Appearance Based User-Independent Gaze Estimation

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

ABSTRACT

An ideal gaze user interface should be able to accurately estimates the user's gaze direction in a non-intrusive setting. Most studies on gaze estimation focus on the accuracy of the estimation results, imposing important constraints on the user such as no head movement, intrusive head mount setting and repetitive calibration process. Due to these limitations, most graphic user interfaces (GUIs) are reluctant to include gaze as an input modality. We envision user-independent gaze detectors for user computer interaction that do not impose any constraints on the users. We believe the appearance of the eye pairs, which implicitly reveals head pose, provides conclusive information on the gaze direction. More importantly, the relative appearance changes in the eye pairs due to the different gaze direction should be consistent among different human subjects. We collected a multimodal corpus (MSP-GAZE) to study and evaluate user independent, appearance based gaze estimation approaches. This corpus considers important factors that affect the appearance based gaze estimation: the individual difference, the head movement, and the distance between the user and the interface's screen. Using this database, our initial study focused on the eye pair appearance eigenspace approach, where the projections into the eye appearance eigenspace basis are used to build regression models to estimate the gaze position. We compare the results between user dependent (training and testing on the same subject) and user independent (testing subject is not included in the training data) models. As expected, due to the individual differences between subjects, the performance decreases when the models are trained without data from the target user. The study aims to reduce the gap between user dependent and user independent conditions.

Categories and Subject Descriptors

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

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

http://dx.doi.org/10.1145/2663204.2666288.

Keywords

User-independent gaze estimation; eigenspace analysis; tensor analysis; domain adaptation

1. INTRODUCTION

Gaze is a fast and natural way to identify the vision attention of the user. Accurate gaze estimation enable human computer interfaces to have targeted and effective interactions[1, 4]. Gaze estimation has drawn attention from many different research communities [7, 11, 13]. An ideal gaze tracking system should be accurate and easy to use. Although accurate wearable eye tracking devices with flexible setting do exist, they come with high cost. Common gaze tracking system requires tedious calibration process or imposes constraints on the users. These constraints are required to estimate parameters related to the subject, data acquisition equipments and system settings. The calibration process usually involves letting the users to look at several reference points while calculating the relative eye movement [5]. Additional constraints, such as limited or no head movement, or wearing a head mount device, are added to ensure the reliability of the calibrated parameters. Without these constraints, calibration process has to be re-estimated whenever the parameter changes (head pose, user position).

Most current gaze estimation approaches focus on feature based methods, where the goal is to extract eye pupil and glint (bright reflection of the active light from the eye). Although good results have been achieved in this direction, this method involves the calibration process and requires both camera and active light source. An alternative direction is the appearance based method where the explicit pupil and iris features estimation is not required. This approach is ideal for the gaze estimation system that we envision, since it can be built with normal webcam. Many appearance-based gaze estimation studies focus on user-dependent conditions that require calibration. Baluja and Pomerleau [2] trained a neural network with the eye images and achieved 1.5° angular error gaze estimation. Tan et al. [14] proposed a local linear interpolation approach in a sparse eye image sample space, which had less than 0.5° angular error in their dataset.

In this proposal, we aim to develop an appearance based user independent gaze approach which eliminates the calibration process. Few studies have addressed user independent gaze estimation. Shiele and Waibel [12] relied on head pose to estimate a coarse gaze direction using neural network. Their result shows 12° angular error on average for the test data. Rikert and Jones [10] considered morphable

models to estimate user-independent gaze, and achieved better results. Overall, due to the individual difference, a gap between the user dependent and user independent results is observed. The proposed study aims to reduce this gap. A multimodal corpus (MSP-GAZE) is collected to understand various factors that affect the performance of appearance-based gaze estimation approaches, and to develop robust user independent gaze estimation system.

The primary contributions of this proposal include:

- We collect a novel multimodal corpus (MSP-GAZE) to study and evaluate appearance based gaze estimation methods.
- We study the effect of different factors (individual differences, head movement, distance between the user and the interface's screen) on the appearance based gaze estimation approaches.
- We evaluate different model adaptation strategy to improve the performance of user independent gaze estimation, where no calibration is required. The result of this study can lead to a convenient and practical gaze estimation system for human computer interaction.

2. PRELIMINARY RESULTS

This section summarizes the preliminary results including the description of the MSP-GAZE database, the difference between user dependent and independent gaze estimation results, and our initial attempt to improve the performance of the user independent model by using similarity measures in the eye pair appearance eigenspace.

2.1 MSP-GAZE Database

The MSP-GAZE database is collected in a laboratory setting where sufficient and steady illumination condition is guaranteed. The data collection process involves letting the participants look at and click at randomly projected points displayed on a monitor, and having their glancing behavior and mouse movement recorded. The system includes a regular 22-inch HP monitor, a commercial webcam (Logitich C920) and a Microsoft Kinect sensor. Fig. 1 illustrates the data collection setting and samples of the collected video frames. As shown in the figure, the webcam is placed on top of the monitor and the Kinect sensor is placed below the monitor. Both devices are aligned with the center of the monitor. The webcam and Kinect sensor capture the subject glance behavior from different perspective relative to the field of interest. By comparing the captured data, we can 1) understand and evaluate the effect of camera position; 2) build a robust gaze estimation model using both devices since both the webcam and the Kinect sensor have their own limited view, especially for the vertical gaze position prediction. In addition, the Kinect sensor provides depth estimation which can be used to estimated the user head pose and the distance between the user and the monitor.

A total of 46 students from different disciplines at the University of Texas at Dallas participated in the data collection. The average age of the participants is 22.7 (min 19, max 35). They are balanced between genders. Moreover, they are balanced between a diverse ethnic group including Caucasian, Asian, Indian, and Hispanic. The diverse ethnic

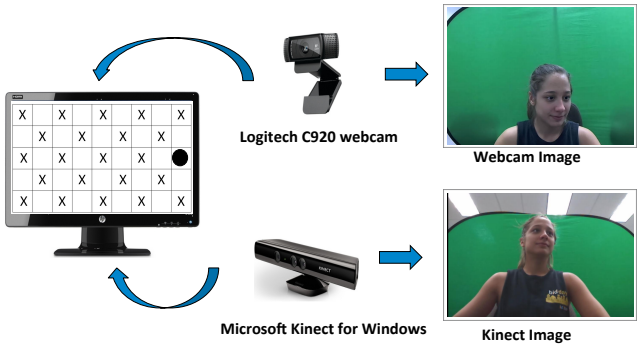


Figure 1: The data collection includes a 22-inch HP monitor, a Logitech C920 webcam and a Microsoft Kinect for Windows. A green screen is placed behind the subject to provide uniform background.

Table 1: Recordings conditions for each session.

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

representation introduces a variety of facial appearances, allowing us to evaluate the effect of individual differences on the appearance model. To simplify the eye detection process and focus on the gaze estimation modeling, we did not consider the cases where the participant wears glasses.

Each subject participated in 2 data collection sessions, each at different days following the same process. The purpose of collecting data from the same subject on different days is to evaluate the consistency of the appearance model across time. The average interval between the sessions is five days.

Each data collection session consists of 14 recordings with different conditions list in Table 1. The first two recordings are collected under no constrains where the subjects decide the user-monitor distance ("User-defined"). The goal is to mimic the normal user computer interaction for each individual. Recordings 3 to 7 uses predefined user-monitor distance where "Near" is 0.4 meter, "Medium" is 0.5 meters and "Far" is 0.6 meters. Recordings 8 to 14 further add the no head movement constraint, where the subjects are required to fix their head pose. Since head movement and user monitor distance affect the appearance model, the constrained data recordings allow us to evaluate the effect of these factors.

The data collection considered two random point projection schemes, which results in "Training" (recordings 3-7 and 9-14) and "Testing" (recordings 1 and 8) recordings. The "Training" projection pattern emphasizes the coverage of the monitor screen (as mush as possible) such that a full mapping between the eye appearance and different ground truth gaze position can be established. We achieve this by dividing the screen into a 5 by 9 grid and projecting points

randomly inside the 23 highlighted grids (marked with 'X' in Fig. 1). As shown in the figure, the 23 grid span the whole screen area vertically and horizontally. In addition, we design the projection process in a way that each of the 23 grids is visited 4 times, generating 92 total random points in each "Training" recording. For the "Testing" projection pattern, the purpose is to evaluate the gaze estimation performance thus does not consider any constraints. Points are randomly projected over the whole screen. A detailed description can be found in Li and Busso [6]. Notice that similar point projection patterns have been used in previous studies [8, 9, 14].

2.2 Appearance-Based PCA Approach

We considered the eye pair image to build the appearance model since it implicitly models the head pose. Also, detecting eye pair is more robust than single eye detection. We use the implementation of Viola-Jones objection detection in the *open computer vision library* (OpenCV) to automatically extract the eye pairs from the collected video frames [3]. We only consider the video frames when the user clicks on the projected point (3 frames before and after the mouse click) during which the mapping between the eye appearance and the projected point is accurate.

Using these extracted eye pair images, we build an orthonormal basis of the eye pair appearance. The orthogonal basis corresponds to the eigenvectors of the covariance matrix Σ estimated from the training images. The projections onto the PCA basis are regarded as a compact representation of the eye appearance, thus can be used to train a regression model to predict the gaze position (screen's coordinates). Two linear regression models are separately built to estimate the gaze position in the horizontal (x) and vertical (y) coordinates.

Under matched training/testing settings (user-screen distances and head movement), we evaluate the performance using the correlation between the ground truth and the prediction (ρ_x, ρ_y), as well as the average between the two (ρ).

This evaluation was done for both user-dependent (training and testing on the same subject) and user independent (training and testing on different subjects) conditions. For the user independent result, we use a *leave-one-subject-out* (LOSO) cross-validation approach, where in each fold we train with data from all subjects , with the exception of the test subject. Table 2 lists the results. It shows that the proposed approach is not affected much by the subject's head motion. The evaluations with and without head movements provide similar gaze estimation performance. We also notice that the distance between the user and the computer monitor has little affect on the accuracy of the system. In addition, we evaluate the consistency of the subject dependent model across time, where the data in one session is trained and tested on another session. We observed similar performance between the two. Overall, the consistent results under different conditions show promising potential for human computer interaction applications.

A better measurement of the gaze estimation performance is the angular error θ_{error}. Accurate estimation of θ_{error} requires accurate measurement of the distance between the eyes and monitor. As the future work, we will use the Kinect depth image to estimate the user to the monitor distance d_{u-mc}. Using this information, along with the fact that the users' eye center is aligned with the monitor center, we can estimate the angular error θ_{error} by the following equation:

$$\theta_{error} = \cos^{-1}\left(\frac{d_{u-tp}^2 + d_{u-pp}^2 - d_{error}^2}{2d_{u-tp}d_{u-pp}}\right) \quad (1)$$

where d_{u-tp} and d_{u-pp} stand for the distance between the user eye center to the true gaze point and the predicted gaze point, respectively. d_{error} stands for the distance between the true gaze point and the predicted gaze point. This calculation is based on the *Law of Cosine*, where the triangle is formed by the user eye center, the true gaze position and the predicted gaze position. Notice that d_{u-tp} and d_{u-pp} can be estimated by the *Pythagorean theorem*, as shown in the following equations,

$$d_{u-tp} = \sqrt{d_{u-mc}^2 + d_{mc-tp}^2} \quad (2)$$

$$d_{u-pp} = \sqrt{d_{u-mc}^2 + d_{mc-pp}^2} \quad (3)$$

where d_{mc-tp} and d_{mc-pp} stand for the distance between the monitor center to the true gaze point and the predicted gaze point.

2.3 Similarity Measures in the Eye Pair Appearance Eigenspace

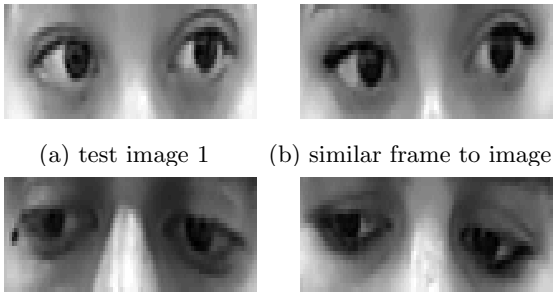
Due to individual difference between the users, usually the appearance based gaze estimation approach does not consider user independent case. This effect is clearly shown in Table 2 where the results of user independent models is worse than the user dependent models. In our case, the differences between subjects affects the eye pair image covariance matrix Σ, from which we derive the orthogonal basis representation of the test subject eye appearance. One way to reduce this difference is to predict the test subject gaze using a subset of training samples $\mathcal{S}$ that have similar eye appearance as the test subject. Following this direction, we propose to find similar training data to the test subject eye appearance, and build the target subject image covariance matrix $\Sigma_{\mathcal{S}}$ using the selected data samples $\mathcal{S}$.

We first estimate the covariance matrix using all the training data, which we referred to as Σ_{All}. For any test subject eye appearance, we search for similar training eye appearance images in the eigenspace defined by Σ_{All}. Specifically, we use the projection of the images to the top 30 eigenvectors of Σ_{All} as a compact representation in a 30 dimension space, and calculate the Euclidean distance between the projected 30D coordinates to measure the similarity between images. Fig. 2 shows the nearest neighbor training sample of two given test images. Although the difference between the selected training sample and test subject exists, they both show similar gaze directions. In fact, the ground truth gaze positions associated with the selected training samples and the test image are very close, with (68, 160) for Fig. 2(a) and (75, 99) for Fig. 2(b), (1600, 967) for Fig. 2(c) and (1526, 931) for Fig. 2(d).

Based on the similar training sample search scheme, we consider two ways to define $\mathcal{S}$: similar n frames and similar n subjects. The similar n frames method considers n nearest neighbors of the given test image in the eigenspace of Σ_{All}, while the similar n subjects method considers all training samples from the n subjects that have the closest average distance to the test image in the eigenspace of Σ_{All}.

Table 2: Appearance based PCA gaze estimation results.

	User Dependent						User Independent					
	Without head motion			With head motion			Without head motion			With head motion		
Distance	ρ_x	ρ_y	ρ	ρ_x	ρ_y	ρ	ρ_x	ρ_y	ρ	ρ_x	ρ_y	ρ
Near	0.90	0.85	0.87	0.91	0.84	0.87	0.85	0.76	0.81	0.87	0.75	0.81
Medium	0.89	0.84	0.87	0.91	0.83	0.87	0.86	0.75	0.81	0.85	0.74	0.79
Far	0.88	0.83	0.86	0.90	0.83	0.87	0.85	0.68	0.77	0.85	0.73	0.79
User-Defined	0.89	0.82	0.86	0.88	0.82	0.85	0.85	0.78	0.82	0.86	0.70	0.78



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

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

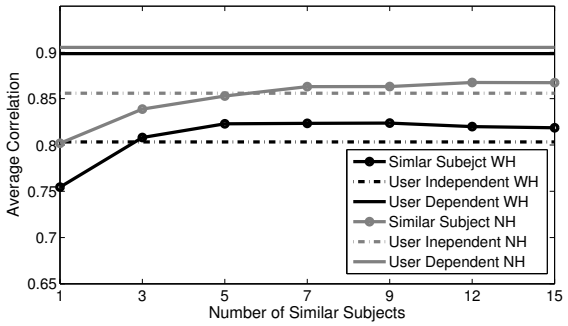
Figure 2: Two testing images and their closer patches in the training set, as measured by the Euclidean distance in the projected eigenspace.

Table 3: 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).

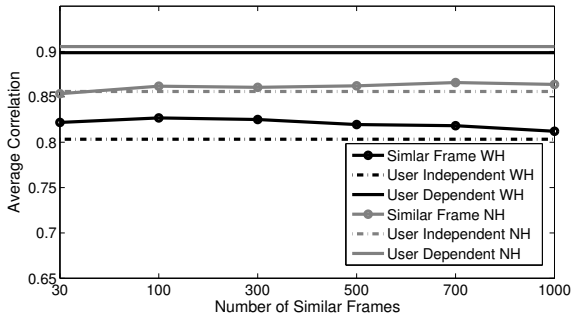
		W/O		W	
	Distance	ρ_x	ρ_y	ρ_x	ρ_y
Subjects	Far	0.90	0.79	0.88	0 .72
	Medium	0.93	0.82	0.90	0.74
	Near	0.93	0.87	0.93	0.78
	User-Defined	0.91	0.78	0.88	0.74
Frames	Far	0.90	0.81	0.88	0.72
	Medium	0.92	0.80	0.90	0.74
	Near	0.93	0.82	0.93	0.77
	User-Defined	0.92	0.81	0.90	0.76

We evaluate the proposed approach with different n values for both the similar n frames (n =30, 100, 300, 500, 700, 1000) and similar n subjects (n =1, 3, 5, 7, 9, 12, 15). The average correlation $\rho = \frac{\rho_x+\rho_y}{2}$ across different settings is shown in Figure 3. In the figure, we also highlight the performance of user dependent and independent models with straight lines. The goal of the proposed study is to achieve performance close to the dotted straight line (user dependent model). We observe that both similar search models can achieve better result than the user independent model. The best performance is achieved when n = 100 for similar n frames and n =7 for the similar n subjects, respectively.

Between the similar n frames and similar n subjects methods, it appears that the n similar frames scheme has better performance. One interesting result about the n similar frames approach is that it achieve better performance in terms of angular error when the subjects were free to move their head. This finding suggests that the selected training images have similar head pose as the test subject, improving the accuracy of the gaze detection system. (see Fig. 2).



(a) Similar subject approach



(b) similar frame approach

Figure 3: Evaluation of the training scheme implemented with similar *subjects* and similar *frames*.

Table 3 shows the correlation (ρ_x, ρ_y) results where the best performance is achieved with 100 similar frames and with 7 similar subjects. As observed before, similar frame approach performs slightly better than the similar subject approach, especially when the setting is with head movement. Although the results are still lower than the ones observed for user-dependent case, we observe an improved performance using the proposed method, compared with the results for use-independent case. We expect further improvement when more subjects are involved in the training set (i.e., finding better matches leading to better gaze estimation).

3. PROPOSED RESEARCH PLAN

We consider the following three directions to reduce the gap between the user dependent and independent models:

- **Incoorperate multimodal dataset**

  The MSP-GAZE corpus contains data collected from both a webcam and Kinect sensor. Although the study

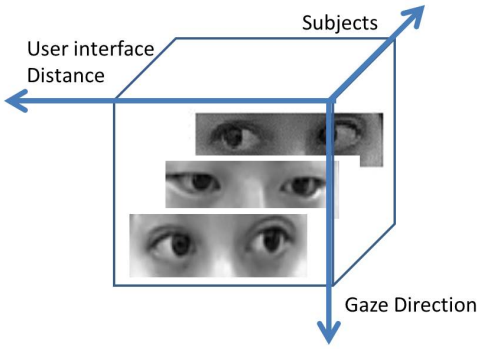


Figure 4: Tensor decomposition of the gaze data.

so far 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. We believe it will be helpful for the vertical gaze estimation since the Kinect sensor is placed below the screen, see Figure 1. The depth information can be used to estimate the head pose, and distance between the monitor and the participant.

In addition, We can also include the mouse click and mouse movement information to refine and update the gaze prediction result. For the human computer interaction, mouse click provide useful information about the visual interest of the user.

- **Tensor analysis**

  The MSP-GAZE corpus include various factors: gaze direction, individual difference, head movement, and distance between the user and the interface's screen. Since most of these factors are labeled (the last two parameters can be estimated from the kinect depth data), we can apply tensor analysis to decompose the database into the multiple subspaces, each corresponding to a particular factor. Figure 4 illustrate the idea. The subspace learnt by the training data have better representation than the PCA approach, thus it can improve the user independent gaze estimation.

- **Domain adaptation**

  General machine learning approaches assume same underlying distribution between the training and testing data. In our case, it may not be true due to individual difference. To be specific, the PCA of Σ_{All} and Σ_{Target} may not represent similar variation factors in the dataset. Domain adaptation approaches consider both shared or no shared support between source and target data, and aims to use the unlabeled data from the target domain to help learning. The idea of user-independent gaze estimation is very much in line with this concept. In particular, it belongs to the learning under covariance shift problems, where the goal is to adapt the covariance Σ_{All} to be as close as possible to Σ_{Target}. Our initial approach tries to estimate Σ_{Target} with Σ_S where training data are selected based on the similarity to the target data. We believe different domain adaptation approaches can be considered for this problem. Since it requires unlabeled test data, this approach is suitable for online user-independent gaze estimation.

4. REFERENCES

[1] G. Anders. Pilot's attention allocation during approach and landing- eye-and head-tracking research in an a 330 full flight simulator. *Focusing Attention on Aviation Safety*, 2001.

[2] 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.

[3] 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.

[4] C. Ghaoui. *Encyclopaedia of Human Computer Interaction*. ITPro collection. Idea Group Reference, 2006.

[5] 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.

[6] 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.

[7] Y. Matsumotot, T. Ino, and T. Ogsawara. Development of intelligent wheelchair system with face and gaze based interface. In *IEEE International Workshop on Robot and Human Interactive Communication*, pages 262–267, Bordeaux and Paris, France, September 2001.

[8] Y. Ono, T. Okabe, and Y. Sato. Gaze estimation from low resolution images. In *Advances in Image and Video Technology*, pages 178–188. Springer, 2006.

[9] T. Proscevičius, V. Raudonis, A. Kairys, A. Lipnickas, and R. Simutis. Autoassociative gaze tracking system based on artificial intelligence. *Electronics and Electrical Engineering.–Kaunas: Technologija*, (5):101, 2010.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.