($\mathbf{X}$ and $\mathbf{P}$), i.e. a

latent variable approach to modeling the covariance structures in these two spaces. The PLS formulation is as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{15}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{16}$$

where $\mathbf{T}$, $\mathbf{U}$ are the projections of $\mathbf{X}$ and $\mathbf{Y}$ in the latent space, respectively, and $\mathbf{P}$, $\mathbf{Q}$ are orthogonal loading matrices, and $\mathbf{E}$, $\mathbf{F}$ are the error terms which are assumed to be independent and identically distributed random normal variables. The decompositions of $\mathbf{X}$ and $\mathbf{Y}$ are made so as to maximize the squares of covariance between $\mathbf{T}$ and $\mathbf{U}$ by finding weight (basis) vectors $\mathbf{w}$ and $\mathbf{c}$ such that:

$$\begin{aligned}[cov(\mathbf{T},\mathbf{U})]^2 &= [cov(\mathbf{X}\mathbf{w},\mathbf{Y}\mathbf{c})]^2 \\ &= max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{X}\mathbf{r},\mathbf{Y}\mathbf{s})]^2\end{aligned} \tag{17}$$

where $cov(\mathbf{T},\mathbf{U}) = \mathbf{T}^{\mathbf{T}}\mathbf{U}/n$ denotes the sample covariance between score vectors $\mathbf{T}$ and $\mathbf{U}$. Weight vectors $\mathbf{w}$ and $\mathbf{c}$ are computed by the NIPALS algorithm [38] and stored into the projection matrices $\boldsymbol{W}$ and $\boldsymbol{C}$, respectively. Then, input and output data can be projected into the latent space by using these projections: $\hat{\boldsymbol{x}} = \boldsymbol{W}^T\boldsymbol{x}$ and $\hat{\boldsymbol{y}} = \boldsymbol{C}^T\boldsymbol{y}$.

4) Gaussian Process Regression (GPR): A Gaussian process (GP) is a statistical distribution for which any finite linear combination of samples has a joint Gaussian distribution. Therefore, any linear functional applied to the sample function will give a normally distributed result.

Given observed samples $\mathbf{D} = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n\} = (\mathbf{X}, \mathbf{Y})$, we formulate the GPR as follows:

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon_i \tag{18}$$

$$f \sim GP(\cdot|0, \mathbf{K}) \tag{19}$$

$$\epsilon_i \sim \mathcal{N}(\cdot|0, \sigma^2) \tag{20}$$

where f is the GP function which is distributed as a GP with zero mean and covariance function $\mathbf{K}$:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = k_1 * exp(-\frac{1}{2}\frac{|\mathbf{x}_i - \mathbf{x}_j|}{k_2^2}) + k_3\sigma^2 \tag{21}$$

where k_i are the hyperparameters of the GPR.

5) Weighted LSR (WLSR): In classical regression methods, each sample has the same impact on the regression. However, this is not completely valid for the subject-specific calibration in eye tracking. The quality of the calibration data is heterogeneous over the monitor for different calibration points due to several factors such as the viewing angle, subject's concentration and feature detection noise (Fig. 8). In fact, even the samples of the same point cluster may have different qualities. Therefore, we propose to extend the classical LSR to weighted LSR schemes in which calibration point clusters and/or individual samples have different impacts in the regression according to their quality (weights). In this study, we propose two weighting schemes.

i) *Cluster Weighting ($WLSR_{CW}$)*: We assign weights, $\mathbf{cw}$, to the point clusters according to the cluster variance. If the samples of a point cluster is concentrated, in other words, the cluster variance is low, a high weight is assigned to the point cluster (individual samples of the cluster have the same weights) and vice versa as follows:

$$\mathbf{cw_n} = \frac{1}{\frac{1}{K}\sum_k^K (x_n^k - \mu_n)^2} \tag{22}$$

where x_n^k is the k-th sample of the n-th calibration point, μ_n is the mean of the n-th calibration point.

ii) *Individual Sample Weighting ($WLSR_{IW}$)*: We assign weights, $\mathbf{iw}$, to each individual sample according to the distance to the corresponding point cluster's mean as well as the cluster variance. Samples with lower distance to the cluster mean are assigned with higher weights as follows:

$$\mathbf{iw}_n^k = \mathbf{cw}_n * w_n^k \tag{23}$$

$$w_n^k = \frac{\sum_k^K ||(x_n^k - \mu_n)||}{||(x_n^k - \mu_n)||} \tag{24}$$

where x_n^k is the k-th sample of the n-th calibration point, μ_n is the mean of the n-th calibration point, and $\mathbf{cw}_n$ is the normalized cluster weight of the n-th calibration point.

Once the weights are computed, they are stored in the weight matrix, $\mathbf{W}$, and used for the calculation of the regression parameters by the modified version of Eq. (9) as follows:

$$\widehat{\beta} = (\mathbf{X}\mathbf{W}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{W}\mathbf{Y}^T \tag{25}$$

6) Iteratively Reweighted LSR: Another factor that affects the calibration quality is the outliers caused by feature detection errors and user distractions during the data acquisition. In order to handle the outliers, our system detects some of the (global) outliers prior to the subject-specific calibration. It checks each individual calibration sample's distance to its point cluster center. If the distance is larger from a certain threshold distance, i.e., 4° of visual angle, the sample is filtered out as an outlier. This global outlier filtering is applied to detect outliers caused by momentary feature detection errors. However, it can possibly not detect outliers caused by user distractions or long-lasting feature detection errors. For example, if a user is distracted, i.e. change his/her gaze, for a second during data acquisition, or the feature detection fails to accurately detect the features for several frames, the system, in the meanwhile, will capture several bad samples (Fig. 9a).

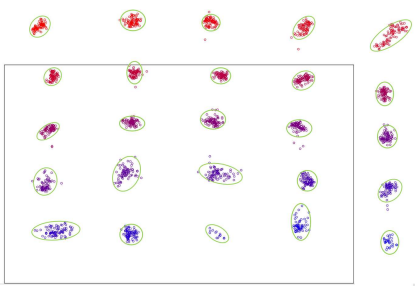


Fig. 8. Raw gaze calibration data obtained from a sample user.

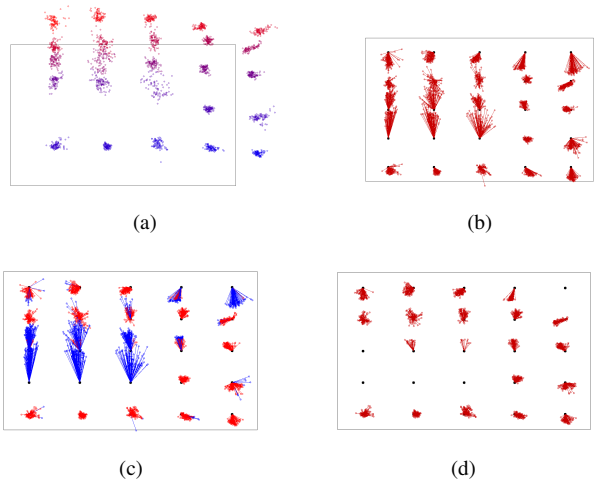


Fig. 9. Distracted user case **(a)** raw gaze data; **(b)** projection of samples with the model learned at the first iteration; **(c)** detected outliers (blue); **(d)** projection of samples with the final model. This figure is best viewed in color.

These samples may highly influence the cluster center. In such situations, global filtering fails to detect these outliers, which results in a decrease in the calibration quality.

To overcome such situations, we propose an iteratively reweighted calibration approach in which the weights of outliers are set to zero. The first iteration performs any of the calibration methods described previously. Then, instead of storing the learned model as the final calibration model, it first projects the samples using the learned model (Fig. 9b). In ideal conditions the projected samples should create dense clusters around the calibration points. It checks how close the samples are projected to their corresponding stimuli points. If the projected sample is farther from a certain threshold distance, i.e., 2° of visual angle, to its corresponding stimuli point, the sample is filtered out as an outlier and its weight is set to zero (Fig. 9c). The same applies to all the samples. The iteration is completed by (re)learning the calibration with the updated weights. The iterations continue until no further outliers are detected, and the final calibration model is stored as the final calibration model.

In fact, the situations in which the global outlier filtering is not sufficient arise rarely. We encountered only one case (out of ten cases) in our user experiments. Fig. 9 displays the visualization of such a situation and the impact of the proposed iterative method. The algorithm discards several bad samples from the calibration data, in fact, it may even discard a calibration point, in order to obtain a better calibration.

In this study, we investigate three iteratively reweighted LSR methods, namely, iterative Ridge, iterative $WLSR_{CW}$ and iterative $WLSR_{IW}$.

E. Adaptive Fusion Scheme

The proposed setup operates with low resolution eye data so as to provide a more realistic experience to users allowing free head movement. Even though it causes an accuracy drop for an eye due to the low resolution, it enables to capture and process both eyes simultaneously. In order to compensate for the accuracy decrease we propose to combine the estimated PoRs obtained from each eye, and output an overall PoR per frame. The proposed adaptive fusion scheme improves the overall estimation accuracy compared to the performance achieved using single eye. The adaptive fusion scheme performs a weighted averaging of individual PoRs obtained from both eyes as follows:

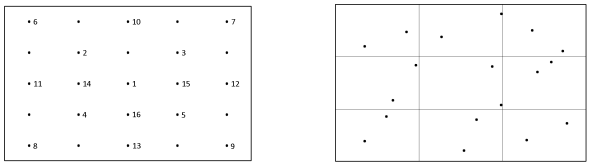
$$PoR_{overall} = \sum_i PoR_i * W_i, \quad \sum_i W_i = 1, \quad i \in \{L, R\},$$

where W_R and W_L are the weights for the right and left eye's PoRs 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 don't report an overall PoR in case both PoRs are unavailable for a given frame. In this study, we assign weights according to the reliability of the detected gaze features. We determine the feature reliability based on an assumption that better aligned features yield a more reliable PoR. So, given the features (glints and pupil center) of both eyes, a higher reliability is assigned to the eye whose dark pupil features (glints) and bright pupil feature (pupil center) are better matched. To this effect, we measure, for each eye, the distance between the pupil center and the center of the trapezoid formed by four glints. Then, we assign weights (reliabilities) inversely and linearly proportional to the measured distances. The proposed weight assignment is robust against outliers, and therefore, achieves a higher performance than uniform weighting. Yet, it is not an optimal weighting strategy as the assumption is invalid when users gaze at the edges of the monitor. In our future work, we plan to employ more accurate weighting methods. For instance, the calibration data and statistics (similar to [30]), users' eye dominance and head pose angles can alternatively be considered to determine the weights.

III. Evaluation of Investigated Methods

We conducted several experiments both on simulated data and real-data in order to evaluate the performances of all the investigated methods. We measured and reported the performances as gaze estimation accuracy error, which is defined as the average displacement in degrees of visual angle between the target stimuli points and the estimated PoRs, using all raw samples, i.e., no temporal smoothing or post-processing is applied. We chose to report the estimation errors in degrees of visual angle since it is invariant to the user distance to the monitor.

Both simulations and user experiments consisted of acquiring the calibration and test data. In calibration data acquisition, users were asked to look at 25 uniformly distributed target stimuli points on the screen as shown in Fig. 10a. The target points were displayed in a left to right and top to bottom sequence in a 5×5 grid on the monitor. Out of these 25 points, we formed 5 different calibration configurations, i.e., *5*, *9*, *13*, *16* and *25 points calibration*, according to the amount of calibration data in order to examine its impact on the test performance. In Fig. 10a, the numbers displayed on the right of the points indicate how these configurations were formed. For instance, the points from 1 to 5, from 1 to 9 and from 1



(a) (b)

Fig. 10. Target stimuli points **(a)** the calibration set; **(b)** a sample test set.

to 13 constitute *5*, *9* and *13 points calibration*, respectively. Hence, the sets with more points contain the smaller sets.

For the test data acquisition, we introduced a new testing protocol where we place the test points independently of the calibration points so as to avoid the problem of overfitting. Users were asked to look at 18 target points in a 3×3 grid covering the whole screen. Fig. 10b displays the positions of a sample set of test points. The positions of the target points in a region were randomly determined. We ensured the display of two points in each region in order to cover the whole screen. The display order of the regions and the points were also randomly determined. This enables to avoid reporting false test results due to overfitting on the calibration point locations. In addition, it simulates a more natural and realistic test condition.

Note that in our evaluation comparisons, we give higher priorities to the calibration configurations with lower number of target points, especially to *5 points calibration* configuration, among all the configurations since we target a convenient and user-friendly calibration which corresponds to as little effort as possible from the users. Hence, our conclusions on the performance comparisons rely mostly on the results of the *5 points calibration* configuration, and the statistical significance of the arisen differences has been checked by means of a paired sample T-test.

The further details of the evaluation are explained in the following subsections.

A. Simulation Setup

Simulation data was generated using the simulator developed by [39]. The simulator allows for detailed modeling of different components of the hardware setup and user eye in 3D. However, it also has a few limitations. For instance, the simulated eye model uses a spherical cornea model rather than an ellipsoidal one and it does not model the refraction on the posterior cornea surface. The camera does not exhibit any lens distortion or other imperfections. The light model does not account for the spatial extent of the light sources that the apparent shape of the light source can change depending on the direction from which it is viewed. Despite its limitations, the simulator provides a realistic simulation framework and the source code[1] is publicly available.

In our simulations, we tried to simulate the same environment and protocol that we used for the user experiments. The simulated monitor has a size of 24-inch, and four LEDs were placed on the corners of the monitor. A camera was located 2 cm below the monitor, and a user eye was located 70 cm away from the monitor. We simulated the user eye with cornea radius of 7.98 mm and we located the pupil center 6.2 mm from the cornea center. Furthermore, in order to simulate the real data containing certain noise and outliers, we include, for certain simulation setups, feature position errors, which introduce noise to the feature positions in camera images (e.g., 0.3 pixels per feature) and we added artificial outliers into the generated data.

For each calibration point, we generated 100 samples with and without noise and outliers depending on the simulation setup. On the contrary, we generated a single sample without any noise for each test point. With different simulation setups we examined the effects of the investigated calibration methods on the noise-free, noisy and outliered data.

B. Simulation Results

The results of simulation experiments are presented in Figs. 11 and 12. The figures illustrate the performances of the investigated regression-based calibration methods as well as a widely accepted homography-based method, namely, normalized homography (NHOM) proposed by [24]. Each figure demonstrates the average gaze estimation accuracy errors of different methods for each calibration configuration. Fig. 11 shows the results of the first simulation setup where the calibration data has neither noise nor outliers. As expected, non-linear regression methods (Ridge regression with polynomial kernel and GPR) do not perform better compared to linear regression methods when there are few calibration points. GPR requires more than 9 points to achieve an acceptable calibration quality. Ridge regression with polynomial kernel significantly outperforms all other methods when using 9 and more calibration points. However, its performance is significantly lower than NHOM and linear regression methods when there is only 5 calibration points. The classical linear regression methods show similar performances to NHOM for all the calibration configurations. The proposed weighted and iterative regression methods are not plotted in Fig. 11 because their performances are exactly equal to Ridge regression method because the data variance is absent. Hence, the first simulation experiment, in which the calibration data does not contain any noise or outliers, states NHOM outperforms regression-based methods for *5 points calibration* configuration.

On the other hand, Fig. 12a shows the results of another simulation setup where we introduce feature noise but no outliers, in the calibration data. The behavior of non-linear and linear regression-based methods and homography method stay the same as in Fig. 11 even though a performance drop of $\sim 0.1^\circ$ is observed due to the introduced feature noise. The major difference is observed in the performances of the proposed weighted LSR methods. Both $WLSR_{CW}$ and $WLSR_{IW}$ outperform NHOM when the calibration data contains certain noise. The performance difference between weighted LSR methods and NHOM gets higher when outliers are also included in the calibration data as can be seen in Fig. 12b.

[1]The MATLAB source code of the simulation framework can be downloaded at http://webmail.inb.uni-luebeck.de/inb-toolsdemos/FILES/et-simul-1.01.zip

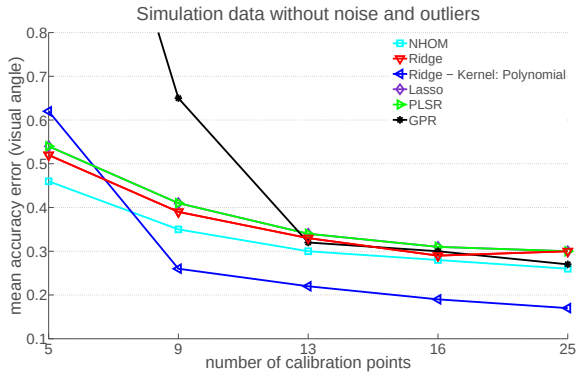
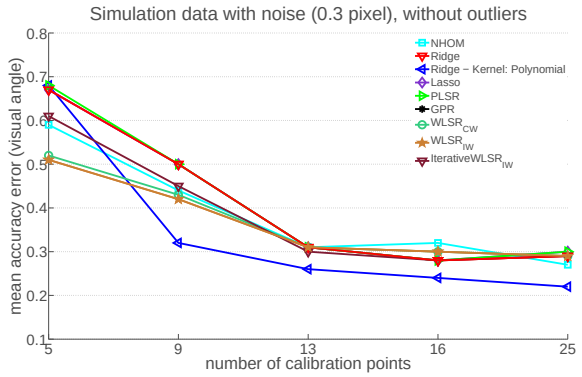
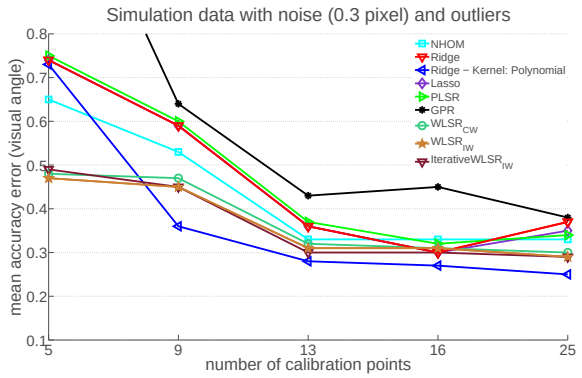


Fig. 11. Comparison of the calibration methods in case the simulation data has neither noise nor outliers.



(a)



(b)

Fig. 12. Comparison of the calibration methods in case the simulation data contains **(a)** feature noise; **(b)** feature noise and outliers.

Conducted simulation experiments illustrated effectiveness of the proposed weighting methods against all other methods, especially NHOM, when the calibration data contain certain noise and outliers, as in real data. However, further experiments on the real data is necessary to validate the results since the real data contains not only feature noise and outliers, but also other factors that were not taken into consideration in our simulations such as a spherical cornea model, no refraction on the cornea surface, no spatial extent of the light sources, user behavior (blinks, distractions, vision disorders) and perhaps the most importantly fixed head pose and location. Hence, we put more emphasis on real user experiments in our evaluations, the next section gives the details of the conducted experiments.

C. User Experiments

We conducted user experiments using the hardware setup described in Section II-A to evaluate and to compare the performances of the investigated subject-specific calibration methods. Ten users, nine of whom had no previous experience with any gaze tracking system, participated in our experiments. Since we targeted a generic and natural HCI environment, the ground truth data is collected in a natural manner where the users were asked to look at the target stimuli points naturally the way they feel comfortable. Therefore, we did not require the use of chin rest to keep the user's head still and to keep user's one of the eyes within the camera's FOV to capture high resolution eye data. The users are positioned in a distance of $\sim$ 70 cm to the monitor. Note this study focuses on generic regression-based user calibration for eye tracking, we do not explicitly address the head pose robustness issue as in [25], [27]. However, the proposed the system can decently deal with head pose variations in the range of natural user machine interaction. Table I shows the average head pose variation statistics, which are obtained using the pose estimation method described in [40].

In user experiments, each target stimulus point (Fig. 10) is displayed for 100 frames (3.33 seconds), and the data of both eyes during this period is captured. The size of the circular target varies continuously from an initial radius of 30 pixels to a final radius of 20 pixels to serve as visual stimulus.

The gaze estimation process starts with face tracking on the frames where we extract eye regions of size $\sim$130$\times$70 pixels. Eye region extraction is followed by feature detection where we detect four glints and a pupil center. Next, we apply CR-based gaze estimation with the detected features to calculate the initial PoR. This procedure provides us the raw gaze data. In the calibration process, we learn an estimation bias correction model on the raw gaze data obtained from the calibration session by minimizing the distances between the initial PoRs and the real target points, as stated in Eq. 7. The calibration is performed for each eye and for each user separately. In the test process, we apply the learned models to correct the raw gaze data estimated from the test session. The corrected PoRs of each eye are combined by the proposed adaptive fusion scheme to output an overall PoR for each frame.

In our user experiments, we also analyzed some important factors which highly affect the performance of gaze estimation systems such as the data resolution and amount of used eye

TABLE I
HEAD POSE STATISTICS (IN DEGREE) OBTAINED BY THE FACE TRACKER ON THE COLLECTED EXPERIMENTAL DATA.

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

data. Moreover, we compared the effectiveness and performances of the investigated methods as well as the state-of-the-art methods.

1) The Effect of Eye Data: Firstly, we examined the effect of the amount of eye data used for the overall PoRs estimation. Since the proposed hardware configuration enables to process both eyes simultaneously for a given frame, it is possible to use either or both of the eyes for gaze tracking. In this manner, we obtained results by altering the eye data i.e., *Single eye* (either left or right), *Strictly both eyes* and *Adaptive fusion*. *Adaptive fusion*, as defined in Section II-E, corresponds to calculating the overall PoR using all the available gaze data obtained from both eyes. If the gaze data is not available for both eyes, the gaze data of only available eye is used to set the overall PoR. On the contrary, *Strictly both eyes* calculates the overall PoR only if the gaze data is available for both eyes. Besides, to illustrate the impact of the proposed fusion weighting method, we calculated the fusion in two ways such as simple averaging (uniform weighting) and feature reliability-based weighting.

Table II and Fig. 13 illustrate the effect of the amount of eye data used for the overall PoRs estimation. The results are obtained for different calibration configurations using $WLSR_{IW}$ as the calibration method. We observe that individual eyes perform differently due to several factors such as different illumination (shading and reflection of ambient light or LEDs), head pose and eyeball rotations with respect to the camera and the gazed point on the monitor, or vision disorders (ambly-opia, eye laziness). More importantly, the results demonstrate that utilizing both eyes does not only improve the overall estimation performance significantly, but it also increases the estimation availability of the system compared to utilizing single eye data. The reason is that the data obtained from a single eye may not be reliable enough to output a PoR for some of the test points, especially those where we observe higher head pose or eyeball rotation. The gaze estimation availability of the system is defined as the percentage of frames in which the system is able to compute an overall PoR. As shown in Table II, the system outputs a PoR for 96.3% of all frames while a natural eye blink is detected for 1.86% of all the frames. Therefore, the system could not output a PoR only 1.84% of all the frames due to missing or bad features. Note that *Strictly both eyes* performs slightly better than *Adaptive fusion by simple averaging*, but the availability significantly drops. Using *Adaptive fusion by weighting* keeps the gaze availability higher while reaching to the performance of *Strictly both eyes*. Hence, the proposed weighting method

TABLE II
AVERAGE ESTIMATION ACCURACY ERRORS (IN DEGREE) AND GAZE AVAILABILITIES WHEN USING DIFFERENT EYE DATA.

Eye Data	Accuracy Error (°)	Gaze Availability(%)
Single eye (left)	1.08	90.7
Single eye (right)	1.18	95.1
Strictly both eyes	0.89	87.8
Adaptive fusion by simple averaging	0.92	96.3
Adaptive fusion by weighting	0.89	96.3

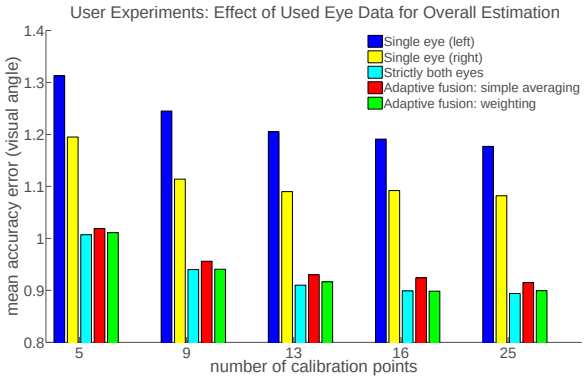


Fig. 13. The effect of used eye data for the overall estimation.

is more practical for a real-time gaze tracking system with higher estimation accuracy and availability.

Moreover, all the results consistently prove that the estimation error reduces with increasing number of calibration points used. However, increasing the number of calibration points greatly harm the user experience as discussed previously.

2) The Effect of Data Resolution: Secondly, we analyzed the impact of data resolution on the estimation accuracies in order to examine the system's flexibility in terms of face resolution change. Even though the proposed eye tracking system operates with relatively lower data resolution compared to the most of the previous work, we further downscaled the resolution to observe the robustness against data resolution. Sample eye regions extracted from an original frame and downscaled frames are shown in Fig. 14. The extracted eye region (Fig. 14a) from the original frame (1280×1024) has a resolution of 130×70 pixels, and the polygon formed by the glints is around 12×7 pixels. The original frames are downscaled in each dimension by 75% (960×768), 60% (768×614) and 50% (640×512) to generate different resolution data. The same feature detection and calibration methodology is applied on the generated data. We note that no particular parameter tuning is performed according to the resolution.

Table III and Fig. 15 illustrate the resolution impacts on the overall estimation accuracies when $WLSR_{IW}$ is used as the calibration method. The results show that downscaling by up to

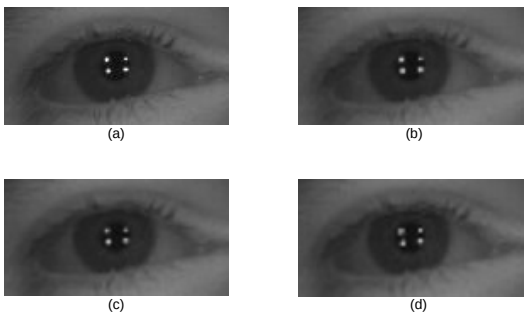


Fig. 14. Sample eye regions extracted from **(a)** an original frame; and downscaled frames by **(b)**75%; **(c)** 60%;**(d)** 50%.

TABLE III
AVERAGE GAZE ESTIMATION ACCURACY ERRORS (IN DEGREE) AND GAZE AVAILABILITIES WHEN USING DIFFERENT DATA RESOLUTIONS.

Resolution	Calibration		Gaze
	5 Points	25 Points	Availability(%)
Original frame	1.01	0.89	96.3
Downscaled by 75%	1.06	0.90	95.8
Downscaled by 60%	1.16	1.05	93.1
Downscaled by 50%	1.68	1.52	82.4

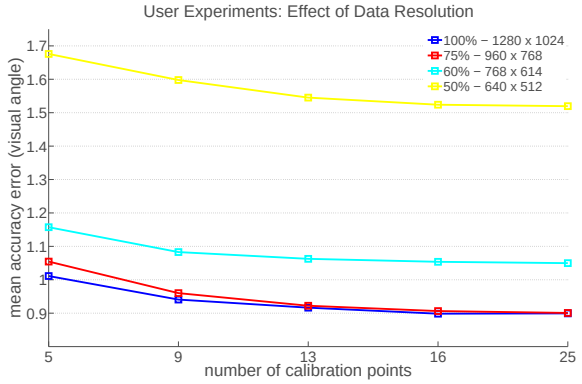


Fig. 15. The effect of the data resolution.

75% does not significantly affect the overall estimation accuracies. Towards 60% downscaling, the accuracy drop starts to get higher, and more than 60% downscaling results in a very significant performance decrease. We also observe that the impact stays consistent among different calibration configurations. Hence, the results indicate the system can tolerate a lower face resolution up to downscaling by around 75% without sacrificing too much accuracy. For further downscaling, we observe that the feature extraction, especially detection of glints, is highly affected by the low resolution. Therefore, less precisely detected features result in lower accuracies.

3) Comparison of Weighted and Iterative LSR Methods: Results from Fig. 16 show the comparison of Ridge regression with the proposed weighted and iterative LSR methods on real user data. The major observation is that the weighted LSR methods, i.e., $WLSR_{IW}$ and $WLSR_{CW}$, bring a performance improvement to the classical Ridge regression-based method, especially for *5 points calibration* configuration. $WLSR_{IW}$ seems to perform slightly better than $WLSR_{CW}$, but the difference is not significant according to the paired t-test.

Furthermore, even though iterative LSR methods bring additional computational burden for the calibration, they do not provide significant performance enhancement. In fact, the only improvement is achieved by iterative Ridge method over the classical Ridge method. We do not observe the same effect for iterative $WLSR_{IW}$ and iterative $WLSR_{CW}$ methods. We believe the effectiveness of the iterative methods greatly depends on the data. We designed the iterative methods to overcome the problem of outliers caused by user distractions and persistent feature misdetections during the calibration data acquisition as explained in Section II-D6. However, such situations arise

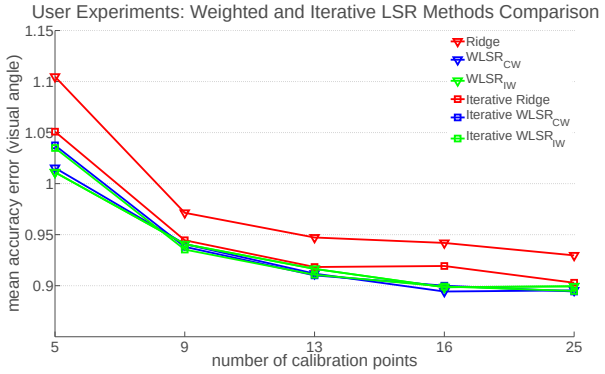


Fig. 16. Comparison of the proposed weighted and iterative LSR-based calibration methods.

rarely. In our experiments, we have encountered only one case out of ten users. Even though this particular subject's results are improved by the iterative methods, the influence on the overall results is negligible. In addition, another reason could be that iterative learning tends to overfit the calibration data since certain samples providing the data variance are eliminated during the iterations. Yet, they would provide a better calibration model for certain applications where the user data is rather noisy and contains a lot of outliers. In this paper, among all the proposed methods we suggest to utilize $WLSR_{IW}$ as the subject-specific calibration approach since it is efficient and computationally simpler. Note that only $WLSR_{IW}$ is plotted as the proposed method in the following sections for the clarity of the figures.

D. Comparison of Investigated Regression-based Methods

Fig. 17 shows the overall comparison of the investigated non-linear and linear regression methods as well as the homography method (NHOM). First of all, all linear regression methods significantly outperform NHOM and the non-linear regression methods, i.e., Ridge with polynomial kernel and GPR. This proves that linear regression methods are superior to non-linear methods because non-linear methods tend to overfit on the calibration data. Secondly, when there is few calibration data linear regression methods provide significantly better generalization capabilities than the homography-based

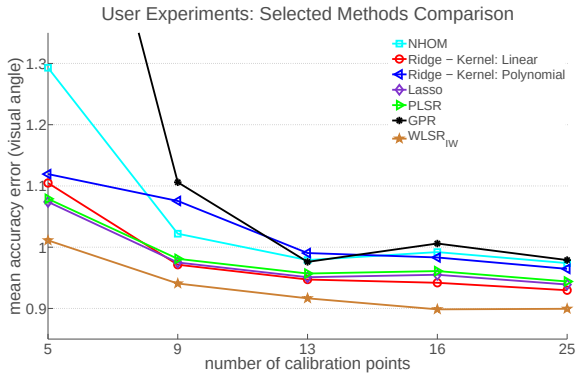


Fig. 17. Comparison of the investigated calibration methods.

TABLE IV
COMPARISON OF THE INVESTIGATED METHODS. AVERAGE ESTIMATION ACCURACY ERRORS ARE REPORTED IN DEGREES OF VISUAL ANGLE.

Method	Required Eye	Number of Calibration Points 5	9	16	25	Gaze Av. (%)
No calib [20]	Single	6.63	-	-	-	90.7
GPR [24]	Single	1.91	1.11	1.01	0.98	96.3
NHOM [24]	Single	1.39	1.14	1.09	1.07	90.7
NHOM [24]	Either	1.27	1.02	0.98	0.97	96.3
BHF [26]	Both	1.23	1.00	0.97	0.95	87.8
Ridge (poly)	Either	1.12	1.08	0.99	0.96	96.3
PLSR (poly)	Either	1.10	0.99	0.97	0.96	96.3
Ridge (linear) [28]	Either	1.10	0.97	0.94	0.93	96.3
PLSR (linear)	Either	1.08	0.98	0.96	0.94	96.3
Lasso	Either	1.07	0.98	0.96	0.94	96.3
Iterative Ridge	Either	1.05	0.95	0.92	0.9	96.3
Iter. $WLSR_{CW}$	Either	1.04	0.94	0.9	0.89	96.3
Iter. $WLSR_{IW}$	Either	1.03	0.94	0.9	0.89	96.3
$WLSR_{CW}$	Either	1.02	0.94	0.89	0.89	96.3
$WLSR_{IW}$	Either	1.01	0.94	0.9	0.9	96.3

methods due to reduced model parameters and relaxed constraints as detailedly explained in Section II-D1.

Furthermore, the proposed weighted LSR method, $WLSR_{IW}$, achieves the lowest average estimation accuracy error for all the calibration configurations. Especially for *5 points calibration*, the performance enhancement is noteworthy. This enables the method to be reliably utilized for a convenient subject-specific calibration for a user-friendly eye tracking system.

In addition, the performances of the other classical linear regression methods such as Ridge, Lasso regression and PLSR are all very similar. Different regularizations or utilization of a latent space for LSR does not seem to greatly influence the quality of the regression for the calibration problem since the number of input variables is small.

E. Comparison with Previous Work

We compare the performance of our proposed best performing calibration framework based on $WLSR_{IW}$ with the previous state-of-the-art calibration methods, namely, normalized homography (NHOM) [24], Gaussian process regression (GPR) [24], binocular homography fusion[2] (BHF) [26] and Ridge regression in our previous work [28]. A detailed comparison is listed in Table IV. Note that we did not include the comparison of earlier methods [21]–[23] due to two reasons: first, some methods require additional hardware material and second, the ones compared have been proven to perform better than the earlier methods. We did not also include the comparison with the methods ([25], [27]), that are explicitly proposed to bring robustness against large head movements. In fact, both methods are variations of NHOM method which are adapted for the head movement compensation. Therefore, the improvement over NHOM is marginal when there is not large head movements. Note that we focus more on a regression-based user calibration framework in this study and we leave to future work an explicit head movement robustness study.

[2]Experiments of [26] show that *binocular fusion* provides a significant improvement over the best single eye. However, the improvement is marginal over *averaging*. In our experiments, NHOM with *strictly both eyes* corresponds directly to *averaging* in [26]. Therefore, we consider the results of NHOM with *strictly both eyes* as the results of BHF in our comparisons. In fact, the real BHF is expected to achieve slightly better performance.

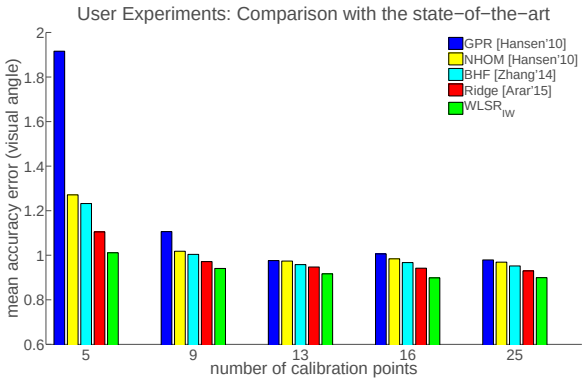


Fig. 18. Comparison with the state-of-the-art user calibration methods for the CR-based gaze estimation.

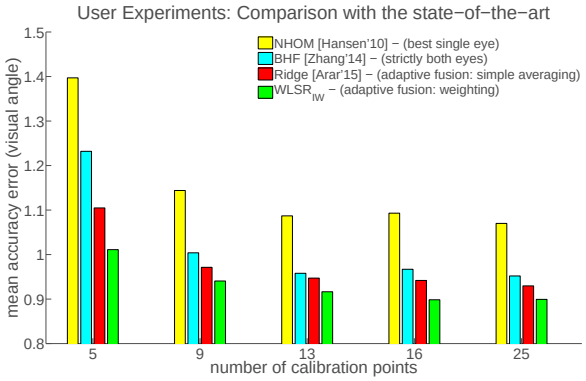


Fig. 19. Comparison with the original NHOM method.

The overall comparison of methods by altering the number of calibration points is shown in Fig. 18. In this figure, the proposed adaptive fusion of both eyes is applied to compute the overall PoRs. The results demonstrate that the proposed calibration approach, $WLSR_{IW}$, achieves the best estimation performances in all the calibration configurations. Especially for *5 points calibration* configuration, there is a notable enhancement achieved by both Ridge regression and its weighted extension, $WLSR_{IW}$, compared to NHOM and BHF. Note also that the statistical significance between Ridge regression and $WLSR_{IW}$ is verified by the paired t-test ($p < 0.05$). In addition, contradictory to the findings of [24], GPR performs worse than NHOM in our experiments. We think that this is possibly due to our testing protocol in which we chose the test points independent of the calibration points. Our experiments show that non-linear regression methods, i.e., Ridge regression with polynomial kernel and GPR, tend to overfit on the chosen calibration points.

Moreover, as can be seen from Table IV, utilization of both eyes through adaptive fusion scheme highly boosts the results. For instance, even though the improvement from NHOM to BHF does not seem significant from Fig. 18, there is, in fact, a significant increase compared to the original NHOM,

which utilized single eye data. For a more fair comparison, we should compare NHOM with single eye data against the others utilizing both eyes (BHF, Ridge and $WLSR_{IW}$). In this manner, assuming that the original NHOM uses the best performing single eye, the performance comparison could be as in Fig. 19. Therefore, a notable enhancement over NHOM is achieved by the other methods. On the other hand, it is important to note that BHF's gaze estimation availability (87.8%) is lower than those of Ridge and $WLSR_{IW}$ (96.3%). The reason is that BHF requires both eyes to be available to output a PoR while Ridge and $WLSR_{IW}$ can also output even if there is only single eye available.

F. Real-time Implementation

The software was developed on Windows platform in C++ language. OpenCV library is used for image processing, gaze features detection and the implementation of NHOM method. A publicly available Gaussian process library[3], which uses lapack routines for the matrix operations, is used for the implementation of GPR-based calibration. We implemented all other investigated calibration methods ourselves. In addition, SDM face tracker[4] is used for the localization of facial features prior to gaze features detection.

The computational complexity of the system is lower than 3D model-based methods as the gaze estimation is based on simple two-dimensional (2D) CR geometry. This enables to achieve a real-time implementation without requiring any performance optimization. In our implementation, the most computationally expensive process is face detection/tracking. PoRs estimation from both eyes using CR algorithm, proposed WLSR$_{IW}$ bias correction and adaptive fusion processes require much less computational effort. For instance, these three processes take only $\sim$8 ms on a PC with Intel i7 3.2GHz processor whereas face tracking itself takes $\sim$24 ms. Our current system can simultaneously output PoRs for both eyes as well as an overall PoR at $\sim$30 fps with a mean estimation accuracy error $\sim 1^\circ$ of visual angle ($\sim$50 pixels on the monitor) with only 5 calibration points[5]. Note that the computationally expensive face tracking process can be replaced with a simpler eye region detector, e.g., OpenCV eye detector, in order to achieve higher frame rates. As the feature extraction process does not require precisely located facial landmarks from a face tracker, but only requires a rough estimate of the eye region, simpler basic eye detectors would be employed to reach higher frame rates while achieving similar estimation accuracies.

IV. Discussion

In eye tracking, in addition to achieving high estimation accuracy, providing robustness against changing factors, such as ambient illumination, head and eye movements, eye wear, eye type, easiness of the calibration process, amount of calibration data, data resolution and working volume (availability) of the system, constitutes an important quality evaluation criteria. Robustness against most of these factors such as illumination, movement, eye wear and type, are highly related to the system design and choice of methods and algorithms for feature extraction and gaze estimation. In this manner, each eye tracking approach has advantages and disadvantages when compared with the others. For instance, 3D model-based systems are more robust against head and eye movements, whereas, appearance-based systems are more tolerant to the reflections on eye glasses. CR-based and 3D model-based systems are more robust against illumination changes while appearance-based methods are more vulnerable to illumination. Therefore, one can choose the most suitable approach depending on the desired application scenario. On the other hand, user calibration is inevitable and is required by all approaches to achieve high estimation accuracies. Undoubtedly, some of the mentioned factors, e.g., calibration data amount (number of calibration points required) and easiness of the calibration process, are more related to the calibration method. Since the major contribution of this work is the proposed user calibration methods, we chose to focus on the calibration related robustness issues. The analysis and discussions regarding these issues are given in Section III.

Together with the proposed user calibration methods, we suggested to use a CR-based gaze estimation algorithm due to its certain advantages over other methods as described in Section I. However, the proposed calibration method can be easily applied in any other state-of-the-art gaze estimation approaches, such as a 3D model-based approach, in order to improve the PoR accuracy.

A comparison of representative eye tracking techniques in several aspects such as accuracy, calibration requirements, hardware requirements and user friendliness, is given in Table V. Since a direct accuracy comparison with certain techniques is not possible due to particular hardware requirements, we listed accuracies in two separate columns. The first column (**Exp.**) lists the accuracies obtained from the experiments on our own dataset with *5 points calibration*, therefore, a direct performance comparison can be made. On the other hand, the second column (**Reported**) lists the reported accuracy errors achieved on individual datasets of corresponding studies, so a direct comparison would not be fair. Yet, the provided information helps us to make the following inferences. First of all, we observe that the popularity of appearance-based systems, which have lighter hardware requirements, have been increasing recently in parallel with the recent advancements in the synthesizing and rendering technology. Although their accuracies and head movement tolerances are currently not sufficient for precise eye tracking, their potential is likely to be exploited in the foreseeable future. Secondly, 3D model and CR-based systems undoubtedly outperform the appearance-based methods in terms of the accuracy. However, they mostly require particular hardware (i.e., IR cameras and light sources). Especially, 3D model-based systems need fully calibrated setups consisting of multiple cameras (or a kinect-like sensor) to accurately model the eye in 3D. Thirdly, CR-based systems have an important advantage over 3D model-based systems, that they require only an uncalibrated camera to accurately

[3]Library is publicly available at www.cs.umass.edu/~vidit/Code/GPR.tgz

[4]SDM face tracker library is available at www.humansensing.cs.cmu.edu/intraface/download_functions_cpp.html

[5]Note that the visual stimuli points displayed to users are circular targets with a varying radius of 20 to 30 pixels. Yet, the target PoRs are simply considered as the centers of the circles in our estimation error calculations.

TABLE V
COMPARISON OF REPRESENTATIVE REMOTE EYE TRACKING SYSTEMS. ACRONYMS: EXP (EXPERIMENTED), HR (HIGH RESOLUTION), CR (CROSS-RATIO-BASED SYSTEMS), 3D (3D MODEL-BASED SYSTEMS), AP (APPEARANCE-BASED SYSTEMS)

Method	Category	Acc. Error(°)		Calib.	Light	Camera(s)		Zoomed	Free Head
		Exp.	Reported	Points	Source(s)	#	Resolution	(HR) Eye	Movement [a]
Ours	CR	1.01°	∼1°	5	4+1	1 IR	1280 × 1024	×	✓
Coutinho [25]	CR	-	∼0.5°	9	4	1 IR	640 × 480	✓	×
Zhang and Cai [26]	CR	1.23°	∼0.6°	?	8	1 IR	1280 × 1024	×	×
Huang et al. [27]	CR	-	∼0.8°	25	8	1 IR	1280 × 1024	×	×
Hansen et. al [24]	CR	1.39°	∼1	5	4	1 IR	1280 × 1024	×	✓
Arar et. al [28]	CR	1.1°	∼1.1°	5	4+1	1 IR	1280 × 1024	×	✓
Yoo and Chung [21]	CR	-	∼1.3°	25	4+1	1 IR	640 × 480	✓	✓
Lai et. al [18]	3D	-	0.8°	9	2	2 IR	1600 × 1200	✓	✓
Villanueva and Cabeza [16]	3D	-	∼1°	1	2-4	1 IR	640 × 480	✓	×
Sun et. al [17]	3D	-	∼2°	10+ [b]	-	kinect	-	×	✓
Chen and Ji [19]	3D	-	∼3°	? [c]	2	1 IR	640 × 480	✓	✓
Mora et. al [41]	AP	-	1.7° [d]	42	-	kinect	-	×	✓
Lu et. al [10]	AP	-	2.4°	33 + 100	-	1	1280 × 1024	×	×
Lu et. al [11]	AP	-	2.5°	33 + 4	-	1	640 × 480	×	✓
Wood et. al [12]	AP	-	<10° [e]	-	-	-	-	×	✓

operate. Despite their uncalibrated setups and less complicated (2D) eye models, the accuracies are competitive with 3D model-based systems. It is clear that there is a performance gap between using chin rest and with free head movement for CR-based systems. The proposed calibration method is an effort to reduce the gap with an accuracy of ∼1 degree with minimum calibration effort. A further systematic robustness study on head movements is planned to be conducted in the future.

V. CONCLUSIONS AND FUTURE WORK

In this paper we present a real-time gaze estimation system which requires no effort in terms of the camera and geometric system calibration since the estimation of the gaze relies on simple 2D cross-ratio geometry. In addition, as opposed to most of the previous work, the system does not require high-resolution eye data to operate. In fact, operating with low-resolution data enables the system to output PoRs from each eye simultaneously. Obtained PoRs from both eyes are combined through an adaptive fusion scheme in order to achieve improved overall estimation accuracy. Furthermore, the estimation availability is enhanced compared to operating with single eye.

Moreover, an extensive investigation of different regression techniques for user calibration is carried out so as to compensate for the subject-specific estimation bias. A weighted least squares regression-based method is proposed for the purpose of a more convenient and user-friendly calibration process. The proposed method and system requires a few calibration points to achieve high estimation accuracy, and therefore, it increases the easiness of the calibration process. Both simulations and user experiments are conducted within a new evaluation scheme, where the test points are chosen independently of the calibration points, in order to avoid overfitting and increase the reliability of the results. The effectiveness of the proposed weighted regression-based calibration framework has been validated by both simulations and user experiments. The framework has been shown to outperform the state-of-the-art approaches as well as other investigated methods, especially when few points are used for calibration.

The results show that the average accuracy of the presented gaze estimation system is around 1°. The system offers a natural and personalized gaze tracking at 30 fps. Thus, it is highly suitable for several HCI applications.

Our future research directions focus on another significant challenge in eye tracking, that is the robustness of eye tracking systems against varying illumination, head pose changes and head/body movements, use of eye wear and different eye types.

ACKNOWLEDGMENT

This project is supported by the Swiss Commission for Technology and Innovation (CTI) under grant number 13594.1 PFFLR-ES and Logitech. The authors would also like to thank Yves Moser, Regis Croisonnier and Olivier Theytaz from Logitech Europe SA for their valuable contributions.

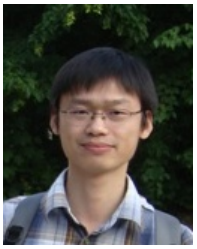
REFERENCES

[1] G. Underwood, *Cognitive processes in eye guidance*. Oxford University Press, 2005.
[2] Z. Zhu and Q. Ji, "Eye and gaze tracking for interactive graphic display," *Machine Vision and Applications*, vol. 15, no. 3, pp. 139–148, 2004.
[3] X. Yang, J. Sun, J. Liu, J. Chu, W. Liu, and Y. Gao, "A gaze tracking scheme for eye-based intelligent control," in *World Congress on Intelligent Control and Automation (WCICA)*, July 2010, pp. 50–55.
[4] M. Reale, S. Canavan, L. Yin, K. Hu, and T. Hung, "A multi-gesture interaction system using a 3-d iris disk model for gaze estimation and an active appearance model for 3-d hand pointing," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 474–486, June 2011.
[5] R. Valenti, N. Sebe, and T. Gevers, "What are you looking at?" *IJCV*, vol. 98, no. 3, pp. 324–334, 2012.
[6] C. Topal, S. Gunal, O. Kocdeviren, A. Dogan, and O. Gerek, "A low-computational approach on gaze estimation with eye touch system," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 228–239, Feb 2014.

[a] Use of a chin rest during the experiments. (×): used, (✓): not used
[b] Online calibration, on average more than 10 calibration points needs to be shown to reach the reported accuracy.
[c] Online calibration, on average 20 to 200 frames are shown to users for calibration to reach the reported accuracies.
[d] Reported error for the static head pose and person-specific regression
[e] Training set contains one million synthesized images. Cross-dataset evaluation for *in-the-wild* scenarios is performed.

[7] D. W. Hansen and Q. Ji, "In the eye of the beholder: a survey of models for eyes and gaze," *PAMI*, vol. 32, no. 3, pp. 478–500, 2010.
[8] X. Brolly and J. Mulligan, "Implicit calibration of a remote gaze tracker," in *CVPRW*, 2004.
[9] D. Hansen and A. Pece, "Eye tracking in the wild," *CVIU*, vol. 98, no. 1, pp. 182–210, 2005.
[10] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE PAMI*, vol. 36, no. 10, pp. 2033–2046, 2014.
[11] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Gaze estimation from eye appearance: a head pose-free method via eye image synthesis," *IEEE Transactions on Image Processing*, 2015.
[12] E. Wood, L.-p. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *ETRA*, 2016, pp. 131–138.
[13] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," in *CVPR*, 2003, pp. 451–458.
[14] J. Shih, S.-W.and Liu, "A novel approach to 3D gaze tracking using stereo cameras," *Transactions on Systems, Man, and Cybernetics*, vol. 34, no. 1, pp. 234–245, 2004.
[15] B. Noureddin, P. Lawrence, and C. Man, "A non-contact device for tracking gaze in a human computer interface," *CVIU*, vol. 98, no. 1, pp. 52–82, 2005.
[16] A. Villanueva and R. Cabeza, "A novel gaze estimation system with one calibration point," *Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 38, no. 4, pp. 1123–1138, 2008.
[17] L. Sun, M. Song, Z. Liu, and M.-t. Sun, "Realtime gaze estimation with online calibration," *IEEE Multimedia*, pp. 1–6, 2014.
[18] C. C. Lai, S. W. Shih, and Y. P. Hung, "Hybrid method for 3-D gaze tracking using glint and contour features," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.
[19] J. Chen and Q. Ji, "A probabilistic approach to online eye gaze tracking without explicit personal calibration," *IEEE Transactions on Image Processing*, 2015.
[20] 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 *FGR*, 2002.
[21] D. H. Yoo and M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," *CVIU*, vol. 98, no. 1, pp. 25–51, 2005.
[22] F. Coutinho and C. Morimoto, "Free head motion eye gaze tracking using a single camera and multiple light sources," *Brazilian Symposium Computer Graphics and Image Processing (SIBGRAPI)*, pp. 171–178, 2006.
[23] J. J. Kang, E. D. Guestrin, W. J. Maclean, and M. Eizenman, "Simplifying the cross-ratios method of point-of-gaze estimation," in *Canadian Medical and Biological Engineering Conference*, 2007.
[24] D. W. Hansen, J. S. Agustin, and A. Villanueva, "Homography normalization for robust gaze estimation in uncalibrated setups," in *ETRA*, 2010.
[25] F. L. Coutinho and C. H. Morimoto, "Improving head movement tolerance of cross-atio based eye trackers," *IJCV*, vol. 101, no. 3, pp. 459–481, 2013.
[26] Z. Zhang and Q. Cai, "Improving cross-ratio based eye tracking techniques by leveraging the binocular fixation constraint," in *ETRA*, 2014.
[27] 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 *ETRA*, 2014.
[28] N. M. Arar, H. Gao, and J.-P. Thiran, "Towards convenient calibration for cross-ratio based gaze estimation," in *WACV*, 2015, pp. 642–648.
[29] N. M. Arar, H. Gao, and J.-P. Thiran, "Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras," in *FGR*, 2015.
[30] N. M. Arar and J.-P. Thiran, "Estimating fusion weights of a multi-camera eye tracking system by leveraging user calibration data," in *ETRA*, 2016, pp. 225–228.
[31] J. J. Kang, M. Eizenman, E. D. Guestrin, and E. Eizenman, "Investigation of the cross-ratios method for point-of-gaze estimation," *Transactions on Biomedical Engineering*, vol. 55, no. 9, pp. 2293–302, 2008.
[32] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," *CVPR*, 2013.
[33] P. Viola and M. Jones, "Robust real-time face detection," *IJCV*, vol. 57, pp. 137–154, 2004.
[34] Y. Ebisawa, "Improved video-based eye-gaze detection method," *Transactions on Instrumentation and Measurement*, vol. 47, no. 4, pp. 948–955, 1998.
[35] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2005.
[36] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[38] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Conference on Subspace, Latent Structure and Feature Selection*, 2006.
[39] M. Boehme, M. Dorr, M. Graw, T. Martinetz, and E. Barth, "A software framework for simulating eye trackers," in *ETRA*, 2008.
[40] S. Chen, C. Wu, S. Lin, and Y. Hung, "2d face alignment and pose estimation based on 3d facial models," in *Conference on Multimedia & Expo (ICME)*, 2012.
[41] K. A. Funes-Mora and J.-M. Odobez, "Gaze estimation in the 3d space using rgb-d sensors," *IJCV*, pp. 1–23, 2015.

Nuri Murat Arar received his B.Sc. from Bilkent University, Turkey in 2010 and M.Sc. from Bogazici University, Turkey in 2012 in computer engineering. Since then he is pursuing a PhD degree at the Signal Processing Laboratory (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. His main research interest include eye tracking, gaze estimation and facial image analysis. He is a student member of the IEEE since 2011.



Hua Gaol received the Dipl.-Inf. and Ph.D. degrees in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2008 and 2013, respectively. This work was done when Hua was working as post-doc researcher at EPFL, Lausanne, Switzerland. His research interests include the fields in facial image processing, e.g. face tracking, 3D face reconstruction, facial expression recognition and face recognition.

Jean-Philippe Thiran is Associate Professor of Image Processing and director of the Signal Processing Laboratory (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an Associate Professor position with the Department of Radiology of the University Hospital Center (CHUV) and University of Lausanne (UNIL). His research field is image analysis and multimodal signal/image processing, with applications in many domains including medical image analysis, human-computer interaction, remote sensing of the Earth and surveillance. Dr Thiran is author of co-author of more than 150 journal papers, 9 book chapters, more than 210 papers in peer-reviewed proceedings of international conferences, and holds 4 international patents. He is a senior member of the IEEE.