A New Reconstruction Method in Gaze Estimation with Natural Head Movement

Yi Liu, Bu-Sung Lee
School of Computer Science and Engineering
Nanyang Technological University, Singapore
{yliu028, ebslee}@ntu.edu.sg

Martin McKeown
Department of Medicine
University of British Columbia, Canada
martin.mckeown@ubc.ca

Abstract

We present a novel reconstruction method for the appearance-based gaze estimation that allows inferring persons' gaze under natural head movement. We first study that the locally linear combination in the respective manifolds consisting of stable left and right eye appearances is efficient. The local structure of the manifolds is destroyed when there is head movement. This is due to the destruction of the intrinsic relations between the two eyes(left and right) when we do locally linear combinations. We then introduce a new combination of both eye appearances, which maintains the relation embedding into the reconstruction of the training stage. Through comparison with other well known methods, we show that the proposed method achieves an optimal performance with head pose variation.

1 Introduction

Eyes and their movements play an important role in expressing human beings desires, cognitive processes, emotion states, and interpersonal relations. Eye tracking has been applied in various applications such as human-computer interaction, marketing and advertisement, human behavior analysis, and go beyond that to aid disable people [7]. However, the main technology of eye-tracking applications, gaze estimation still performs poorly in natural environments, e.g. the additional eye-tracking device is necessary, and head movement must be limited to small degree [9]. These limitations prevent eye tracking from becoming a pervasive technology.

As webcam is becoming a standard component in computers, replacing the eye tracker with a webcam would simplify the setup of eye-tracking applications, and also be promising to facilitate the prevalence. In addition, with the development of computer vision technology in face detection and recognition, gaze estimation directly using original eye appearance images has been proven to be feasible with considerable performance accuracy [8].

However, gaze estimation with head movement is still a major research challenge. There is one fundamental problem that has yet to be studied extensively in existing appearance-based approaches that how to integrate the left and right eye images, and model the relation of the both appearance spaces. Researchers [11, 10, 8] usually consider the left and right eye images separately in the reconstruction space. Thus, losing the intrinsic relation properties that exist between the pair of eyes. In fact, the relation might compensate the impact of head pose variation. Without addressing this issue, existing appearance-based approaches show the limited performance under natural head movement/pose.

Therefore, in this paper we propose a new reconstruction method that aligns the two appearance spaces of both eyes into the same local structure. We then investigate the performance of different combinations of both eyes images. In addition, we also analyze the impact of degree of head pose variation. The experimental results show that our proposed method outperforms existing combination methods under natural head movement.

2 Related Work

In general, research in eye tracking mainly refers to two areas: *eye localization* and *gaze estimation* [5]. The main task of eye localization is detecting the existence of eyes, accurately interpreting the eye positions in the images, or tracking the eyes from frame to frame from video images [12, 15].

Instead of only localizing where the eyes are in face images, gaze estimation is determining what a person is looking at by detected eyes in images or videos. It can be either a gaze direction in the 3D space, or a gaze point that the intersection of gaze direction and a 2D plane. Generally the two terms "eye tracking" and "gaze tracking" are inter-changeable, where gaze tracking is a process of consecutive gaze estimations from frame to frame in a video or real time.

Gaze estimation modelling focuses on the mapping from the eye-area image to the gaze direction/point. Typically the gaze estimation approaches are divided into three categories, 3D-model-based, feature-based, and appearance-based approaches [5, 3, 4, 11]. The 3D-model-based and feature-based approaches require highly accurate feature detection, and also a high-resolution camera and infrared light. Instead of explicit local feature extraction , appearance-based approaches focus on the whole eye image content as an input mapping to the gaze coordinate (or gaze direction), and the setup is more flexible, where a single webcam with relative low resolution is sufficient. Thus the appearance-based method is becoming a popular gaze estimation technique.

Appearance-based approaches directly take the eye-area image as a high dimensional vector (figure 1), and resemble the interpolation-based methods to train a regression function, such as multi-layer neural network [1], Gaussian process [13], and manifold learning [11, 10]. Recently, along with the prevalence of deep learning, some research work also attempted to introduce deep neural network into gaze estimation training with millions of eye images [14, 6].

In the paper, we mainly focus on a fundamental is-

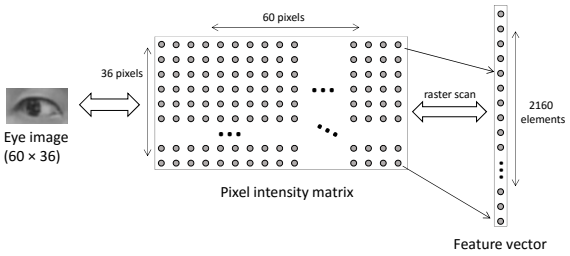


Figure 1: Eye appearance vector extraction.

sue in modelling the relation of both eyes images in manifold learning, so that the local linear combination in the two spaces can be aligned resulting in decreasing error during head movement.

3 Methodology

Appearance manifold is a continuous set of appearance feature points embedding in the high dimensional space, and any point in the manifold can be linearly interpolated by its local neighbours. The assumption of appearance-based approaches is that eye-appearance feature points constitute a manifold in the high dimensional space (the appearance space), and its relative local interpolation is maintained in the low dimensional space, which has been well verified already [10, 8].

Thus rather than global mapping directly from high-dimensional appearance data to the low-dimensional gaze coordinates, appearance-based gaze estimation is reconstructing the query appearance vector $\widehat{\boldsymbol{x}}$ by the local linear combination of its neighbouring appearance vectors $\{\boldsymbol{x}_i\}$ in the appearance space. Then it infers the corresponding gaze point $\widehat{\boldsymbol{p}}$ of $\widehat{\boldsymbol{x}}$ by using the same linear combination of corresponding gaze points $\{\boldsymbol{p}_i\}$ of $\{\boldsymbol{x}_i\}$ in the gaze-coordinate space:

$$\begin{aligned} \boldsymbol{X}\boldsymbol{w} &= \widehat{\boldsymbol{x}} \\ \boldsymbol{P}\boldsymbol{w} &= \widehat{\boldsymbol{p}} \end{aligned} \tag{1}$$

Where $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k\}$, $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_k\}$, and $\boldsymbol{w} = \{w_1, w_2, ..., w_k\}$ denotes the locally linear combination, which avoids directly estimating the mapping between $\boldsymbol{X}$ and $\boldsymbol{P}$. However, since the equation is overdetermined, it contains estimation errors between the locally linear combination $\boldsymbol{X}\boldsymbol{w}$ and the query appearance vector $\widehat{\boldsymbol{x}}$. Therefore, we have the optimization function with minimizing the estimation error:

$$\begin{aligned} \widetilde{\boldsymbol{w}} &= \arg\min \left\| \widehat{\boldsymbol{x}} - \sum_i^k w_i \boldsymbol{x}_i \right\| \;\; s.t. \sum_i^k w_i = 1 \\ \widehat{\boldsymbol{p}} &= \sum_i^k w_i \boldsymbol{p}_i \end{aligned} \tag{2}$$

Where $\widetilde{\boldsymbol{w}}$ is optimal locally linear combination parameter minimizing the linear combination error.

Then without loss of generality, let $\widehat{\boldsymbol{x}}^L$, $\widehat{\boldsymbol{x}}^R$ denote the left and right eye appearance vectors of the query face image; $\boldsymbol{X}^L$, $\boldsymbol{X}^R$ denote the neighbouring appearance vectors of $\widehat{\boldsymbol{x}}^L$ and $\widehat{\boldsymbol{x}}^R$; $\boldsymbol{w}^L$, $\boldsymbol{w}^R$ denote corresponding linear combination parameters. Thus, the actual reconstruction should be given by:

$$\begin{aligned} \widetilde{\boldsymbol{w}}^L &= \arg\min \left\| \widehat{\boldsymbol{x}}^L - \boldsymbol{X}^L \boldsymbol{w}^L \right\|, \; s.t. \; \mathbf{1}^T \boldsymbol{w}^L = 1 \\ \widetilde{\boldsymbol{w}}^R &= \arg\min \left\| \widehat{\boldsymbol{x}}^R - \boldsymbol{X}^R \boldsymbol{w}^R \right\|, \; s.t. \; \mathbf{1}^T \boldsymbol{w}^R = 1 \\ \widehat{\boldsymbol{p}}^L &= \boldsymbol{P}^L \widetilde{\boldsymbol{w}}^L, \widehat{\boldsymbol{p}}^R = \boldsymbol{P}^R \widetilde{\boldsymbol{w}}^R, \widehat{\boldsymbol{p}} = (\widehat{\boldsymbol{p}}^L + \widehat{\boldsymbol{p}}^R)/2 \end{aligned} \tag{3}$$

Where $\widehat{\boldsymbol{p}}$ is averaging estimated gaze points of both left and right linear combination. Note that $\boldsymbol{X}^L$ and $\boldsymbol{X}^R$ are in different sets, so the functions are optimized in respective spaces. Thus, let "Left"/"Right" denote the method using $\widehat{\boldsymbol{p}}^L/\widehat{\boldsymbol{p}}^R$ as the estimated point, and "Average" denote the method using average point $\widehat{\boldsymbol{p}}$.

In the stable environment, e.g. no head movement, invariant illumination, the appearances of both eyes are relatively stable, and the two manifolds constituted by respective appearances would be homogeneous in their respective spaces. If any left appearance is a neighbour of the left query appearance, its corresponding right appearance should be also the neighbour of the right query appearance, and vice versa. That means that the pair of neighbours ($\{\boldsymbol{x}_i^L\}$,$\{\boldsymbol{x}_i^R\}$) are from the same face image ($\{\boldsymbol{f}_i\}$), i.e. $\forall i, \boldsymbol{x}_i^L \& \boldsymbol{x}_i^R \in \boldsymbol{f}_i$. Thus $\boldsymbol{P}^L = \boldsymbol{P}^R$, and then linear combination parameters $\boldsymbol{w}^L$ and $\boldsymbol{w}^R$ are approximate to each other. The relation of left and right appearance vector is not necessary to be modelled.

However, under natural head movement, the appearance images would be much different while even looking at the same position under different head poses. Likewise, the two similar eye appearances might correspond to two different gaze points which are far away from each other. The left and right neighbouring appearances might not be from the same face images, which causes $\boldsymbol{P}^L \neq \boldsymbol{P}^R$. Thus potential head pose information described by the relation between the pair of left and right eyes is missing. The reconstruction error would be large using individual linear combination in respective spaces with head pose variation.

In order to maintain the relation between the pair of eyes, we should restrict that the neighboring appearances should be pairs of eyes from the same face image ($\boldsymbol{P}^L = \boldsymbol{P}^R$). In addition, the same local structure should be guaranteed, i.e. the same locally linear combination in the two appearance spaces:

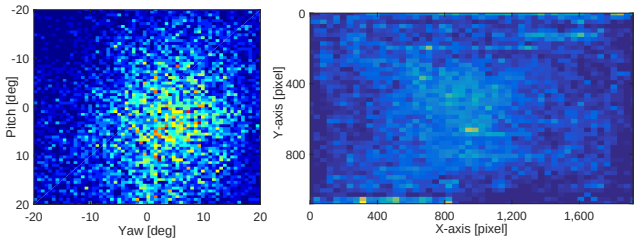
$$\begin{aligned} \widetilde{\boldsymbol{w}} = \arg\min \left\| \widehat{\boldsymbol{x}}^L - \boldsymbol{X}^L \boldsymbol{w} \right\| + \arg\min \left\| \widehat{\boldsymbol{x}}^R - \boldsymbol{X}^R \boldsymbol{w} \right\| \\ s.t. \quad \mathbf{1}^T \boldsymbol{w} = 1; \; \forall i, \boldsymbol{x}_i^L \& \boldsymbol{x}_i^R \in \boldsymbol{f}_i \\ \widehat{\boldsymbol{p}} = \boldsymbol{P}\widetilde{\boldsymbol{w}} \end{aligned} \tag{4}$$

Where $\boldsymbol{X}^L$ and $\boldsymbol{X}^R$ contain left and right eye-pair appearances of the same face images respectively, and the number of the neighboring appearances are the same. Thus the above equation(4) can be simplified to:

$$\begin{aligned} \widetilde{\boldsymbol{w}} = \arg\min \left\| \widehat{\boldsymbol{x}}^E - \boldsymbol{X}^E \boldsymbol{w} \right\| \quad s.t. \; \mathbf{1}^T \boldsymbol{w} = 1; \\ \widehat{\boldsymbol{p}} = \boldsymbol{P}\widetilde{\boldsymbol{w}} \end{aligned} \tag{5}$$

Where $\widehat{\boldsymbol{x}}^E = \begin{pmatrix} \widehat{\boldsymbol{x}}^L \\ \widehat{\boldsymbol{x}}^R \end{pmatrix}$, $\boldsymbol{X}^E = \begin{pmatrix} \boldsymbol{X}^L \\ \boldsymbol{X}^R \end{pmatrix}$. As we can see, while solving the optimization equation, this combination is the same with concatenating left and right appearance vectors together, which unifies left and right appearance spaces into the same space, so that the relation of both eye images are implicitly modelled in the same locally linear combination. Thus, let "Concatenate" denote the proposed method.

We also study another two methods applying the basic mathematical operation to the left and right vectors. In equation 5, $\widehat{\boldsymbol{x}}^E = (\widehat{\boldsymbol{x}}^L - \widehat{\boldsymbol{x}}^R)$, $\boldsymbol{X}^E = (\boldsymbol{X}^L - \boldsymbol{X}^R)$,



(a) Head pose (b) Mouse click

Figure 2: Distribution of head pose and mouse click

which is denoted as "Minus". $\widehat{\boldsymbol{x}}^E = (\widehat{\boldsymbol{x}}^L + \widehat{\boldsymbol{x}}^R)$, $\boldsymbol{X}^E = (\boldsymbol{X}^L + \boldsymbol{X}^R)$, which is denoted as "Plus". The methods can be regarded as the dimensionality reduction of "Concatenate" where it reduces half reconstruction space.

4 Experiment

In order to evaluate the gaze estimation performance of our method with head movement, we conduct an experiment that all subjects worked in natural, and collected valid samples with natural head movement. In the experiment, our objectives are two fold: 1). decreasing the estimation error using the proposed method comparing with other methods. 2). studying the impact of the degree of head pose variation.

4.1 Data Collection

We developed the system on a desktop computer with a 21-inch LED-lit monitor attaching an off-the-shelf webcam (30fps), and the eye tracker (Tobii EyeX, 60Hz) was put on the bottom of the monitor. Five volunteers (2 male, 3 female) from the local university participated in the experiment. To guarantee natural head movement, the system was installed on their own computer in the office, so they worked on their daily work as usual with free movement of head and body, and they were also allowed to leave and come back as usual. The experiment of each participant covered a duration of two days.

The experiment procedure is as follows: the participants first carried out the calibration with the eye tracker. Later, the gaze points estimated by the eye tracker were taken as validation data. The mouse clicks were taken as the trigger event. Once the participant clicked the left mouse button, the system simultaneously recorded the frontal face image captured by the webcam, the current gaze point estimated by the eye tracker, and the mouse click position. If the distance of mouse click position and estimated gaze point is within 80 pixels (the eye tracker accuracy is around $2^\circ \sim 3^\circ$ with head movement), the pair of face image and the mouse click position was selected as a valid sample. In total, we collected one thousand valid samples for each participant.

To demonstrate the head variability of the participants, we used the method from [2] to estimate the head pose of the face images. Figure 2a shows the distribution of head pose that indicates natural head movement; figure 2b shows that mouse clicks distribute among the large area of the screen.

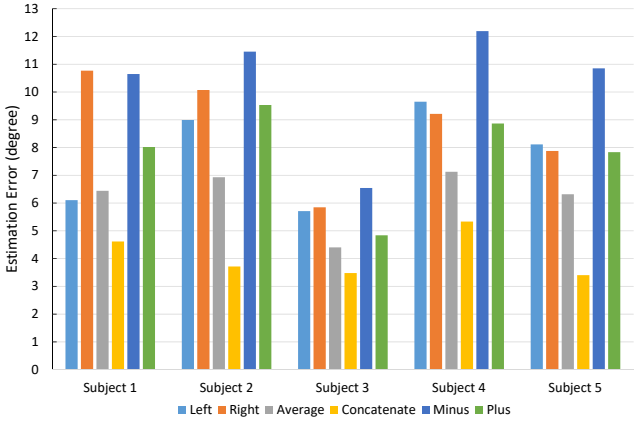


Figure 3: The evaluation results. "Left", "Right" indicates only left or right eye images is used; "Average" indicates using average estimated points of both left and right eye images; "Concatenate" indicates our proposed method; "Minus"/"Plus" indicates left eye-image vector minus or plus right eye-image vector.

Then we follow the procedure in [8] to extract eye appearance features ,where the eye-area images were cropped at the same fixed aspect ratio as shown in figure 4a. Finally the feature vector of the eye appearance is generated by raster scan of the eye intensity images (figure 1). The leave-one-out cross validation was employed, and the performance was evaluated by the mean estimated angular error as:

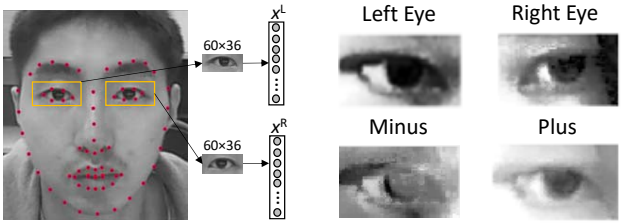
$$error = \frac{1}{n}\sum_{i=1}^{n} \arctan\left(\frac{\left\|\widehat{p}_i - p_i\right\|_2}{d}\right) \quad (6)$$

where $\|\widehat{p}_i - p_i\|$ is the Euclidean distance between estimated gaze position $\widehat{p}_i$ and actual gaze position p_i. d is the distance between participant's eyes and the screen. In the experiment, we take the average distance as 60cm).

4.2 Results

Figure 3 shows the evaluation results, where x-axis indicates different methods in each subject, and y-axis indicates the estimated angular error. There are some observations: 1). Although the performance of using the only left or right eye images has big gap on subject 1, it does not show significant difference among subjects ($p\text{-}value = 0.3981 > 0.05$), but averaging is always better than using the only one eye images. 2). for applying the mathematical operation on eye-image vector, the plus operation could slightly decrease the error, but the minus operation leads to larger error. 3). our proposed method always shows the lowest error, which outperforms averaging method by 12.97% to 45.28% ($p\text{-}value = 0.0082 < 0.05$).

These observations help us understand the intuition of the appearance-base method. "Average" can be regarded as a post-combination method, where it does not process the original image vectors, but deal with the estimation results of both eye images. The method is capable of decreasing the error by reducing the bias of the respective reconstruction spaces, but it is losing the local relation between them. Thus, the pre-combination method is restricting the local relation of



(a) Eye area cropping (b) Image effect

Figure 4: Eye area cropping and the image effect applying the minus/plus operation

both eye images by processing the original image vectors, i.e. "Minus" and "Plus", in which the two individual spaces are integrated to the same space. However, such mathematical operations could affect the pixel-level information of both eye images, which causes the different performance. In order to better understand the impact of the operators, we plot the resultant images in figure 4b (normalized to 0-255). As we can see, after doing the "minus" operation, the image becomes blurred and not clearly recognized. The "plus" operation enhances the same and reduces the difference, so that the pupil area which is related to gaze becomes more prominent. Our proposed method is not pre-combination (original vector processing) or post-combination (estimated gaze points averaging), but an embedding combination which is to learn the "hidden" relation between left and right eye images better matching the space mapping with head movement.

In order to further analyze the performance of different combinations, we investigated the impact of degree of head poses variation. we used a head tracker in the toolkit [2]. The tracker estimated the head pose of each face image, and output 6-dimensional head pose vector. The head-pose similarity is calculated using Euclidian distance of corresponding vectors, and the K-means algorithm is used to do the clustering task. In the experiment, we iteratively set K from 1 to 10. Given a new face image, the gaze point is estimated in the closest cluster. Figure 5 shows the evaluation results of subject 4. When K=1, the result is the same as figure 3. As K increases, i.e. controlling the head pose gradually, the error first drops dramatically, and then taper off. As we can see, the relative performance of these combinations does not change, but "Average" is approximating "Concatenate", which verified the above statement that with relatively stable appearance, the two manifolds constituted by left and right eye appearances would be homogeneous, so that the "Average" method is able to handle it. While head pose variation is large ($K < 3$), our proposed method is still better.

5 Conclusion

In this paper, we have presented a new reconstruction method that is able to robust to head movement in appearance-based gaze estimation. The proposed method is embedding into the reconstruction stage, aligning the left and right appearance spaces into the same local structure, which is able to maintain the relation of both eyes appearance under head movement. In addition, we conducted an experiment to investigate the performance of different combinations of eye images. The experiment results demonstrate the proposed method outperforms others significantly. The observations of the results also provide new sights of intuition of the appearance-based method, which could help to extract efficient eye appearance features even in deep learning.

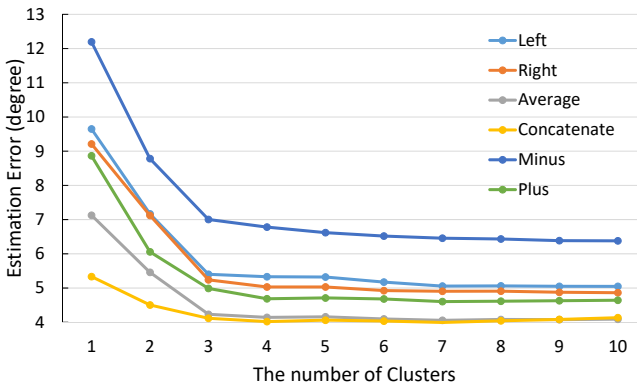


Figure 5: The degree of head pose variation

References

[1] Shumeet B. and Dean P. Non-intrusive gaze tracking using artificial neural networks. In *NIPS*, volume 6. 1994.

[2] T. Baltru, P. Robinson, and L.P. Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016.

[3] J.X. Chen and Q. Ji. Probabilistic gaze estimation without active personal calibration. In *CVPR*, 2011.

[4] Y.M. Cheung and Q.M. Peng. Eye gaze tracking with a web camera in a desktop environment. *THMS*, 45(4):419–430, 2015.

[5] D.W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *PAMI*, 32(3):478–500, 2010.

[6] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *CVPR*, 2016.

[7] Y. Liu, B.S. Lee, and M.J. McKeown. Robust eye-based dwell-free typing. *IJHCI*, 32(9):682–694, 2016.

[8] Y. Liu, B.S. Lee, A. Sluzek, D. Rajan, and M. Mckeown. Feasibility analysis of eye typing with a standard webcam. In *ECCV*, 2016.

[9] Y. Liu, C. Zhang, C. Lee, B.S. Lee, and A.Q. Chen. Gazetry: Swipe text typing using gaze. In *OzCHI*, 2015.

[10] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *ICCV*, 2011.

[11] K.H. Tan, D. J Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *WACV*, 2002.

[12] R. Valenti and T. Gevers. Accurate eye center location through invariant isocentric patterns. *PAMI*, 34(9):1785–1798, 2012.

[13] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the s^ 3gp. In *CVPR*, 2006.

[14] X.C. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015.

[15] Z.W. Zhu and Q. Ji. Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. *CVIU*, 98(1):124–154, 2005.