Calibration Free, User Independent Gaze Estimation with Tensor Analysis

Nanxiang Li and Carlos Busso
The University of Texas at Dallas
800 W Campbell Rd, Richardson, TX 75080

Abstract

Human gaze directly signals visual attention, therefore, estimation of gaze has been an important research topic in fields such as human attention modeling and human-computer interaction. Accurate gaze estimation requires user, system and even session dependent parameters, which can be obtained by calibration process. However, this process has to be repeated whenever the parameter changes (head movement, camera movement, monitor movement). This study aims to eliminate the calibration process of gaze estimation by building a user-independent, appearance-based gaze estimation model. The system is ideal for multimodal interfaces, where the gaze is tracked without the cooperation from the users. The main goal is to capture the essential representation of the gaze appearance of the target user. We investigate the tensor analysis framework that decomposes the high dimension gaze data into different factors including individual differences, gaze differences, user-screen distances and session differences. The axis that is representative for a particular subject is automatically chosen in the tensor analysis framework using LASSO regression. The proposed approaches show promising results on capturing the test subject gaze changes. To address the estimation shift caused by the variations in individual heights, or relative position to the monitor, we apply domain adaptation to adjust the gaze estimation, observing further improvements. These promising results suggest that the proposed gaze estimation approach is a feasible and flexible scheme to facilitate gaze-based multimodal interfaces.

Preprint submitted to Image and Vision Computing *May 2, 2018*

Keywords: User-independent gaze estimation, tensor analysis, LASSO regression, domain adaptation, human computer interaction

1. Introduction

Gaze is a natural and fast way for users to interact with a computer. Visual attention from gaze estimation can be used for either diagnostic analysis or interactive applications [1]. The former uses gaze to understand users' visual attention process, while the latter applies gaze to respond or interact with the user. In addition, the gaze estimation has been used to further understand high level human behaviors such as cognitive distractions [2, 3]. Due to these potentials, the areas of eye tracking and gaze estimation have been extensively studied [4, 5, 6].

The main goal for a gaze-based interface is to map the user's gaze behavior to the coordinates in the user interface. This mapping is usually accomplished by a calibration process, during which the user looks at the interface's screen and has his/her gaze behavior recorded [6]. Important parameters are estimated through the calibration process and used for gaze estimation. For example, one common approach is to use the reflections in the corneal from an infrared / near-infrared non-collimated light to establish a reference position. The vector between the pupil center and this reference position is used to infer the gaze [7, 8]. Depending on the systems, important parameters including the camera setting parameters, user head pose, user relative position to the camera, or even user eye curvature are needed for gaze estimation. Some of these parameters are closely related to the user. Thus, they need to be re-estimated for different users or when the current user moves. Although adding constraints to the user or using intrusive devices can reduce the calibration requirement, these systems do not provide comfortable setting for practical usage. They also require the collaboration of the user, limiting potential applications. The constrained settings and the repetitive calibration processes prevent the application of gaze estimation for advanced *human-computer interactions* (HCIs).

This study proposes a user-independent, appearance-based method for gaze tracking where the main goal is the elimination of the calibration process. This system provides a flexible setting for gaze-aware HCIs. The key contributions of this study include: (1) We propose the tensor analysis framework to reduce the gap in performance between user-dependent and user-independent conditions. It decomposes the high dimension gaze data into different factors including individual differences, gaze differences, user-screen distances and session differences. The approach automatically selects important axis to the test subject using LASSO regression. (2) We apply domain adaptation to the gaze estimation results to further reduce the subject dependent error, achieving improved results. The proposed method reduces the gap in performance in estimating the gaze using the user-dependent and user-independent conditions. These promising results suggest that the proposed system can be effectively used in gaze-aware multimodal interfaces.

The paper is organized as follows. Section 2 reviews the state-of-the-art in the field of gaze estimation, emphasizing appearance-based approaches. It suggests open challenges in this research area, highlighting the contributions of this study. Section 3 presents the protocol behind the data collection of the MSP-GAZE database that is used in this study. The section describes the various factors considered for the recordings. Section 4 motivates the need for the proposed tensor-based approach to capture various sources of variability. Section 5 explains our proposed approach based on tensor analysis, LASSO regression, and domain adaptation. It reports the performance of the aforementioned method, comparing our results with other related approaches. Section 6 concludes the paper with final remarks, limitations of the study, and our future research directions in this area.

2. Related Work

Gaze estimation has been an important research topic, where many studies have advanced the state-of-the-art in this area. There are several approaches to

estimate gaze direction. Our proposed framework is an appearance-based model, so the focus of this section is on describing related appearance-based methods, and establishing the contributions of our work. For more comprehensive surveys on gaze estimation methods, we refer the readers to the studies of Duchowski [1] and Hansen and Ji [6].

2.1. Gaze Estimation

To achieve high accuracy in gaze estimation, many studies applied intrusive equipments including head mount devices, chin rest set and contact lenses [9]. These devices aim to reduce the variability of the system and improve the parameter estimation for the target user. For example, chin rest set limits users' head movement; head mount devices fix the relative position between the camera and the users' eyes. These intrusive equipments pose constraints to the users, leading to limited practical solutions for multimodal interfaces.

Non-intrusive techniques mostly consider adding light sources and cameras. One of the most commonly used light source is *infrared* (IR) light. Since it is invisible to human eyes, infrared camera can robustly capture the glint (reflections of light off the cornea) against various illumination conditions [6]. In fact, several studies have used the glint for eyes detection [10, 11, 12]. The glint provides a reference point of the human eye to the IR light sources. The gaze estimation can be implemented based on the gaze vector between the pupil/iris center and the glint [7, 8]. A 3-D eye model can be estimated when the system uses multiple cameras [10, 12].

An alternative non-intrusive gaze estimation approach relies on appearance-based techniques. The idea is to capture the image of the eye and build a mapping between the eye appearance and the gaze position / direction. Some studies considered direct mapping between the high dimensional eye image and the gaze position [13, 14, 15], while others extracted eye image features such as pupil, iris and eye corner [16, 17]. Then, the mapping can be established between the eye appearance and the gaze position. This is usually achieved by regression models where the image or image features are used as independent variables and

the corresponding gaze position is used as the dependent variable. Different regression models have been considered such as *support vector regression* (SVR) [18], localized linear regression [19, 13, 14], and Gaussian processes [20]. Instead of the eye image features, some studies used ellipse fitting approaches to model the iris/pupil contours [21, 22, 23], predicting gaze based on the fitting results. Other studies have also considered segmenting the eye image into the iris, sclera (white part of the eye), and the surrounding skin. The resulting regions can then be pixel-wise matched with 3-D rendered eyeball models using different parameters [24, 25].

Due to its simple system requirement, there has been an increased interest in appearance-based methods for gaze estimation. However, the predominant approach followed by many studies considers user-dependent conditions that require calibration. S. Baluja and Pomerleau [26] used neural networks to estimate the gaze position using the eye image. Tan et al. [27] used local linear interpolation among sparse appearance samples to approximate the gaze. Few studies have considered appearance-based gaze estimation models under user-independent conditions. Schiele and Waibel [28] studied user-independent gaze estimation with neural network, where they estimated coarse gaze directions based on head pose without considering the iris position. Rikert and Jones [29] used morphable models to estimate user-independent gaze, where the main challenge is the initial match between the model parameters and the image.

2.2. Contributions of this Work

This study aims to eliminate the repetitive calibration process while simultaneously maintaining the accuracy in gaze estimation in the context of HCI. We achieve this goal with a tensor-based framework that combines the tensor decomposition and LASSO regression to represent the user-independent gaze model. During gaze estimation, we observed a shift of the gaze estimation among different users. To address this problem, we applied domain adaptation based on the assumption that the user looks at the center of the screen most of the time [30, 31]. This approach achieves improved performance.

Our approach differs from the majority of previous techniques, in that it models various factors that affect the appearance-based gaze estimation approaches. The approach relies on a principled framework relying on tensor analysis and LASSO regression. Moreover, the proposed approach relies on modeling user-independent gaze estimation using a large training data, removing the need to calibrate the system.

Among existing approaches, the appearance-based gaze estimation approach proposed by Lu et al. [19] shares similar ideas. However, instead of directly relying on large training data, they applied an *adaptive linear regression* (ALR) model to find a subset in the training samples. In particular, they formulated a l_1 optimization problem where the objective was to minimize the error between a linear combination of a subset of training images and the test image, predicting the gaze. In section 5.3, we compare our approach with this framework [19]. The experimental evaluation shows the benefits of the proposed tensor-based method, which achieves significant better performance.

2.3. Connection to our Previous Work

Our previous studies considered *principal component analysis* (PCA) as a framework to represent the essential components to reconstruct the eye appearance for various gaze positions [15]. While the focus was on user-dependent conditions, we noticed significant decrease in performance when we extended the PCA model to user-independent conditions (see Tables 2 and 3). These results suggest that important factors for gaze estimation are introduced by user variability, which were not properly captured by the PCA-based models. In Li and Busso [32], we proposed the idea of finding similar training images to the testing image in a PCA representation space to reduce the variability between the users. While we improved the performance, we still observe a performance gap between the user-dependent and user-independent models. In contrast, the proposed gaze estimation method relies on a novel tensor-based framework, providing a principled approach to address this problem. By combining this approach with domain adaptation, we have significantly reduced the performance

gap between user-dependent and user-independent conditions.

3. MSP-GAZE Database

The study relies on the MSP-GAZE corpus [15], a multimodal database to design gaze estimation systems for HCI. The purpose of the MSP-GAZE database is to evaluate appearance-based gaze estimation methods against various factors, including individual eye appearance differences, head movements, and various distances between the user and the interface's screen. The data was collected in a controlled laboratory environment. It involves letting the participants look at and click on randomly generated points displayed on the computer screen, while recording their gaze behavior. The system includes a standard 22-in HP monitor with a screen resolution set to 1680×1050, a commercial webcam (Logitech C920) placed on top of the monitor, and a Microsoft Kinect sensor for Windows placed below the monitor, as shown in Figure 1. The center of the webcam and the RGB sensor of the Kinect are aligned with the center of the monitor. Both devices record the subjects from different angles and are synchronized using a clapping board at the beginning of each recording. This study relies only on the videos from the webcam.

3.1. Sources of variabilities included in the corpus

3.1.1. Appearance

To cover different eye appearance in the database, we considered a gender balanced data collection from 46 participants (23 male and 23 female). The average age of the participants is 22.7, with the oldest being 35 years old and the youngest being 19 years old.

Another important factor for the eye appearances is the ethnic background. The participants in the MSP-GAZE cover the dominant ethnic groups at the University of Texas at Dallas including Caucasian, Asian, Indian and Hispanic participants. Specifically, there are 10 subjects from each of the Asian, Indian and Hispanic groups. There are 16 subjects from the Caucasian group. The

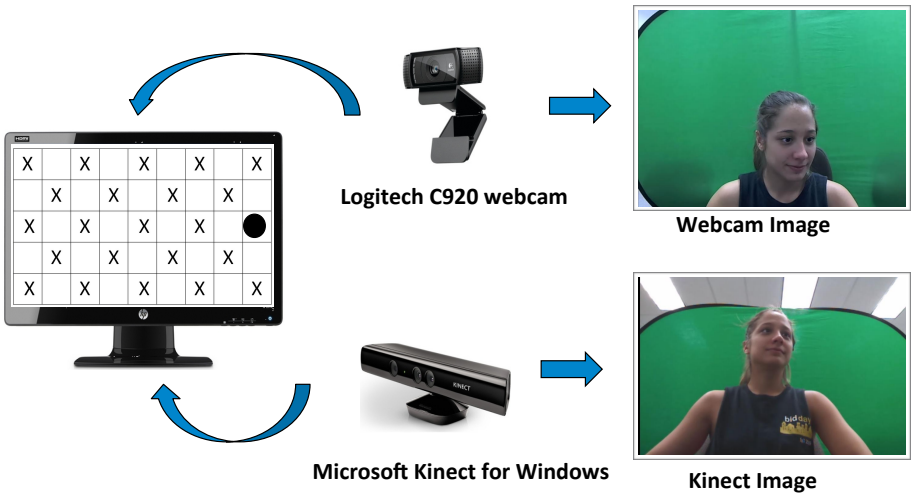

Figure 1: The data collection includes a 22-inch HP monitor, a Logitech C920 webcam and a Microsoft Kinect for Windows. A green screen is placed behind the subject to provide uniform background.

corpus is also gender balanced within each ethnic group. For example, there are five male and five female participants from the Asian group.

3.1.2. Session

To evaluate the robustness of the appearance gaze model against time, each subject participated in two sessions for different days. On average, the sessions were collected with a 5-day interval. Some sessions were recorded in the early morning, and some sessions were collected in the late afternoon. Over different sessions, we noticed that some participants appeared quite differently between the two sessions (e.g. different makeups, hair styles). These differences allow us to evaluate how reliable the appearance-based gaze model is for user-dependent condition across sessions.

3.1.3. Head movement

In each session, 14 recordings are collected following the conditions listed in Table 1. The first seven recordings allow participants to freely move their head to capture normal user computer interaction. During the last seven recordings,

Table 1: Recording conditions for each session. The data collection follows the order listed in the table.

Recording	Head Movement	Distance	Pattern
1	Yes	User-defined	Testing
2	Yes	User-defined	Training
3	Yes	Near	Training
4	Yes	Medium	Training
5	Yes	Medium	Training
6	Yes	Far	Training
7	Yes	Far	Training
8	No	User-defined	Testing
9	No	User-defined	Training
10	No	Near	Training
11	No	Medium	Training
12	No	Medium	Training
13	No	Far	Training
14	No	Far	Training

we asked the participants to maintain their head still while completing the tasks. We did not use a head mount device to make the data collection comfortable for the participants. The differences between these two sets of data allow us to evaluate the effect of head movement on appearance-based gaze estimation in user-computer interaction.

3.1.4. Distance

To evaluate the effect of user-screen distance on gaze estimation, the MSP-GAZE database considers four different distance settings. As shown in Table 1, the data collection starts with the "User-defined" distance setting where a subject selects his/her preferred distance to the monitor. Then, we consider three predefined distances that cover the range of common user-screen distance: "Near" (0.4 m), "Medium" (0.5 m) and "Far" (0.6 m). This process is repeated

under both head movement settings (i.e. with and without head motion).

In our data collection, we observed that the "User-defined" distance always lies within the "Near" and "Far" distances, suggesting that these distance settings cover the range of common user-computer interaction distances. Most participants perceived the "Near" distance setting as too close to the monitor, especially for the head constrained recordings where the participants only used eye movement to look at the target points. For this reason, we only collected one recording in this distance setting. The same protocol shown in Table 1 is repeated for the second session, which is collected on a different day.

Overall, we collect about 90 min of data over the two sessions from each subject. The subjects were encouraged to take short breaks between recordings. We also give 10-min breaks at the middle of the session to reduce fatigue.

3.2. Train and test datasets

The data collection should efficiently cover different areas on the screen to develop an appearance-based gaze estimation system that maps the captured eye image to the screen position. Following previous studies [33, 34, 27], we divide the screen into 5 by 9 grids as illustrated in Figures 1. This division defines 45 grids, from which we only use 23 grids, marked with 'X' in Figure 1. We randomly generate a white point inside one of the 23 highlighted grid areas (i.e., for a given grid, the points appear at different locations within the region). The subject is asked to click the point with the mouse cursor. The point turns green once the user clicks it, and stays still for 1 second before jumping to a different location. We introduce the mouse click action to get the time at which the participant is looking at the target point, avoiding transient frames in which the point jumps from grids. It also ensures that the subject is not distracted during the data collection (i.e., looking at the time, missing points). The target point appears four times in each of the 23 marked grid areas in random order. This design ensures enough sample data, while limiting the duration of the recording (i.e., 92 points are collected in approximately 3 min). We record the videos from the cameras, the actual location of the points, the

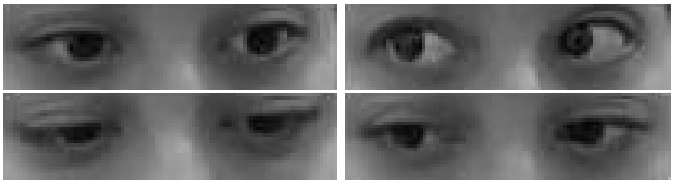

Figure 2: Eye pair samples extracted with the Viola-Jones algorithm. These examples correspond to cases where the subject was looking at points in the corners.

mouse cursor locations, and the mouse click actions. This protocol is used in the 12 recordings labeled as "Training" in Table 1.

Two recordings in Table 1 are collected with a slightly different protocol. For these recordings, which are referred to as "Testing", 92 different points are randomly shown on the monitor without considering the grid areas. The data collected with the "Testing" pattern is used to evaluate the performance of the regression model using the proposed methods (see Section 5). We follow this pattern in the first (with head movement) and the eighth (without head movement) recordings, as listed in Table 1.

3.3. Eye detection

We extracted patches with both eyes, following the approach proposed in our previous work [15] (see Fig. 2). Considering both eyes reduces eye detection errors, improving the robustness against head motion. The eye pair image is automatically extracted using the Viola-Jones object detection framework. We use the implementation provided by the *open computer vision library* (OpenCV) with the eye-pair detector developed by Castrillón et al. [35]. For each point displayed on the screen, we extract the eye pair images from the three frames (≈ 0.14s) immediately after the user clicks on the target points. For each of the 14 recordings per session, we consider 92 points resulting in 276 eye pair images (92 points $\times$ 3 frames). These detected eye-pair images are resized to 25 $\times$ 100 pixel images. Due to equipment malfunctions, few video files were not correctly formatted. As a result, this study considers 44 out of the 46 subjects whose data are completely recorded.

4. Motivation: Appearance Based Framework for Gaze Estimation

Principal component analysis (PCA)-based approaches have been successfully applied to human face recognition in the form of eigenfaces [36] and Fisherfaces [37]. For human gaze related studies, it has been used for eye detection [38, 39] and gaze estimation [33, 34]. The approach consists of representing a set of N aligned images with an orthonormal basis estimated from the covariant matrix of the images. This basis is computed using eigenvalue decomposition according to equation 1, which defines a new coordinate system using the bases $[\mathbf{u_1}, \ldots, \mathbf{u_n}]$. The eigenvectors associated with the larger eigenvalues define directions with the higher variability. By considering only eigenvectors with larger eigenvalues, we can reduce the dimension of the feature vector while capturing the appearance of the image.

$$\begin{aligned} \Sigma &= \begin{bmatrix} \mathbf{u_1} & \ldots & \mathbf{u_n} \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u_1} & \ldots & \mathbf{u_n} \end{bmatrix}^T \\ &= U \Lambda U^T \end{aligned} \tag{1}$$

Our previous study considered PCA as a framework to represent the essential components to reconstruct the eye appearance for various gaze positions [15, 32]. Our assumption is that the main variable that causes changes in eye pair images is the gaze direction, especially for the user-dependent gaze condition, where the train and test data are from the same subject. In this case, the height, head movement preference and eye appearance are similar in train and test data.

We used the projections into the *principal components* (PCs), as features for two linear regression models $(p_1, p_2, \ldots, p_N)$. These two models are separately built to estimate the gaze position in the horizontal (x) and vertical (y) coordinates, respectively. N is the number of PC projections considered in the model. The projections are used as independent variables. The mapped screen positions (horizontal coordinate x and vertical coordinate y) are used as the

dependent variables (see Eqs. 2 and 3). The output of these regression models are limited to be within the screen size.

$$x = \begin{cases} 0, & \text{if } x < 0 \\ 1680, & \text{if } x > 1680 \\ \beta_{x0} + \beta_{x1}p_1 + \cdots + \beta_{x30}p_{30}, & \text{else} \end{cases} \tag{2}$$

$$y = \begin{cases} 0, & \text{if } y < 0 \\ 1050, & \text{if } y > 1050 \\ \beta_{x0} + \beta_{y1}p_1 + \cdots + \beta_{y30}p_{30}, & \text{else} \end{cases} \tag{3}$$

We used the angular error between the true and predicted gaze positions (θ_{error}) to assess the performance of the proposed gaze estimation approach. We assume that the eye pair center of the subjects is aligned with the center of the monitor. Therefore, the distance between the subject eye pair center and the estimated gaze point on the monitor, $d_{eye-estimate}$, and the distance between the subject eye pair center and the ground truth gaze point on the monitor, $d_{eye-true}$, can be estimated with Equations 4 and 5, respectively. We use the following definitions: d_{eye-mc} is the distance between the user and the monitor center; $d_{mc-estimate}$ is the distance between the monitor center and the predicted gaze position; and $d_{mc-true}$ is the distance between the monitor center and the ground truth gaze position. We estimate d_{eye-mc} using the original size of the eye pair image. For each subject, we estimate the average size of the detected eye pair images for the "Near" and "Far" recordings during the same session. Then, d_{eye-mc} is estimated by linearly interpolating the width of the eye pair image between the corresponding values for the "Near" and "Far" conditions. Notice that d_{eye-mc} can also be derived from the depth images provided by the Kinect sensor. For future work, we will rely on the depth images for more accurate distance information.

$$d_{eye-estimate} = \sqrt{d_{eye-mc}^2 + d_{mc-estimate}^2} \tag{4}$$

$$d_{eye-true} = \sqrt{d_{eye-mc}^2 + d_{mc-true}^2} \tag{5}$$

Using these estimation, the gaze angular error θ_{error} can be estimated using the law of cosines:

$$\theta_{error} = \arccos\left(\frac{d_{eye-estimate}^2 + d_{eye-true}^2 - d_{error}^2}{2 \times d_{eye-estimate} \times d_{eye-true}}\right) \tag{6}$$

where d_{error} is the distance between the ground truth and estimated gaze point.

We considered both user-dependent model (train and test datasets are from the same subject) and user-independent model (train and test datasets are from different subjects). For user-dependent evaluation, we calculate the angular error for each of the 44 subjects and report the average performance. For user-independent evaluation, we used a leave-one-user-out cross-validation approach, where in each fold data from 43 subjects are used for training the model, and data from the remaining subject is used for testing the models. To understand the effect of the user-screen distance and head movement on the PCA based model, we train separate models for each distance ("Near", "Medium", "Far" and "User-defined") and each head movement (with and without head movement) condition by considering only the training data collected during the corresponding setting. Then, we test the model on two sets of data (with and without head movement) where the distance is defined by the user. Following our previous study where we showed that the eigenvectors associated with the largest 30 eigenvalues captures over 90% of the variability in the eye pair patches [15], we considered $p_1, p_2, \cdots, p_{30}$ to build linear regression models.

Table 2 provides the results for the PCA appearance-based framework. For user-dependent model, the matched distance condition ("User-defined" distance) provides the best performance, as expected. By comparing the performance across different distance settings, we observe that it affects the performance when the distance is too far or too close to the computer screen. When the user is too close to the monitor, both the head movement and the eye movement are involved causing large variance on the eye pair image. When the user is too far from the monitor, the eye appearance change is less obvious, thus PCA

Table 2: Evaluation of gaze estimation using the PCA approach. The training conditions consist of various distance setting and head movement constrains (wH - with head movement, w/oH - without head movement). The performance is measured with the angular error θ_{error} (°).

	Test Setting					
	User Dependent			**User Independent**		
Training Setting	**wH**	**w/oH**	**Avg**	**wH**	**w/oH**	**Avg**
Near (wH)	8.1	9.2	8.6	10.0	10.1	10.0
Medium (wH)	5.7	6.3	6.0	9.1	9.1	9.1
Far (wH)	7.1	6.8	6.9	9.4	9.4	9.4
User-defined (wH)	4.9	6.8	5.9	9.2	9.2	9.2
All (wH)	4.9	5.6	5.3	9.9	10.0	10.0
Near (w/oH)	9.2	8.3	8.8	10.9	10.3	10.6
Medium (w/oH)	7.7	6.2	6.9	9.5	9.2	9.4
Far (w/oH)	8.0	6.8	7.4	10.0	9.7	9.9
User-defined (w/oH)	7.5	4.6	5.4	10.6	10.3	10.4
All (w/oH)	6.3	4.6	5.4	10.6	10.3	10.4
All	5.0	4.6	4.8	11.7	11.8	11.7

fails to capture the details. To study the effect of head movement, we compare the performance on the test data with and without head motion. We found that matched conditions provide better results. When the training data only consists of natural gazing behavior (without head movement constrains), the performance is better on the test data under the same condition. These results suggest that the distance and head movement affect the eye-pair appearance, changing the PCA representation of the user-dependent gaze behavior. The best performance is achieved when the training data includes all training data across different conditions. This result suggests that when all training data are combined, the major variation in the training data is the gaze direction and the effect of other factors is reduced.

We observe very different results when we train with user-independent mod-

els. First, the performance significantly drops when using the PCA-based approach with a user-independent setting. Second, the effect of distance and head movement is less important when compared to user-dependent models. Third, including all training data does not improve the performance. On the contrary, the performance drops when using the data from all subjects. We believe these observations can be explained by the variability among different individuals. For user-independent model, the main challenge is the individual differences. The variation across different subject dominates the variance of the training data. PCA captures the overall variability in the data, regardless of the factors responsible for it. Therefore, the approach is less effective in user-independent condition where users' differences account for most of the variability. This also explains the reduced effect of head movements and user-screen distances.

To reduce the gap between user-dependent and user-independent performance, we consider the tensor analysis framework.

5. Proposed User-Independent Gaze Estimation Framework

This section describes the proposed framework for appearance-based gaze estimation to reduce the gap in performance between user-dependent and user-independent conditions. The approach relies on tensor analysis (Section 5.1), and LASSO regression (Section 5.2). It also presents the domain adaptation scheme proposed to reduce user dependent bias (section 5.3).

5.1. Tensor Analysis Framework

The tensor analysis framework offers a natural approach to modeling the multi-factor nature of the eye appearance image ensembles. Several algorithms based on tensor representation have been proposed for various problems [40, 41, 42]. A tensor is a multidimensional array. The order of a tensor is the number of dimensions (modes) of the array. For example, a first order tensor is a vector and a second order tensor is a matrix. Here, we denote a $N-th$ order tensor $\mathcal{A}$ as $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. The matricization of a tensor is the process of converting

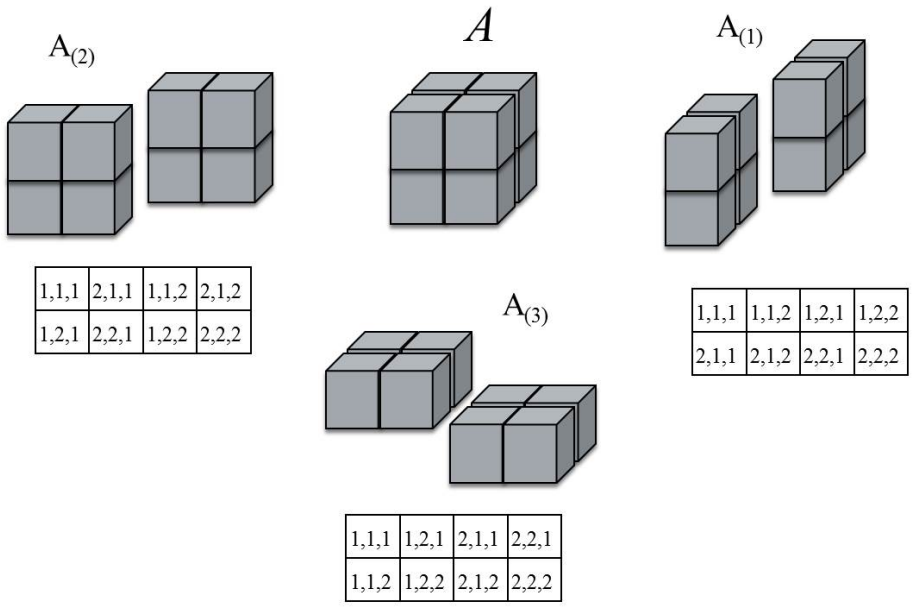

Figure 3: Tensor matricization. Tensor $\mathcal{A}$ can be converted into 3 matrices: $\mathcal{A}_{(1)}$, $\mathcal{A}_{(2)}$ and $\mathcal{A}_{(3)}$.

a tensor into a matrix, also known as unfolding or flattening. For example a $5 \times 2 \times 3$ tensor can be converted into either a 10×3 or a 5×6 matrix. The mode-n matricization of a tensor $\mathcal{A}$, denoted as $\mathcal{A}_{(n)}$, is a map between the tensor element $(i_1, i_2, \ldots, i_N)$ and the matrix element (i_n, j), where j is defined as

$$j = 1 + \sum_{k=1, k\neq n}^{N} \left((i_k - 1) \prod_{m=1, m\neq n}^{k-1} I_m \right) \tag{7}$$

where I_m stands for the order of each dimension. Figure 3 illustrates a simple interpretation of the above equation. For a given $3rd$ order tensor $\mathcal{A}$, it can be converted into 3 different matrices $\mathcal{A}_{(1)}$, $\mathcal{A}_{(2)}$ and $\mathcal{A}_{(3)}$, as indexed by its mode-n matricization.

Using the definition of matricization, the mode-n product of a tensor $\mathcal{A}$ and a matrix X is defined as

$$\mathcal{B} = \mathcal{A} \times_n X \tag{8}$$

resulting in a new tensor $\mathcal{B}$ where $\mathcal{B}_{(n)} = X\mathcal{A}_{(n)}$. Therefore the eigenvalue de-

composition presented in Equation 1 can be rewritten using the mode-n product as

$$\Sigma = \Lambda \times_1 U \times_2 U \tag{9}$$

As a natural extension of PCA for matrix, Tucker decomposition of the tensor is a form of high-order PCA. It decomposes a tensor into a core tensor, $\mathcal{C}$, multiplied by a matrix along each mode. For example, a 3 order tensor $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$can be decomposed into

$$\mathcal{A} = \mathcal{C} \times_1 P \times_2 Q \times_3 R \tag{10}$$

where $P \in \mathbb{R}^{I \times X}$, $Q \in \mathbb{R}^{J \times Y}$, and $R \in \mathbb{R}^{K \times Z}$ are the factor matrices and can be regarded as principal components in each mode. $\mathcal{C}$ is the *core tensor* that represents the interaction between different components.

The basic idea to compute the Tucker decomposition is to find the components that best capture the variation in mode n, independent of the other modes. This approach is also known as *high-order* SVD (HOSVD). The details of the method to decompose a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ are described as follow:

1. For $n = 1, \ldots, N$, compute the SVD of the mode-n matricization of a tensor $\mathcal{A}$, and assign the leading left singular vectors to U_n.
2. Repeat the above steps for all U_n.
3. Solve the core tensor $\mathcal{C}$ using the following equation:

$$\mathcal{C} = \mathcal{A} \times_1 U_1^T \times_2 U_2^T \times \cdots \times_N U_N^T \tag{11}$$

As discussed earlier, the eye appearance can be affected by the gaze directions (23 grids) - U_g, user-screen distances (4 distance settings) - U_d, head movements (with and without) - U_h, session differences (2 sessions of data collection) - U_s and individual appearance differences (44 subjects) - U_{pe}. By vectorizing the detected eye pair images (25 $\times$ 100 pixels) - U_{pi}, we formulate the eye appearance image tensor $\mathcal{D}$ as a 6-dimensional tensor (23 $\times$ 4 $\times$ 2$\times$ 2 $\times$ 44 $\times$ 2500).

The HOSVD decomposition of this tensor yields

$$\mathcal{D} = \mathcal{C} \times_1 U_g \times_2 U_d \times_3 U_h \times_4 U_s \times_5 U_{pe} \times_6 U_{pi} \tag{12}$$

where $\mathcal{C}$ dictates the interaction between the 6 factors. The 23×23 matrix U_g spans the space of different gaze directions. The 4×4 matrix U_d spans the space of different user-screen distance. Similarly, the matrices U_h, U_s, U_{pe} and U_{pi} span their corresponding factor spaces (head movement, session, people, image).

The n-mode product of a tensor with a matrix is related to a change of basis in the case where the tensor defines a multilinear operator, analog to PCA. The product of $\mathcal{C} \times_6 U_{pi}$ transforms the eigenimages in matrix U_{pi} into the principal axes of variation across the various modes (gaze directions, user-screen distances, head movements, session differences and individual appearance differences). By multiplying the core tensor with the parameter matrices $\mathcal{G} = \mathcal{C} \times_2 U_d \times_3 U_h \times_4 U_s \times_5 U_{pe} \times_6 U_{pi}$, we can obtain the principal axes of variation of the eye appearance image ensemble for each gaze direction across user-screen distance, individual, session and head movement factors.

5.2. LASSO Regression

The principal axes capture the variation across different gaze direction and are potentially useful for detecting user-independent gaze. For a particular setting (e.g. "Near", Subject 4, Session 1, Without head movement), the sub tensor $\mathcal{G}_{d,h,s,pe}$ of dimension $23 \times 1 \times 1 \times 1 \times 1 \times 2500$ can be flatten along the gaze direction mode to obtain a 23×2500 matrix providing 23 principal components. As a result, the proposed tensor framework provides $23 \times 4 \times 2 \times 2 \times 44 = 16192$ principal components. Directly applying linear regression using the projection to these axes can easily lead to over-fitting due to the overwhelming number of independent variables. Unfortunately, unlike SVD where the most important principle components are decided by their corresponding eigenvalues, it is difficult to identify important principle components among them. To solve this

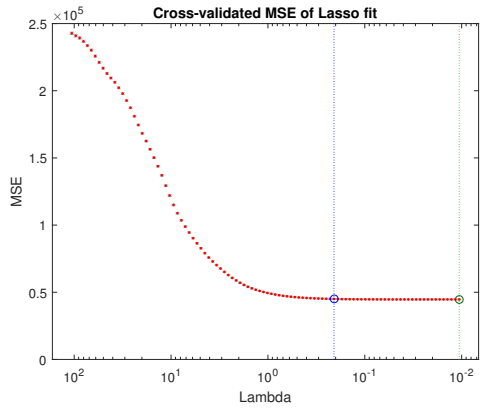

Figure 4: The effect of the value of λ in the LASSO model (Equation 13). The green line locates the value of λ with minimum cross-validation error. The blue line locates the value of λ with minimum cross-validation error plus one standard deviation.

problem, we applied LASSO regression instead of the original linear regression [43].

LASSO regression is a regularized linear regression method with (l_1) optimization. Given N observations $\{(x_1, y_1) \dots (x_N, y_N)\}$, it solves the following equation for the intercept and coefficient vector:

$$min_{\beta_0,\overline{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \overline{\beta}\right) + \lambda \sum_{j=1}^{n} |\beta_j| \right\} \tag{13}$$

LASSO regression provides a systematic solution to address the difficulties in finding good principal components from tensor analysis. Due to (l_1) regularization, it chooses few coefficients as indicated by the nonzero coefficient in the regression model. This is especially helpful when the independent variables are highly correlated (i.e. gaze variations among different subjects). The value of λ in Equation 13 dictates the number of features in the model. To understand the effect of this parameter in the LASSO regression model, we evaluate the *mean square error* (MSE) as a function of λ. We estimate this evaluation using training set using tenfold cross-validation. Figure 4 illustrates the results. The performance is very stable for a wide range of values for λ. The minimum

Figure 5: Tensor components with the highest coefficients selected by the LASSO model.

cross-validation error is marked with a green lane. We select the value of λ such that the degree of freedom of the LASSO regression model is equivalent to the number of independent variables of the tensor model implemented using the user-dependent condition (without U_{pe}, i.e., $23 \times 4 \times 2 \times 2 = 368$). We highlight this value in Figure 4 with a blue lane.

We visualize the tensor components selected by the LASSO framework to get a deeper understanding of the eye representation created by the proposed tensor-based approach, Figure 5 describes the results for the three most important components. They represent the most important factors for user-independent gaze estimation. These principal components capture the structure of the eyes, showing wide eye appearance variations and nose angles. For example, the figure in the middle includes angular variations which we hypothesize are useful for head orientation.

We report the results using the leave-one-user-out cross-validation approach in Table 3. For each fold, we used all the data from 43 subjects to form the eye appearance tensor and build the regression model with LASSO to predict the gaze of the remaining subject. On average, the approach achieves an angular error of 8.9°, which is 2.8° better than the performance of the PCA approach.

Due to its ability to capture multiple factors of the image ensemble, we also applied the tensor analysis for the user-dependent case where equation 12 is modified as follow, removing the individual difference factor.

$$\mathcal{D} = \mathcal{C} \times_1 U_g \times_2 U_d \times_3 U_h \times_4 U_s \times_5 U_{pi} \tag{14}$$

As shown in Table 3, we achieve over 1° improvement over the original PCA approach for user-dependent gaze estimation. This suggests that applying tensor

Table 3: Evaluation of gaze estimation using tensor analysis and LASSO regression. The training setting consists of two head movement constraints (wH - with head movement, w/oH - without head movement). The performance is measured with angular error θ_{error} (°). The standard deviation of the results are reported in brackets.

	UD			UI		
Approach	**wH**	**oH**	**Avg**	**wH**	**oH**	**Avg**
PCA	5.0 (2.0)	4.6 (1.8)	4.8 (1.9)	11.7 (3.0)	11.8 (3.4)	11.7 (3.2)
Tensor	3.9 (1.9)	3.1 (1.3)	3.5 (1.6)	8.9 (2.9)	8.9 (3.2)	8.9 (3.0)

framework for user-dependent model can significantly reduce the user-dependent factors such as head movements, session differences and user-screen distances.

Overall, the proposed tensor-based framework models the underlying factors affecting the eye appearance space. It provides a solution for modeling the complicated joint effect of factors existing in user gaze estimation for both user-dependent and user-independent applications, as illustrated by the improved results.

5.3. Domain adaptation

In addition to the angular error, we also calculate the correlation between the predicted gaze position and the ground truth gaze position while evaluating the performance. Higher correlation indicates that the proposed model can capture the changes in the gaze direction. For the user-independent model, we observed high correlation of the proposed tensor analysis framework. It achieves correlation of $\rho_x = 0.91$ for the horizontal prediction, and $\rho_y = 0.79$ for the vertical prediction.

To understand the performance of the system, Figure 6 shows the prediction and the ground truth of the gaze positions. When the actual gaze direction changes, we observe clear changes in the prediction following the same direction. This result explains the high correlation between the prediction and the ground truth. More importantly, the result in Figure 6 reveals a shift between the prediction and the actual gaze position. We believe that this shift is due to the differences in individual heights or sitting positions, as mentioned earlier.

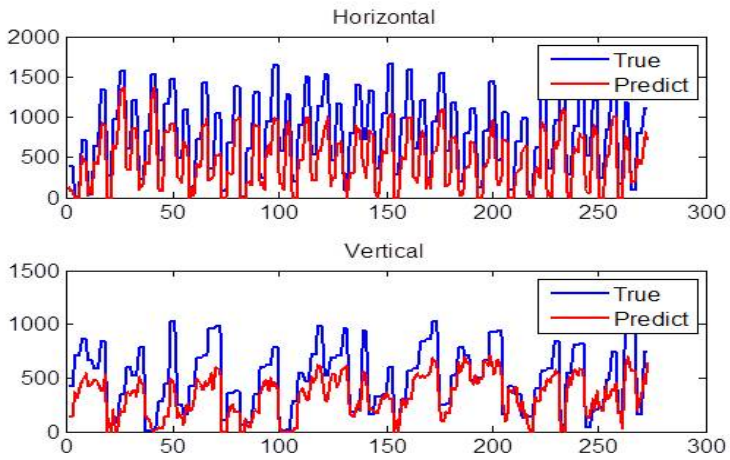

Figure 6: The shift of the gaze prediction due to individual difference.

For example, when the test subject does not sit along the center of the monitor, his/her eye appearance is different from when he/she looks at the monitor center causing the shift of the overall prediction.

We address this problem using domain adaptation (mean and variance). We make the common assumption that users tend to look at the center of the screen most of the time [30, 31]. Therefore, the average of the prediction should be around the center of the monitor. In addition, we also assume that the gaze variance is similar across different subjects. The proposed approach consists of applying the mean and standard deviation adaptation such that the predicted gaze value has similar statistics as the training data. We used the center position of the monitor ($\mu_x = 840, \mu_y = 525$) as the target mean value, we used the variance calculated in the training data as the target variance value ($\sigma_x = 480, \sigma_y = 300$). As a result, the mean and variance adaptation is achieved by the following equations, resulting in the updated position $(\hat{x_p}, \hat{y_p})$.

$$\begin{cases} \hat{x_p} = 840 + (x_p - \mu_{x_p}) \times \frac{\sigma_x}{\sigma_{x_p}} \\ \hat{y_p} = 525 + (y_p - \mu_{y_p}) \times \frac{\sigma_y}{\sigma_{y_p}} \end{cases} \tag{15}$$

where x_p and y_p are the original predicted gaze positions, μ_{x_p} and μ_{x_p} are

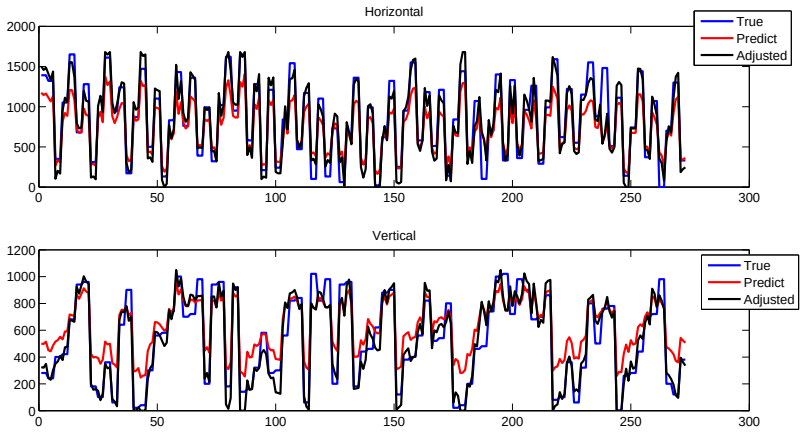

Figure 7: The shift of the gaze prediction due to individual difference are corrected by the proposed domain adaptation.

the mean of the predicted gaze position, and σ_{x_p} and σ_{y_p} are the standard deviation of the predicted gaze positions. In practical applications, the domain adaption calculation can be performed over a fixed length window to update the adaptation parameters. This allows the subject to slightly move near their current position, providing more flexibility to the users. Figure 7 shows the adapted prediction for one of the test data where we used the total test data (273 frames) to adapt the prediction. The result shows that the adaptation method reduces the shift effect, achieving better predictions.

Table 4 shows the overall result using domain adaptation applied to the PCA and tensor analysis framework. As before, we evaluate the approach using leave-one-user-out cross validation. We observe improvement in the performance suggesting that domain adaptation is important to address individual differences. The best performance is achieved with tensor analysis framework with 6.6° angular error. This method reduces more than 40% of the angular error observed from the user-independent PCA-based approach (11.7°). These values are only few degrees worse than the ones achieved with user-dependent PCA method (see Table 2).

Table 4: Evaluation of domain adaptation on the PCA, and the tensor analysis framework results. The performance is measured by angular error θ_{error} (°). The standard deviation of the results are reported in brackets.

	Original			**Domain Adaptation**		
Approach	**wH**	**oH**	**Avg**	**wH**	**oH**	**Avg**
PCA	11.7 (3.0)	11.8 (3.4)	11.7 (3.2)	8.2 (3.0)	7.4 (2.7)	7.8 (2.8)
Tensor	8.9 (2.9)	8.9 (3.2)	8.9 (3.0)	6.7 (2.3)	6.6 (2.4)	6.6 (2.4)

One important parameter for the proposed domain adaptation method is the amount of test data required to estimate the mean and standard deviation. The assumption about user looking at the center of the screen most of the time is more reliable when we use more test data. However, it also implies longer delay for practical usage. In Table 4, we considered all the test data (273 frames) to perform the adaptation. Figure 8 shows the effect of the window size on the proposed adaptation. The performance improves as the window size increases. We notice that most of the improvement can be achieved with 50 frames of test data. We highlight that test data does not need to be labeled. This result suggests that domain adaptation can be applied with limited test data. For practical applications, we can start without adaptation. As more predictions are made, we can update the estimation of the mean and standard deviation to improve the performance.

5.4. Comparison to Other Methods

To illustrate the effectiveness of the proposed approach, we compare our method with a related approach introduced in Section 2.2 proposed by Lu et al. [19] using the ALR model. In their work, they divided a single eye image into 3 by 5 grids, and used the summation of the pixel intensity value in each grid as eye image features. We follow the same approach, with the exception that we consider a 3 by 10 grid, since we consider eye-pair images instead of single eye images, as originally done in Lu et al. [19].

Table 5 shows the comparison between our results and the ALR approach.

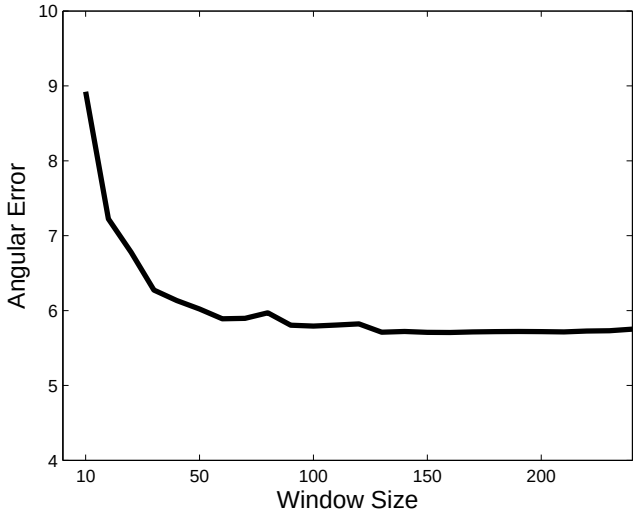

Figure 8: The effect of window size for the domain adaptation.

One difference between our experiment is that we use the image itself for the evaluation. Considering that the feature extraction proposed by Lu et al. [19] may reduce the eye appearance information and reduce the effectiveness of the l_1 optimization result, we also evaluate their approach using the image itself. As we expected, the l_1 optimization approach using the entire image shows better performance, especially after domain adaptation. However, the proposed tensor framework outperforms the ALR approach. We evaluate whether the differences in performances are statistically significant using pairwise t-tests between the conditions (tensor-based framework, ASL with feature, ASL with images). We assert significance at p-value $= 0.05$. The results indicate that the proposed tensor-based method is significantly better than the ASL method implemented with the feature extraction framework proposed by Lu et al. [19]. The differences between our method and the ASL method implemented using the entire image is not significant. More importantly, the superior performance after the adaptation step indicates that the proposed approach can capture the variabilities in the data better than the ALR approach.

Table 5: Comparison with the appearance based gaze estimation proposed by Lu et al. [19]. The performance is measured with the angular error θ_{error} (°). The standard deviation is given in the brackets.

	Original			**Domain Adaptation**		
Approach	**wH**	**oH**	**Avg**	**wH**	**oH**	**Avg**
Tensor	8.9 (2.9)	8.9 (3.2)	8.9 (3.0)	6.7 (2.3)	6.6 (2.4)	6.6 (2.4)
ASL - feature [19]	11.4 (1.8)	11.9 (2.2)	11.7 (2.0)	9.2 (2.1)	9.3 (2.5)	9.3 (2.3)
ASL - image [19]	11.8 (2.2)	11.8 (2.4)	11.8 (2.3)	7.4 (2.5)	7.4 (2.5)	7.4 (2.5)

Figure 9 illustrates the best performance in user independent gaze estimation in terms of pixel error between ground truth and predicted gaze position in the screen. We plot an eclipse around the ground truth gaze position (cross) where its major axis corresponds to the average horizontal pixel error and the minor axis corresponds to the average vertical pixel error. This error is small enough for many applications. Depending on the application, the gaze-aware multimodal interfaces may not require high gaze estimation accuracy, especially when the focus is on simple commands. In these cases, commands can be triggered by detecting gaze in coarse areas on the screen. The proposed approach implements a gaze estimation approach that, while may produce lower accuracy than commercially available systems, (1) it is robust against head motion and individual differences, and (2) it requires no calibration. Such a system will be suitable for gaze-aware multimodal interfaces that require non-intrusive sensors, and for non-cooperative users.

6. Conclusions

The paper presented our efforts to design a calibration free, user-independent gaze estimation framework that does not require cooperation from the user. This study built upon our previous studies on PCA based appearance model to explore different strategies to reduce the gap in performance between user-dependent and user-independent gaze estimation. We found that the performance drops for user-independent gaze estimation due to various factors, such

Figure 9: An illustration of the gaze estimation performance on the 22-inch HP monitor. The cross center indicate the ground true while the eclipse indicate the average pixel error on the prediction. The monitor screen resolution is set to 1920 by 1080. The average horizontal and vertical gaze estimation error by the proposed approach is 110 by 80 in terms of pixel error.

as individual eye appearance differences, head movements, and various distances between the user and the interface's screen. The variation in these factors affects the performance of the PCA approach. To address this problem, we proposed a tensor analysis framework to model the underlying joint effect among the factors including individual differences, head movements, session differences, and user-screen distances. The results suggest that the tensor framework effectively models the multiple factors affecting the system, improving the performance. We noticed a shift in the gaze prediction due to differences in individual heights and sitting positions. We proposed a domain adaptation method to reduce this effect, achieving the best performance of 6.6° angular error for user-independent gaze estimation, which is only 1.8° worse than the user-dependent PCA models.

The center-biased assumption plays an important role for domain adaptation. In real-world scenarios, user gaze behavior depends on the task. For example, the gaze of the user could be concentrated at the center of the screen when watching a movie. In other cases such as working with multiple windows, the variance of the gaze may be higher. However, there are strong reasons to

believe that the central area of the screen concentrates most of the gaze directions. Judd et al.[31] studied salience maps by collecting gaze of people looking at different randomly selected images. The results suggested an important bias toward the center of the images. Zhao and Koch [30] presented similar results. Therefore, we believe that the assumption is reasonable. The train and test data patterns followed in the recordings of the MSP-GAZE dataset represent the worst-case scenario for this assumption, where the participants are asked to randomly look at different points in the screen. Furthermore, the train and test sessions follow different protocols to avoid bias. In the train sessions, we only consider selected regions of the screen. In the test sessions, the points are randomly displayed on the screen. In spite of using the worse case scenario with train-test mismatch, the adaptation approach is very effective, improving the performance of the system.

The MSP-GAZE corpus contains data collected from both a webcam and Kinect sensor. Although the study only considered the webcam images, the Kinect sensor provides additional RGB image and depth information. The RGB image is captured from a different angle and can be used to improve the robustness of the captured eye pair appearance. This information can be helpful for the vertical gaze estimation since the Kinect sensor is placed below the screen. The depth information can be used to estimate the head pose, and the distance between the monitor and the participant. In addition, we can also include the mouse click and mouse movement information to update the gaze prediction result. For the HCI, mouse click provides useful information about a user's visual attention. Likewise, this study only considers frame-by-frame gaze estimation. One future direction to enhance the performance is to incorporate a tracking algorithm. This approach can reduce the errors and provide smooth gaze estimation trajectories.

A limitation of the data collection is that we only recorded subjects without glasses. Also, we relied on good illumination conditions similar to conditions of regular office environment. These conditions are not realistic in many real applications, and robust methods are needed. Our future data collection ef-

fort will include these different challenging conditions. In spite of these challenges, appearance-based models offer a good trade-off between accuracy and complexity for HCI. The MSP-GAZE corpus and the proposed methods can be used as building blocks to address these open challenges towards robust gaze aware interfaces. For this purpose, we intend to release the corpus (`http://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-gaze.html`).

References

[1] A. Duchowski, Eye tracking methodology: Theory and practice, Springer-Verlag, London, UK, 2007.

[2] K. Rayner, C. Rotello, A. Stewart, J. Keir, S. Duffy, Integrating text and pictorial information: eye movements when looking at print advertisements, Journal of Experimental Psychology: Applied 7 (3) (2001) 219–226. doi:10.1037/1076-898X.7.3.219.

[3] N. Li, J. Jain, C. Busso, Modeling of driver behavior in real world scenarios using multiple noninvasive sensors, IEEE Transactions on Multimedia 15 (5) (2013) 1213–1225. doi:10.1109/TMM.2013.2241416.

[4] D. Salvucci, J. Anderson, Intelligent gaze-added interfaces, in: SIGCHI conference on Human Factors in Computing Systems (CHI 2000), The Hague, The Netherlands, 2000, pp. 273–280. doi:10.1145/332040.332444.

[5] H. Skovsgaard, J. Mateo, J. Hansen, Evaluating gaze-based interface tools to facilitate point-and-select tasks with small targets, Behaviour & Information Technology 30 (6) (2011) 821–831. doi:10.1080/0144929X.2011.563801.

[6] D. Hansen, Q. Ji, In the eye of the beholder: A survey of models for eyes and gaze, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (3) (478-500) 478–500. doi:10.1109/TPAMI.2009.30.

[7] E. Guestrin, M. Eizenman, General theory of remote gaze estimation using the pupil center and corneal reflections, IEEE Transactions on Biomedical Engineering 53 (6) (2006) 1124–1133. doi:10.1109/TBME.2005.863952.

[8] T. Ohno, N. Mukawa, A. Yoshikawa, FreeGaze: a gaze tracking system for everyday gaze interaction, in: Symposium on Eye tracking research & applications (ETRA 2002), New Orleans, LA, USA, 2002, pp. 125–132. doi:10.1145/507072.507098.

[9] D. Scott, J. Findlay, Visual Search, Eye Movements and Display Units, IBM UK Hursley Human Factors Laboratory, 1991.

[10] T. Ohno, N. Mukawa, A free-head, simple calibration, gaze tracking system that enables gaze-based interaction, in: Symposium on Eye tracking research & applications (ETRA 2004), San Antonio, TX, USA, 2004, pp. 115–122. doi:10.1145/968363.968387.

[11] B. Noureddin, P. D. Lawrence, C. F. Man, A non-contact device for tracking gaze in a human computer interface, Computer Vision and Image Understanding 98 (1) (2005) 52–82. doi:10.1016/j.cviu.2004.07.005.

[12] D. Beymer, M. Flickner, Eye gaze tracking using an active stereo head, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), Vol. 2, Madison, WI, USA, 2003, pp. 451–458. doi:10.1109/CVPR.2003.1211502.

[13] F. Lu, T. Okabe, Y. Sugano, Y. Sato, A head pose-free approach for appearance-based gaze estimation, in: British Machine Vision Conference (BMVC 2011), Dundee, Scotland, 2011, pp. 126.1–126.11. doi:10.5244/C.25.126.

[14] Y. Sugano, Y. Matsushita, Y. Sato, H. Koike, An incremental learning method for unconstrained gaze estimation, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision (ECCV 2008), Vol. 5304 of Lecture Notes

in Computer Science, Springer Berlin Heidelberg, Marseille, France, 2008, pp. 656–667. doi:10.1007/978-3-540-88690-7_49.

[15] N. Li, C. Busso, Evaluating the robustness of an appearance-based gaze estimation method for multimodal interfaces, in: International conference on multimodal interaction (ICMI 2013), Sydney, Australia, 2013, pp. 91–98. doi:10.1145/2522848.2522876.

[16] E. Pogalin, A. Redert, I. Patras, E. Hendriks, Gaze tracking by using factorized likelihoods particle filtering and stereo vision, in: International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT 2006), Chapel Hill, NC, USA, 2006, pp. 57–64. doi:10.1109/3DPVT.2006.66.

[17] J. Xie, X. Lin, Gaze direction estimation based on natural head movements, in: International Conference on Image and Graphics (ICIG 2007), Chengdu, Sichuan, China, 2007, pp. 672–677. doi:10.1109/ICIG.2007.160.

[18] B. Noris, J. B. Keller, A. Billard, A wearable gaze tracking system for children in unconstrained environments, Computer Vision and Image Understanding 115 (4) (2011) 476–486. doi:10.1016/j.cviu.2010.11.013.

[19] F. Lu, Y. Sugano, T. Okabe, Y. Sato, Inferring human gaze from appearance via adaptive linear regression, in: International Conference on Computer Vision (ICCV 2011), Vol. Barcelona, Spain, November, 2011, pp. 153–160. doi:10.1109/ICCV.2011.6126237.

[20] O. Williams, A. Blake, R. Cipolla, Sparse and semi-supervised visual mapping with the S^3GP, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), Vol. 1, New York, NY, USA, 2006, pp. 230–237. doi:10.1109/CVPR.2006.285.

[21] D. Xia, Z. Ruan, IR image based eye gaze estimation, in: ACIS International Conference on Software Engineering, Artificial Intelligence, Network-

ing, and Parallel/Distributed Computing (SNPD 2007), Vol. 1, Qingdao, China, 2007, pp. 220–224. doi:10.1109/SNPD.2007.237.

[22] C. Colombo, D. Comanducci, A. Del Bimbo, Robust tracking and remapping of eye appearance with passive computer vision, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 3 (4) (2007) 20.1–20.20. doi:10.1145/1314303.1314305.

[23] J. Wang, L. Yin, J. Moore, Using geometric properties of topographic manifold to detect and track eyes for human-computer interaction, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 3 (4) (2007) 3. doi:10.1145/1314303.1314306.

[24] H. Wu, Y. Kitagawa, T. Wada, T. Kato, Q. Chen, Tracking iris contour with a 3D eye-model for gaze estimation, in: Y. Yagi, S. Kang, I. Kweon, H. Zha (Eds.), Computer Vision - ACCV 2007, Vol. 4843 of Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, Tokyo, Japan, 2006, pp. 688–697. doi:10.1007/978-3-540-76386-4_65.

[25] H. Yamazoe, A. Utsumi, T. Yonezawa, S. Abe, Remote and head-motion-free gaze tracking for real environments with automated head-eye model calibrations, in: IEEE CVPR Workshop Human Communicative Behavior Analysis (CVPR4HB), Anchorage, Alaska, 2008, pp. 1–6. doi:10.1109/CVPRW.2008.4563184.

[26] S. Baluja, D. Pomerleau, Non-intrusive gaze tracking using artificial neural networks, Tech. Rep. CMU-CS-94-102, Carnegie Mellon University, Pittsburgh, PA, USA (January 1994).

[27] K. Tan, D. Kriegman, N. Ahuja, Appearance-based eye gaze estimation, in: IEEE Workshop on Applications of Computer Vision (WACV 2002), Orlando, FL, USA, 2002, pp. 191–195. doi:10.1109/ACV.2002.1182180.

[28] B. Schiele, A. Waibel, Gaze tracking based on face-color, in: International

Workshop on Automatic Face-and Gesture-Recognition, Zurich, Switzerland, 1995, pp. 344–349.

[29] T. Rikert, M. J. Jones, Gaze estimation using morphable models, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG 1998), Nara, Japan, 1998, pp. 436–441. doi:10.1109/AFGR.1998.670987.

[30] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, Journal of vision 11 (9) (2011) 1–15. doi:10.1167/11.3.9.

[31] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: International Conference on Computer Vision (ICCV 2009), Vol. Kyoto, Japan, 2009, pp. 2106–2113. doi:10.1109/ICCV.2009.5459462.

[32] N. Li, C. Busso, User-independent gaze estimation by exploiting similarity measures in the eye pair appearance eigenspace, in: International conference on multimodal interaction (ICMI 2014), Istanbul, Turkey, 2014, pp. 335–338. doi:10.1145/2663204.2663250.

[33] Y. Ono, T. Okabe, Y. Sato, Gaze estimation from low resolution images, in: L.-W. Chang, W.-N. Lie, R. Chiang (Eds.), Advances in Image and Video Technology, Vol. 4319 of Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, Hsinchu, Taiwan, 2006, pp. 178–188. doi:10.1007/11949534_18.

[34] T. Proscevičius, V. Raudonis, A. Kairys, A. Lipnickas, R. Simutis, Autoassociative gaze tracking system based on artificial intelligence, Electronics and Electrical Engineering 101 (5) (2010) 67–72.

[35] M. Castrillón, O. Déniz, C. Guerra, M. Hernández, ENCARA2: Real-time detection of multiple faces at different resolutions in video streams, Journal of Visual Communication and Image Representation 18 (2) (2007) 130–140. doi:10.1016/j.jvcir.2006.11.004.

[36] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of cognitive neuroscience 3 (1) (1991) 71–86. doi:10.1162/jocn.1991.3.1.71.

[37] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720. doi:10.1109/34.598228.

[38] P. Hillman, J. Hannah, P. Grant, Global fitting of a facial model to facial features for model-based video coding, in: International Symposium on Image and Signal Processing and Analysis (ISPA 2003), Vol. 1, Rome, Italy, 2003, pp. 359–364. doi:10.1109/ISPA.2003.1296923.

[39] W. Huang, R. Mariani, Face detection and precise eyes location, in: International Conference on Pattern Recognition (ICPR 2000), Vol. 4, Barcelona, Spain, 2000, pp. 722–727. doi:10.1109/ICPR.2000.903019.

[40] L. Wolf, H. Jhuang, T. Hazan, Modeling appearances with low-rank SVM, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, Minnesota, USA, 2007, pp. 1–6. doi:10.1109/CVPR.2007.383099.

[41] H. Pirsiavash, D. Ramanan, C. C. Fowlkes, Bilinear classifiers for visual recognition, in: Advances in neural information processing systems (NIPS 2009), Vancouver, BC, Canada, 2009, pp. 1482–1490.

[42] H. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, MPCA: Multilinear principal component analysis of tensor objects, IEEE Transactions on Neural Networks 19 (1) (2008) 18–39. doi:10.1109/TNN.2007.901277.

[43] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 267–288.