User-Independent Gaze Estimation by Exploiting Similarity Measures in the Eye Pair Appearance Eigenspace

Nanxiang Li and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory
University of Texas at Dallas
800 W Campbell Rd, Richardson, TX 75080
nxl056000@utdallas.edu, busso@utdallas.edu

ABSTRACT

The design of gaze-based computer interfaces has been an active research area for over 40 years. One challenge of using gaze detectors is the repetitive calibration process required to adjust the parameters of the systems, and the constrained conditions imposed on the user for robust gaze estimation. We envision user-independent gaze detectors that do not require calibration, or any cooperation from the user. Toward this goal, we investigate an appearance-based approach, where we estimate the eigenspace for the gaze using *principal component analysis* (PCA). The projections are used as features of regression models that estimate the screen's coordinates. As expected, the performance of the approach decreases when the models are trained without data from the target user (i.e., user-independent condition). This study proposes an appealing training approach to bridge the gap in performance between user-dependent and user-independent conditions. Using the projections onto the eigenspace, the scheme identifies samples in training set that are similar to the testing images. We build the sample covariance matrix and the regression models only with these samples. We consider either similar frames or data from subjects with similar eye appearance. The promising results suggest that the proposed training approach is a feasible and convenient scheme for gaze-based multimodal interfaces.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: Miscellaneous

Keywords

User-independent gaze estimation; similar subject measurement; eigenspace analysis; computer user interface

1. INTRODUCTION

Gaze indicates visual awareness, thus, it is important for understanding human intention and attention. The study of gaze has drawn attentions from many research fields including *human-computer interaction* (HCI), market analysis, social behavior, and driver distraction [5, 8, 10]. For HCI, using gaze as an input modality has unique advantages over other modalities: for the users, gaze is a natural and fast way to indicate their interest and intentions; for the computers, being aware of the gaze allows more effective interventions. However, two limitations of employing gaze as input are the repetitive and tedious calibration process, and the constrained conditions imposed on the user to achieve accurate estimation. User and setting specific parameters are usually calibrated by asking the users to look at several reference points [3]. Although the number of reference points can be reduced by employing more light sources or cameras [12, 6], the process still need to be conducted whenever the setting changes. Using head mount devices addresses some of the problems, but it is an intrusive alternative that requires initial calibration. We envision user-independent gaze detectors that do no require calibration or any cooperation from the user. Appearance-based models are appealing approaches for this purpose.

Many studies proposed appearance-based models for gaze estimation. However, most of them focus on user-dependent conditions that require calibration. Baluja and Pomerleau [1] use a neural network to estimate gaze position using the eye image. Tan et al. [11] 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. Shiele and Waibel [9] study 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 [7] used morphable models to estimate user-independent gaze, where the main challenge is the initial match between the model parameters and the image.

We recently proposed an appearance-based gaze estimation approach that consists in building an eigenspace for patches displaying both eyes [4]. The projections onto the eigenspace are used to train a regression model to predict the target screen coordinates. We implemented the approach operating under user-dependent (i.e., training and testing on the same subject) and user-independent (i.e., training and testing on different subjects) conditions. The results reveal a drop in accuracy for the user-independent condition. Building upon this approach, this study proposes a training scheme to reduce the gap in performance between user-dependent and user-independent gaze estimation.

ICMI'14, November 12–16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2663250.

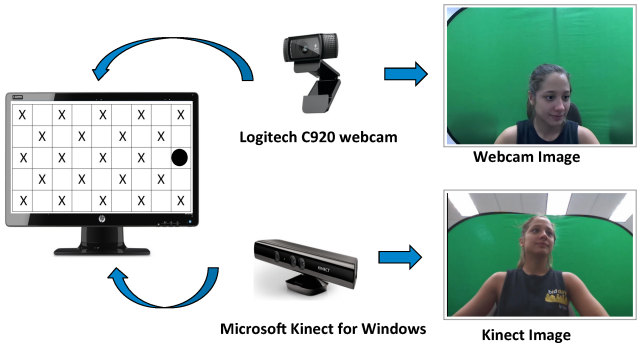


Figure 1: The data collection includes a 22-inch HP monitor, a Logitech C920 webcam and a Microsoft Kinect for Windows. A green screen is placed behind the subject to provide uniform background.

One factor affecting the performance of the appearance-based gaze approach in user-independent mode is that the eigenspace created from the sample covariance matrix does not offer a good representation for the target user. Starting from the premise that *not all data is good data*, we propose to identify a reduced set of frames from the training database that have similar appearance to the samples from the target user. We build the eigenspace and the regression models using this reduced set. This approach offers many advantages (1) it uses data from subjects in the training set with similar eye appearance improving the orthogonal basis to represent the target user; (2) it uses samples where the user is looking to similar screen locations; (3) it uses samples where the user's head position is similar to the target frame. We evaluate the performance of the proposed approach under different conditions (with/without head movement, various user-computer distance). The proposed training scheme improves the performance of the user-independent system, showing that it is a feasible and convenient approach to facilitate gaze-based multimodal interfaces.

2. MSP-GAZE DATABASE

The MSP-GAZE corpus was collected to study appearance-based gaze estimation methods [4]. It considers various factors that affect gaze estimation in HCI: individual appearance differences, head movement, inter-session variability, and distance between user and screen. The data was collected in a laboratory environment using a standard 22-inch HP monitor, a commercial webcam (Logitech C920) placed on top of the monitor and a Microsoft Kinect sensor for Windows placed below the monitor (Fig. 1).

The database includes data from 46 subjects whose average age was 22.7 (min 19, max 35). The corpus is balanced in terms of gender. It also includes a diverse ethnic representation of the students from University of Texas at Dallas (Caucasian, Asian, Indian, and Hispanic populations). We recorded each subject in two different days separated with an average interval of five days. During each of these two sessions, we collected 12 training recordings and 2 testing recordings. The different recordings include the following configuration: with and without head movement; and four user-screen distances (user defined, near 0.4m, medium 0.5m and far 0.6m). The user is asked to click on a randomly generated point displayed on predefined regions. During the training recordings, one of the four predefined user-screen distances is used and the random points are projected in selected regions (See Fig. 1). During the testing recordings, the users selected the distance from the screen and the points are randomly placed on the monitor without any constraint. For each subject, we have about 90 minutes of data over the two sessions. We simultaneously recorded the videos, the location of the target points, the position of the mouse cursor, and the mouse click actions. The details of the corpus are given in Li and Busso [4].

Table 1: Average correlation using appearance-based gaze estimation under user-dependent and user-independent conditions (W: with head motion; W/O: without head motion).

User-Screen Distance	User-Dependent		User-Independent	
	W/O	W	W/O	W
Far	0.89	0.91	0.84	0.79
Medium	0.90	0.92	0.88	0.82
Near	0.89	0.83	0.89	0.84
User-Defined	0.93	0.92	0.85	0.79
Average	0.91	0.90	0.86	0.80

This study relies on data from 24 subjects that has been preprocessed (data cleaning and eye pair extraction). They comprise a gender and ethnic balanced dataset (3 $\times$ 2 gender $\times$ 4 ethnic background). We extracted patches with both eyes, following the approach proposed in our previous work [4] (see Fig. 2). Considering both eyes reduces eye detection errors, and improves robustness against head motion. The eye pair image extraction is done automatically 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. [2]. For each point displayed on the screen, we extract the eye pair images from three frames (< 0.14s) 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).

3. APPEARANCE-BASED GAZE MODEL

In our previous study, we focused on the webcam videos and investigated an appearance-based approach that builds an orthonormal basis of the extracted eye pair images [4]. By resizing all extracted eye pair images to 100 x 25 pixels, a covariance matrix is built and the PCA is implemented. The projections onto the PCA basis are used to train a linear regression model which predicts the gaze location (screen's coordinates). Two linear regression models are built separately to estimate the gaze position in the horizontal (x) and vertical (y) coordinates. This approach was evaluated for both user-dependent (training and testing on the same subject) and user-independent (training and testing on different subjects) conditions. We re-implement the evaluation for the 24 subjects considered in this study. The evaluation considers the top 100 principal components, which capture over 90% of the variance [4]. For the user-dependent condition, we train and test the model with recordings from the same subjects. For the user-independent condition, we use a *leave-one-subject-out* (LOSO) cross-validation approach, where in each fold we train with data from 23 subjects and and test with data from the remaining subject. The evaluation considers matched training/testing settings in terms of user-screen distances and head movement conditions. We

calculate the correlations between the estimated and the actual gaze positions in the horizontal (ρ_x) and vertical (ρ_y) directions, and report the average correlation ρ between the two:

$$\rho = \frac{\rho_x + \rho_y}{2} \quad (1)$$

Table 1 lists the results which show that the average correlation ρ decreases approximately 0.1 from the user-dependent to the user-independent conditions when head movement is allowed. While these results are competitive given that the approach does not require any calibration, this gap can be reduced by implementing better training schemes.

3.1 Proposed Approach: Finding Similar Data

Appearance-based models are sensitive to variation in the patch images. Therefore, the differences between subjects in the training set affect the orthogonal basis to represent the target subject. Following this direction, we propose to build the eigenspace and the regression models using only patches that are similar to the eye appearance from the target subject. The approach offers various advantages. It improves the orthogonal basis to appropriately represent the eigenspace of the target subject. Since similar images most likely share the same gaze direction, it considers training images spanning the target gaze directions. It also increases the robustness against head motion, since the selected frames will tend to have similar head poses.

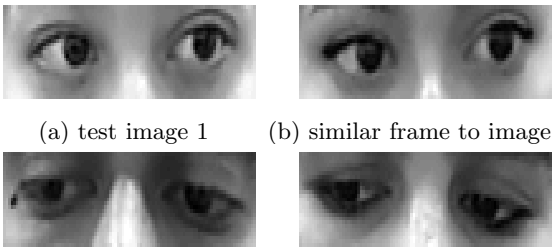
The orthogonal basis corresponds to the eigenvectors belonging to the larger eigenvalues of the covariance matrix Σ. If all N training samples are used, Σ_{All} is calculated with Equation 2, where Φ_i is the mean removed images concatenated as a vectors. If we reduced the training images to a set $\mathcal{S}$ including only similar images, we can build a target user-specific covariance matrix $\Sigma_{\mathcal{S}}$ with Equation 3. The key problem consists in selecting the set $\mathcal{S}$ (i.e., finding similar images to the target user).

$$\Sigma_{All} = \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T \quad (2)$$

$$\Sigma_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in \mathrm{S}}\Phi_i\Phi_i^T \quad (3)$$

We propose the following steps to achieve this goal. First, we estimate Σ_{All} following Equation 2. Then, we project each training image into the first 100 principal components. The testing images are also projected into this 100D space. Finally, we estimate the Euclidean distance between the projections of the training and testing images. Figures 2(a) and 2(c) show two images in the testing set. Figures 2(b) and 2(d) gives the closest images in the training set, respectively. The figures show that the testing and training images have similar head pose and gaze direction. The ground truth gaze positions associated with Figures 2(a) and 2(b) are the screen's coordinates (68, 160), and (75, 99), respectively. For Figures 2(c) and 2(d), the screen's coordinates are (1600, 967) and (1526, 931).

This study considers two criteria to select $\mathcal{S}$: similar subjects and similar frames. For similar subjects, the goal is to find subjects that have similar appearance to the target user. After projecting all the testing images into the eigenspace of Σ_{All}, we select the most similar frame in the training set for each of the target images. From this selected training frames, we can associate the identity of the selected training subjects. By ranking in descending order the training subjects according to the number of selected images belonging to them, we define $\mathcal{S}$ as the set with the data from the top n subjects in the list. This approach guarantees a balanced distribution of the training samples over the screen's coordinates. For similar frames, we search for the top n images in the training set that are similar to each of the testing images. We define $\mathcal{S}$ as the set including these selected images.



(a) test image 1 (b) similar frame to image 1

(c) test image 2 (d) similar frame to image 2

Figure 2: Two testing images and their closer patches in the training set, as measured by the Euclidean distance in the projected eigenspace.

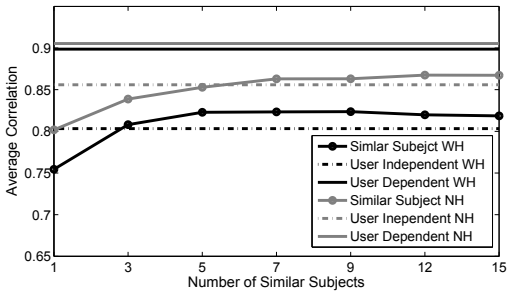
3.2 Experimental Evaluation

By using the reduced set $\mathcal{S}$, we estimate a new eigenspace from $\Sigma_{\mathcal{S}}$ using Equation 3. We train separate linear regression models for the horizontal (x) and vertical (y) coordinates, where the projections onto this new basis are the independent variables, and the corresponding ground truth gaze positions are the dependent variables. The output of these regression models are limited to be within the screen size. Following the approach presented in our previous work, we report the average correlation ρ defined in Equation 1 to evaluate the proposed solution. The evaluation also uses LOSO cross-validation scheme with matched training/testing conditions (head motion and user-screen distance).

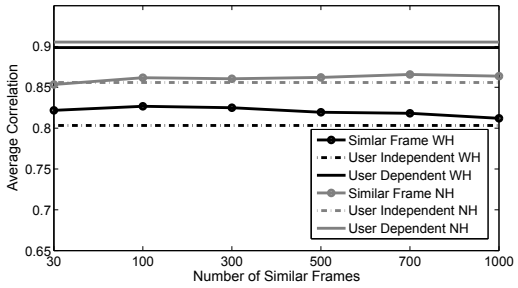
Figure 3(a) shows the average correlation ρ across user-screen distances when $\mathcal{S}$ includes images from the top n similar subjects. We evaluate n =1, 2, 3, 5, 7, and 9. The straight lines in the figures correspond to the performance for user-dependent (solid line), and user-independent (dash line) conditions using Σ_{All} (Table 1). When we consider more than 5 subjects, the performance increases compared with the results in user-independent conditions considering all the training subjects. The best ρ is achieved when n = 12 (ρ=0.87) without head movement, and n = 5 (ρ=0.82) with head movement. Although these results are still worse than the ones achieved under user-dependent conditions, we expect improvements when the pool of subjects in the training set increases (i.e., finding better matches).

Figure 3(b) shows the result when $\mathcal{S}$ includes the top n images that are closer to each testing set. We evaluate n = 30, 100, 300, 500 and 1000. Even with n = 30, we observe better performance than the results for user-independent condition, which demonstrate the potential of this training approach.

Table 2 reports the experimental results for 7-similar subjects and 100-similar frame across different conditions. Both approaches have similar performance across conditions. For the user-defined distance with head movement recordings, using similar frames is slight better than using similar subjects. Both cases are better than the ones for the user-independent condition using all training data (Table 1).



(a) Similar subject approach



(b) similar frame approach

Figure 3: Evaluation of the training scheme implemented with similar *subjects* and similar *frames*. (WH: with head motion; NH: without head motion)

4. CONCLUSIONS AND DISCUSSION

This paper proposed a training scheme to improve an appearance-based, user-independent gaze estimator. The proposed approach identifies similar data samples in the training set and uses this reduced set to build the eye appearance eigenspace, and the regression models to estimate the screen's coordinates. The proposed approach eliminates the requirement for tedious calibration, which is required for most gaze estimation systems. The results demonstrate the robustness against the user's head motion. The promising experimental results show the effectiveness of the proposed approach, reducing the gap between user-dependent, and user-independent gaze estimation accuracy.

One limitation of the approach is that it requires testing samples to define the set $\mathcal{S}$. One possible solution is to start the gaze estimation with user-independent models. As more test samples become available, we can create/update the set $\mathcal{S}$ and estimate $\Sigma_{\mathcal{S}}$. This approach can be implemented for real-time applications. The results presented in this study focus on the case when test samples are available, toward understanding the benefits of the proposed approach compared to the speaker independent mode.

Our future work considers extensions of the proposed training approach to improve the performance of user-independent gaze estimation. This study only considered 24 subjects from the corpus to have a balanced dataset in terms of gender and ethnic background. We will reestimate the system by including all the subjects. By increasing the training set, we expect that the selected samples will be more similar to the testing images. Another direction for user-independent gaze estimation is to take advantage of the test images and perform on-line calibration. We can incorporate this idea by adapting the covariance matrix to the target subject.

Table 2: User-independent gaze estimation results. The reduced set $\mathcal{S}$ includes images from either the top 7 similar *subjects*, or the top 100 similar *frames*. (W: with head motion; W/O: without head motion).

		W/O		W	
Distance		ρ_x	ρ_y	ρ_x	ρ_y
Subjects	Far	0.90	0.79	0.88	0 .72
	Medium	0.93	0.82	0.90	0.74
	Near	0.93	0.87	0.93	0.78
	User-Defined	0.91	0.78	0.88	0.74
Frames	Far	0.90	0.81	0.88	0.72
	Medium	0.92	0.80	0.90	0.74
	Near	0.93	0.82	0.93	0.77
	User-Defined	0.92	0.81	0.90	0.76

5. ACKNOWLEDGMENTS

This work was funded by NSF (IIS 1217104) and Samsung.

6. REFERENCES

[1] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical Report CMU-CS-94-102, Carnegie Mellon University, Pittsburgh, PA, USA, January 1994.

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

[3] D. Hansen and 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, March 478-500.

[4] N. Li and C. Busso. Evaluating the robustness of an appearance-based gaze estimation method for multimodal interfaces. In *International conference on multimodal interaction (ICMI 2013)*, pages 91–98, Sydney, Australia, December 2013.

[5] Y. Matsumotot, T. Ino, and T. Ogsawara. Development of intelligent wheelchair system with face and gaze based interface. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, pages 262–267. IEEE, 2001.

[6] T. Nagamatsu, J. Kamahara, T. Iko, and N. Tanaka. One-point calibration gaze tracking based on eyeball kinematics using stereo cameras. In *Symposium on Eye tracking research & applications*, pages 95–98, Santa Barbara, CA, USA, March 2008.

[7] T. Rikert and M. J. Jones. Gaze estimation using morphable models. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 1998)*, pages 436–441, Nara, Japan, April 1998.

[8] D. Salvucci and J. Anderson. Intelligent gaze-added interfaces. In *SIGCHI conference on Human Factors in Computing Systems (CHI 2000)*, pages 273–280, The Hague, The Netherlands, April 2000.

[9] B. Schiele and A. Waibel. Gaze tracking based on face-color. In *International Workshop on Automatic Face-and Gesture-Recognition*, pages 344–349, Zurich, Switzerland, June 1995.

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

[11] K. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimatio. In *IEEE Workshop on Applications of Computer Vision (WACV 2002)*, pages 191–195, Orlando, FL, USA, December 2002.

[12] A. Villanueva and R. Cabeza. A novel gaze estimation system with one calibration point. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(4):1123–1138, August 2008.