Evaluating the Robustness of an Appearance-based Gaze Estimation Method for Multimodal Interfaces

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

ABSTRACT

Given the crucial role of eye movements on visual attention, tracking gaze behaviors is an important research problem in various applications including biometric identification, attention modeling and human-computer interaction. Most of the existing gaze tracking methods require a repetitive system calibration process and are sensitive to the user's head movements. Therefore, they cannot be easily implemented in current multimodal interfaces. This paper investigates an appearance-based approach for gaze estimation that requires minimum calibration and is robust against head motion. The approach consists in building an orthonormal basis, or eigenspace, of the eye appearance with *principal component analysis* (PCA). Unlike previous studies, we build the eigenspace using image patches displaying both eyes. The projections into the basis are used to train regression models which predict the gaze location. The approach is trained and tested with a new multimodal corpus introduced in this paper. We consider several variables such as the distance between user and the computer monitor, and head movement. The evaluation includes the performance of the proposed gaze estimation system with and without head movement. It also evaluates the results in subject-dependent versus subject-independent conditions under different distances. We report promising results which suggest that the proposed gaze estimation approach is a feasible and flexible scheme to facilitate gaze-based multimodal interfaces.

Keywords

Gaze estimation, eigenspace analysis, computer user interface, multimodal interfaces

1. INTRODUCTION

The design of gaze-based computer interfaces has drawn increasing attention from the research community [13, 19, 21]. Gaze is a natural and fast modality that can be employed to interact with a system, especially for physically impaired individuals [4]. Even if gaze is not used as a user interface, adding gaze estimation to the system can provide useful information about the users' visual attention [1, 5]. Furthermore, these gaze-aware multimodal interfaces can be used to infer the emotional/cognitive state of the user (e.g., frustration, distraction, uncertainty). Although the problem of gaze detection has been studied for over 40 years [15], most *graphic user interfaces* (GUIs) nowadays are still reluctant to include gaze as an input modality. The main challenges of using many of the current gaze detection methods are the tedious calibration process and the sensitivity against variabilities observed in real world applications (e.g., illumination, head movement, individual differences).

The key goal for a gaze-based interface is to map the user's gaze behavior to the screen coordinates. The ideal gaze user interface should have easy, flexible and non-intrusive settings while maintaining high accuracy [20]. The most widely used gaze estimation approach is video-based eye trackers, where a camera is used to capture the eye movement as the user looks at the interface's screen [7]. A common approach consists in using an infrared / near-infrared non-collimated light, which produces reflections in the corneal [6, 16]. These approaches measure the vector between the pupil center and the corneal reflections, which is used to infer the gaze. These approaches require system calibration to estimate parameters related to the subject, data acquisition equipments and the system setting. Some parameters such as camera parameters and human eye curvature are consistent across time and only need to be estimated once. However, other parameters such as the relative location and orientation of the system setting have to be estimated for each session. As expected, these systems are not robust against head motion.

The two most common types of gaze behaviors are fixations and saccades [14]. The former is defined as fixing the gaze on a single location for a short period of time, and the latter as fast eye movements between two or more fixation areas. Both eye movements provide useful information, and have been investigated for different domains such as cognitive science, marketing research, and driver distraction [1, 10, 12]. For multimodal interfaces, fixations are arguably the most relevant gaze behavior. In fact, 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. Since humans eyes have to frequently perform saccade actions to perceive the visual scene, detecting the actual gaze may introduce jumpy trajectories that reduce the effectiveness of the interface. Our

ICMI'13, December 9–13, 2013, Sydney, Australia

http://dx.doi.org/10.1145/2522848.2522876.

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.

long term goal is to implement a gaze estimation approach that, while may produce lower accuracy than commercially available systems, (1) is robust against head motion and individual differences, and (2) requires minimum or no calibration. Such a system will be suitable for gaze-aware multimodal interfaces.

This study proposes an appearance-based method for gaze tracking based on *principal component analysis* (PCA). PCA-based approaches have been used for eye detection [8, 9] and gaze estimation [17, 18]. The approach uses image patches displaying both eyes to estimate an orthonormal basis. For each image, the coordinates with respect to this eigenspace are used to train linear regression models, which predict the target screen location.

A key contribution of this study is the multimodal corpus recorded to build and evaluate the proposed gaze estimation approach, which is named MSP-Gaze corpus. We record a database consisting of 30 participants using a commercial webcam and a Microsoft Kinect for Windows (RGB camera and IR depth-finding camera). The subjects were asked to locate points displayed on the screen under the following conditions: with and without head motion; and, different distances from the monitor. We evaluate the performance of the proposed gaze estimation system with user-dependent and user-independent conditions. We also evaluate the robustness and consistency of the system against head motion. This approach and the use of image patches displaying both eyes make the system robust in the presence of head movement, user-screen distances and subjects characteristics, as validated by the experimental results. These findings suggest that the proposed system can be effectively used in gaze-aware multimodal interfaces.

2. MSP-GAZE DATABASE

This study considers several factors that may affect the performance of appearance-based gaze estimation systems and that are important for gaze-aware multimodal interfaces. In particular, we focus on individual characteristics of the eyes, presence of head movement, and various distance between the user and the interface's screen. We collect the MSP-GAZE corpus to study gaze estimation approaches that are robust against these factors. The database is balanced in terms of gender, and includes a diverse ethnic representation of the students from the University of Texas at Dallas (Caucasian, as well as Asian, Indian, and Hispanic populations). While our target is 46 subjects, the database currently consists of 30 participants whose average age is 22.7. None of the participants used glasses during the recording. This section introduces the data collection protocol and preprocessing methods.

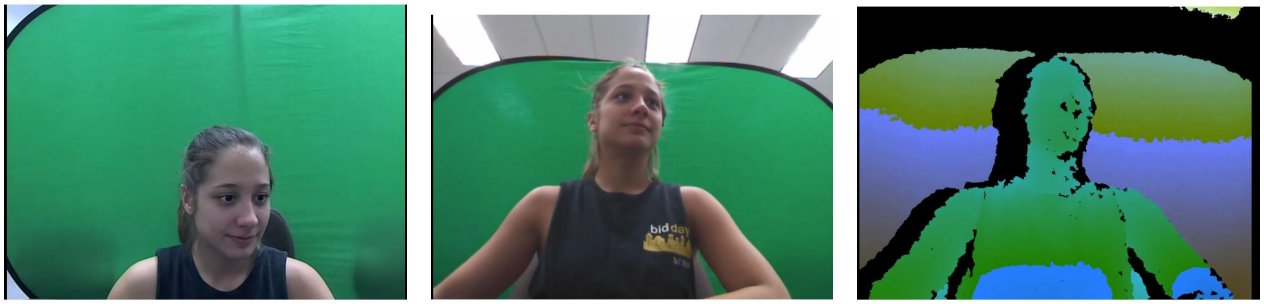
Table 1: Recordings conditions for each session.

Recording	Head Movement	Distance	Pattern
1	Yes	User-defined	Testing
2	Yes	User-defined	Training
3	Yes	Near	Training
4	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

2.1 Data Collection

We collect the MSP-GAZE corpus in our laboratory, which provides similar illumination to regular offices. The recordings consist of tracking the position of points displayed on the computer's monitor (see Fig. 1). We use a standard 22-inch HP monitor with the screen resolution set to 1680×1050. The videos of the subjects are recorded using both a commercial webcam (Logitech C920 – Fig. 2(a)) and a Microsoft Kinect for Windows (RGB camera and IR depth-finding camera – Figs. 2(b) and 2(c)). The resolution of both devices are set to 640×480 pixels. As shown in Figure 1, the webcam is placed on top of the monitor and the Kinect sensor is placed below the monitor. 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. In addition, the Kinect sensor also provides depth information (IR depth-finding camera) which can be used to estimate head pose, and distance between the monitor and the participant. Since the distance between the users and the computer is generally below two meters, we set the Kinect depth sensor to the near mode to estimate accurate depth images. A green screen is placed behind the subject to provide uniform background. Figure 2 shows sample images captured by the cameras.

We record each subject on two different days separated with an average interval of seven days. During each of these two sessions, we collect 14 recordings using the conditions described in Table 1. Each of them lasts approximately 3 minutes, and aims to capture the normal behavior of computer users (more details of the actual task is given in Sec. 2.2). We record unconstrained conditions first, in which the



(a) (b) (c)

Figure 2: Examples of images recorded: (a) Webcam, (b) Kinect RGB image, and (c) Kinect depth image.

subjects are free to move their head as they normally do while interacting with computers (first seven sessions). The only instruction to the subjects is to look at the points on the monitor. The subject selects his/her distance to the monitor for the first two recordings ("User-defined"). Then, the user-monitor distance is adjusted for each recording at "Near" (0.4 meter), "Medium" (0.5 meters) or "Far" (0.6 meters) distance. For consistency across sessions, a tape is placed on the desk to define the monitor positions for these three distances (see white tapes in Fig. 1). The subjects are asked to sit as close as possible to the desk. We collect two recordings per distance, except for the "Near" setting which is limited to one recording, since it was uncomfortable for some of the users. We repeat the same protocol for the second half of the recordings with exception that the participants are asked to maintain a steady head pose during the recordings (i.e., "No" head movement condition). The subjects can take short breaks between recordings. We also give ten-minute breaks at the middle of the session to avoid fatigue. The same protocol is repeated again for the second session, which is collected on a different day. For each subject, we have about 90 minutes of data over the two sessions.

2.2 Calibration Pattern

To develop a gaze estimation system that maps the captured eye appearance to the screen position, the data collection should efficiently cover different areas on the screen. Following previous studies [17, 18, 23], we divide the screen into 5 by 9 grids as illustrated in Figures 1 and 3. This division defines 45 grids, from which we only use the 23 grids marked with 'X' in Figure 3. We randomly generate a white point inside one of the 23 highlighted grid areas (i.e., the points will not appear at the same position in a grid more than once). The subject is asked to click on the point with the mouse cursor. The point turns to green once the user click on it and stays for 1 second, before jumping to a different location. We introduce the mouse click action in order to get the time at which the participant is looking at the target point (i.e., 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 a random order. This design ensures enough sample data while limiting the duration of the recording (i.e., 92 points collected in approximately 3 minutes). The videos from the cameras, the actual location of the points, the mouse cursor location, and the mouse click action are recorded for each frame. This protocol is used in the 12 recordings labeled as "Training" in Table 1.

X		X		X		X		X
	X		X		X		X	
X		X		X		X		X
	X		X		X		X	
X		X		X		X		X

Figure 3: The screen is divided into 45 grids. The "Training" pattern displays points in the 23 grids marked in the figure.

Two recordings in Table 1 are collected with a slightly different protocol. For these recordings, which are labeled 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. This pattern is applied to the first (with head movement) and the eighth recording (without head movement), as seen in Table 1.

3. PROPOSED APPROACH

The appearance of the eye provides useful information about the subject's gaze. An important aspect is to identify a compact representation from the image that is useful to estimate the gaze. We use *principal component analysis* (PCA) as a dimensionality reduction technique. An orthonormal basis is created which produces projections that are used as features in linear regression models. This section describes the preprocessing steps (Sec. 3.1), the eigenspace approach (Sec. 3.2) and the linear regression model to detect the gaze position (Sec. 3.3).

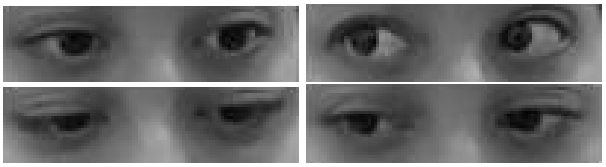


Figure 4: Eye pair samples extracted with the Viola-Jones algorithm. They correspond to cases where the subject was looking at points in the corners.

3.1 Preprocessing Steps

Gaze appearance methods commonly use images displaying an individual eye. In contrast, we propose to use a single image patch describing both eyes (see Fig. 4). The motivation of this approach is trifold. First, the relative location between two eyes and their appearance as captured by the cameras provide information about the subject's head movement. This information cannot be inferred by an image displaying a single eye. Second, in our preliminary analysis we observed that eye pair detection is more robust than single eye detection. Finally, eye pair appearance captures the inherent symmetry and distance heuristic of gazes that cannot be derived from a single eye.

We use the cascade object detector with the Viola-Jones algorithm for the eye pair image detection/extraction [25]. The Viola-Jones object detection framework provides competitive object detection rates using Haar-like features that describe the shape, relative position and orientation of the object-of-interest. We employ the implementation of the eye detector provided by the *open computer vision library* (OpenCV). In particular, we use the eye-pair detector developed by Castrillón et al. [3], which was trained with 7000 positive 11×45 eye pair images. Figure 4 shows examples of the detected eye pair images from the MSP-Gaze corpus.

As discussed in Section 2.2, we simultaneously record the videos, the location of the target points, the position of the mouse cursor, and the mouse click actions. We use the information to map the extracted eye pair image to the screen position. The subjects have to look at the target point before clicking the mouse. Therefore, we consider five eye pair images extracted right after the subject clicked the mouse (0.2 sec). Other eye pair images are ignored since it is not clear that the subject was still looking at the given point. A preliminary observation of our corpus suggests that subjects barely blink immediately after clicking the mouse. Since eyes are closed during eye blinking, these images introduce noise to the appearance-based gaze estimators. Therefore, using only these 5 images not only ensures reliable mapping between eye pair images and accurate screen position for training, but also eliminates the need of detecting eye blinking. Overall, 460 eye pair images (92 animation points $\times$ 5 frames) with their corresponding screen position are extracted for each recording described in Table 1. The images are transformed into gray scale images.

3.2 Eigenspace Approach

We represent the eye's image patch using the eigenspace approach, which has been successfully used in many computer vision problems such as face recognition (i.e., eigenfaces) [11, 24]. The approach consists in representing a set of N aligned images with an orthonormal basis estimated from the covariant matrix of the images. This basis is computed using *principal component analysis* (PCA), which defines a new coordinate system. The eigenvectors associated with the largest eigenvalues define directions with the highest variability. By considering only these eigenvectors, we can reduce the dimension of the feature vector while capturing most of the structure of the data. In this study, we use this approach to represent the essential components to reconstruct the eye pair appearance for various gaze position across different subjects.

The first step consists in organizing a set of N images with the same size. Since the extracted eye pair patches have different sizes, we resize each image patch, Γ_i, such that its size is 100×25 pixels. Then, we estimate the mean image Ψ, and the mean removed images Φ_i. Using the mean removed images, Φ_i, the sample covariance matrix of the eye pair patches can be calculated:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} \Phi_i \Phi_i^T \tag{1}$$

A basis of N orthonormal vectors is derived by computing the eigenvectors of the covariance matrix Σ. Once the eigenvectors are found, we select the eigenvectors associated with the largest 30 eigenvalues. This set of eigenvectors captures 90% of the variability in the eye pair patches. For a new image, we estimate the projections into the basis of this reduced space, generating a 30 dimensional vector.

The eye pair detector may fail to detect the eyes due to eye blinks and occasional extreme behaviors such as sneezing and hand occlusion. We validate the eye pair detection using the PCA analysis. For each recording, all detected eye pair images are normalized to the same size (100×25), and projected into the reduced eigenspace. Then, we reconstruct the image using only the 30 principal components. The error between the original and reconstructed images is used to detect false detections. A threshold is manually determined by studying the false detection images. When the error is greater than this threshold, we assume that the eye pair image is not correctly recognized, and the images are discarded for the analysis.

3.3 Linear Regression

The final step in our approach is to use the projections into the reduced eigenspace, $p_1, p_2, \cdots, p_{30}$, as features for two linear regression models. These models are separately built to estimate the gaze position in the horizontal (x) and vertical (y) coordinates, respectively. 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.

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

$$y = \begin{cases} 0, & \text{if } y < 0 \\ 1050, & \text{if } y > 1050 \\ \beta_{y0} + \beta_{y1}p_1 + \cdots + \beta_{y30}p_{30}, & \text{else} \end{cases} \tag{3}$$

4. EXPERIMENTAL RESULTS

This section evaluates the performance of the proposed approach to estimate the gaze of the subjects. For this study, we only consider eye pair images extracted from webcam videos, and we report results from data extracted from 30 subjects. We train user-dependent models in which data from one subject is used to build and evaluate the approach, and user-independent models, in which general models are validated with data from subjects that were not included in the training set. We also evaluate the effect of head movement, distance to the computer monitor, calibration pattern and consistency of gaze glance behavior.

4.1 Performance Metrics

Two measurements are used to assess the performance of the proposed gaze estimation approach: the correlation between the ground truth and the predicted value of the gaze position (ρ_x, ρ_y), and the angular error between the true and predicted gaze positions (θ_{error}).

To calculate the gaze direction, we assume that the eye pair center of the subjects is aligned with the center of the monitor. Therefore, θ_{error} is determined by:

$$\theta_{error} = \left| tan^{-1}(\frac{d_{p-mc}}{d_{u-m}}) - tan^{-1}(\frac{d_{pp-mc}}{d_{u-m}}) \right| \quad (4)$$

where d_{p-mc} is the distance between the position of the target point and the monitor center, d_{pp-mc} is the distance between the predicted gaze position and the monitor center, and d_{u-m} is the subject-monitor distance. For recordings under the "Near", "Medium" and "Far" settings, d_{u-m} is fixed during the recordings (see Sec. 2.1). For recordings under the "User-defined" setting, we estimate d_{u-m} 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_{u-m} is estimated by linearly interpolating the size of the eye pair image between the corresponding values for the "Near" and "Far" conditions. Notice that d_{u-m} can also be derived from the depth images provided by the Kinect sensor. For future work, we will rely on the depth image for more accurate distance information.

4.2 Subject-Dependent Evaluation

The subject-dependent evaluation corresponds to the most common case in gaze estimation studies (e.g., systems that require to estimate individual parameters such as eye curvature). A calibration process is established for each subject. In our approach, this calibration is used to estimate the eigenspace, and the linear regression models. Then, the system is evaluated with different images extracted from the same subject.

Our first question is to select an appropriate number of grids ($\#grids$) during calibration to estimate an accurate gaze estimation system. We address this question using subject-dependent condition. Notice that we displayed four points per grids in each of the recordings in Table 1. Therefore, the number of calibration points is given by $\#grids \times 4$. We train our model with 4, 9, 15 and 23 grids. The 4-grid case uses the four corners of the screen (see Fig. 3). The 9-grid case considers 3 points in the first, third and fifth rows of the grid pattern. These three points are located in the left, middle and right columns of the grid pattern. The 15-grid case includes all the grids except the ones in the second and fourth rows. The 23-grid case considers all the highlighted grids. In each of these cases, the proposed system is trained and tested with the "User-defined" distance. Figure 5 shows the "within session" results, in which we train and test the gaze detector with data from the same session. We use the 2nd and 9th recordings for training and 1st and 8th recordings for testing for the two head motion conditions (see Table 1). Likewise, Figure 6 shows the "between session" results, in which the training and testing data are the "User-defined" recordings from different sessions (i.e., collected in different days). These figures provide the angular error between the point's position and the estimated gaze position. These figures reveal that the error decreases as the number of grids increases. The results are consistent in both "within session" and "between session" conditions. The 23-grid case provides similar performance than the 15-grid case. However, it requires 32 extra calibration points (92 versus 60). Depending on the application, more grids may provide higher accuracy. To reduce the number of calibration points, we select the 15-grid calibration approach for the rest of the evaluation.

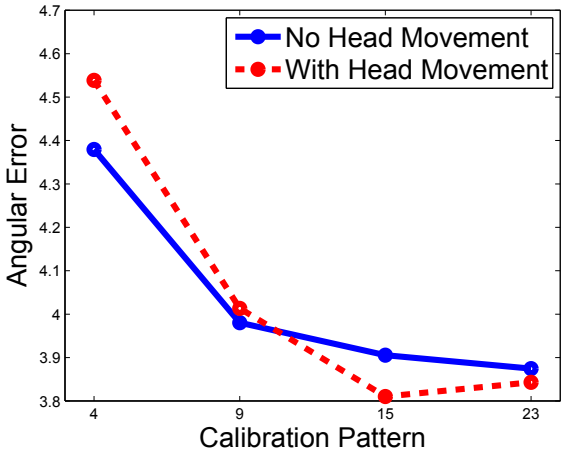


Figure 5: Within session results to evaluate the number of grids needed to train the gaze detection system.

To evaluate the gaze estimation for the subject-dependent models, we consider both "within session" and "between session" conditions. For "within session" evaluation, the models are trained and tested with data from a single subject during the same session. The evaluation considers matched distance and head motion conditions (i.e., training and testing with "Far" distance and without head motion). For each session and for each head motion condition, we collect two recordings for "Medium" and "Far" distances (4rd-7th and 11th-14th recordings in Table 1). For these cases, we train using the first recording and test the results on the second recording. We do not report results for the "Near" distance since we only collect one recording for each head motion condition. For the "User-defined" distance, we train our system using the training grid pattern (i.e., points displayed in the 15 grids) and we evaluate the performance with the testing pattern (i.e., points are randomly displayed in the monitor without following any grid – this is the most challenging case). The results are averaged across the 30 subjects. Ta-

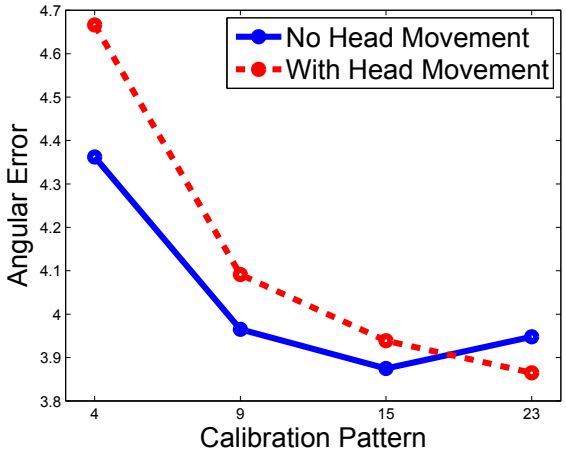


Figure 6: Between session results to evaluate the number of grids needed to train the gaze detection system.

Table 2: Subject dependent results within session. The "Near" condition is not evaluated since we only collect one recording per session (see Table 1).

	Without head motion			With head motion		
Distance	ρ_x	ρ_y	θ_{error}	ρ_x	ρ_y	θ_{error}
Near	–	–	–	–	–	–
Medium	0.89	0.84	4.0	0.90	0.82	4.2
Far	0.88	0.82	3.6	0.90	0.82	3.4
User-Defined	0.89	0.85	3.9	0.89	0.82	3.8

ble 2 shows the results for each distance condition, with and without head motion. the correlation of the regression models is around $\rho_x = 0.89$ for the x coordinate, and around $\rho_y = 0.83$ for the y coordinate, across conditions. This result indicates that the proposed gaze estimation approach is more accurate in the horizontal axis. This result is expected since the eyes present higher appearance differences in the x direction. The angular error is less that 4.3 degrees which is sufficient for many applications.

For "between session" condition, the models are trained and tested with data from different sessions collected in different days but from the same subject. Notice that this condition is ideal for multimodal interfaces since the calibration is conducted only once. After that, the subject can use the system without further calibration. The experiment follows the same matched conditions use for the "within session" condition (i.e., same distance and head motion mode). For the "Middle" and "Far" distances, there are two recordings for each of the two sessions (per head motion mode). We evaluate four permutations in which each of these recordings is used to train the models. These models are evaluated with corresponding recordings from the other session. For the "Near" distance, we only implement two permutations given that we collect only one recording per session and per head motion condition. For the "User-defined" distance, we use the training pattern recordings to build the system and the testing pattern recordings to evaluate the results. This setting defines two permutations per head motion mode. Table 3 reports the averaged results across permutation and subjects.

Table 3: Subject dependent results between sessions.

	Without head motion			With head motion		
Distance	ρ_x	ρ_y	θ_{error}	ρ_x	ρ_y	θ_{error}
Near	0.90	0.85	4.7	0.91	0.84	4.5
Medium	0.89	0.84	3.8	0.91	0.83	3.9
Far	0.88	0.83	3.5	0.90	0.83	3.4
User-Defined	0.89	0.82	3.9	0.88	0.82	3.9

Table 4: Subject Independent results.

	Without head motion			With head motion		
Distance	ρ_x	ρ_y	θ_{error}	ρ_x	ρ_y	θ_{error}
Near	0.85	0.76	7.0	0.87	0.75	6.8
Medium	0.86	0.75	6.0	0.85	0.74	5.9
Far	0.85	0.68	5.3	0.85	0.73	5.2
User-Defined	0.85	0.78	5.9	0.86	0.70	6.0

The results reported in Tables 2 and 3 reveal that the proposed appearance-based gaze estimation approach provides similar performance in "within session" and "between session" conditions. This result is very important since it indicates that once the eigenspace, and the linear regression models are built for a given subject, he/she can use the system without further calibration. This feature of the proposed approach is important for gaze-aware interfaces.

Tables 2 and 3 also show that the proposed approach is not affected by the subject's head motion. The evaluations with and without head movements provide similar gaze estimation accuracy in both tables. This result is also observed in Figure 5 and 6. While the performance of the system is affected when only four grids are used to estimate the eigenspace, the performance for both head motion conditions is similar when 15 or 23 grids are used. We also notice that the distance between the user and the computer monitor has little affect on the accuracy of the system. The fact that the approach is not sensitive to head motion and the user-interface distance suggests that it is suitable for real applications, especially in cases when the gaze is used to estimate the user's attention.

4.3 Subject-Independent Evaluation

This section studies how well the eigenspace and linear regression models can be generalized where the testing user is not considered during training. This case is appealing since it eliminates the calibration process when new users interact with the system (i.e., the models can be trained offline with previous recordings from other subjects).

The methodology used for the subject-independent evaluation is similar to the one described in Section 4.2. We evaluate the approach under distance and head movement matched conditions. The main difference is that training and testing set include data from disjoint groups of subjects. To maximize the usage of the corpus, we conducted a cross-validation approach in which the models are trained with data from 29 subjects, and the performance is evaluated with data from the remaining participant (i.e., 30 different models).

Table 4 shows that the performance drops for the subject-independent condition comparing to the subject-dependent condition. We noticed that the prediction of the vertical position is more affected. However, the correlations of the predicted gaze position are still higher than 0.85 for horizontal direction, and 0.68 for vertical direction. The average angular error is about 6.0 degree across all conditions. These results reveal that the eigenspace derived from multiple subjects provides an appropriate representation of the changes in eye appearance due to glance behavior of individuals that are not considered during training. This promising result suggests that the calibration process can be the completely eliminated, since the models are trained with data from multiple subjects in the corpus. Furthermore, we can include samples from more training grids to improve the performance, since the training is completed offline.

4.4 Comparison to Previous Work

Table 5 compares our results to other appearance-based gaze tracking systems. We compare the results achieved in subject-dependent condition with "User-defined" distance, and with head motion. Notice that the settings are not equal across the studies. For example, Tan et al. [23] estimated the error using "leave-one-out" cross validation on the training data instead of on the independent testing data. Williams et al. [26] used the same grid pattern for the training and testing data. In our case, the testing points are randomly displayed in the screen, even in regions not covered by the highlighted grids. Higher accuracies are observed when the number of calibration points increases. As mentioned before, our approach can be implemented without calibration during testing which is a major advantage for gaze aware interfaces.

5. CONCLUSIONS

This paper proposed an appearance based gaze estimation method that is based on the eigenspace approach and linear regression models. It introduces a new multimodal database recorded from 30 subjects, which is carefully collected to evaluate various sources of variability that are commonly observed in human computer interfaces (e.g., different user-monitor distances and use of head movements). The proposed gaze estimation approach is systematically evaluated under different conditions. The promising experimental results demonstrate that the approach is robust against head motion. The accuracy of the system is not affected when the eigenspace and linear regression models are built/trained with data from one session and tested with data from another session collected during a different day. This result indicates that the calibration step can be implemented only once. The approach is also evaluated under subject - independent conditions (i.e., general models are evaluated with data extracted from subjects not included in the training of the models). While the performance decreases compared with the subject-dependent case, the accuracy of the system is still competitive. This result suggests that the proposed approach can be used without any calibration.

We will explore in our future work techniques to improve the performance of the gaze estimation approach under subject-independent conditions. This condition is the most interesting setting from the perspective of multimodal interfaces, since it does not require calibration. Our goal is to approach the performance achieved with subject-dependent condition. We are exploring various methods to achieve this goal including factor analysis, localized regression and whitening transformation. A limitation of the data collection is that we only recorded subjects without glasses. Also, we provided ideal illumination during the data collection. These conditions are not realistic in many real applications, and robust methods are needed. Our future data collection effort will include these different challenging conditions. Likewise, we will explore the feasibility of implementing the proposed method in mobile devices. One major difference from the current setting is that the position of the screen is not fixed. Therefore, the distance and perspective between the user and interface is time variant. In spite of these challenges, appearance based models may offer a good tradeoff between accuracy and complexity in this context. The MSP-GAZE corpus and the proposed method can be used as a starting point to address these open challenges.

Acknowledgments

This study was funded by Samsung Telecommunications America and the US National Science Foundation under grant IIS-1217104.

6. REFERENCES

[1] G. Anders. Pilot's attention allocation during approach and landing- eye-and head-tracking research in an a 330 full flight simulator. In *International Symposium on Aviation Psychology (ISAP 2001)*, Columbus, OH, USA, March 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] L. Frey, K. W. Jr, and T. Hutchison. Eye-gaze word processing. *IEEE Transactions on Systems, Man and Cybernetics*, 20(4):944–950, July/August 1990.

[5] C. Ghaoui. *Encyclopaedia of Human Computer Interaction*. Idea Group Reference, Hershey, PA, USA, December 2005.

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

[7] 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 2010.

[8] P. Hillman, J. Hannah, and 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)*, pages 359–364, Rome, Italy, September 2003.

[9] W. Huang and R. Mariani. Face detection and precise eyes location. In *International Conference on Pattern Recognition (ICPR 2000)*, volume 4, pages 722–727, Barcelona, Spain, September 2000.

Table 5: Comparison of our approach with other available systems. The number of calibration points for new users in our approach is zero for the subject-independent condition.

System	Angular Error	Calibration Point	Head movement	Light
Tan et al.[23]	$< 0.5^\circ$	252	-	Yes
Baluja and Pomerleau [2]	1.5°	2000	Yes	Yes
Stiefelhagen et al.[22]	$< 2.0^\circ$	4000	Yes	No
Williams et al.[26]	1.3°	16	-	No
Proposed approach	3.9°	60	Yes	No

[10] K. K. Rayner, C. Rotello, A. Stewart, J. Keir, and S. Duffy. Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, 7(3):219–226, September 2001.

[11] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, January 1990.

[12] N. Li and C. Busso. Analysis of facial features of drivers under cognitive and visual distractions. In *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, May 2013.

[13] 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.

[14] G. McConkie and K. Rayner. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586, November 1975.

[15] J. Merchant, R. Morrissette, and J. Porterfield. Remote measurement of eye direction allowing subject motion over one cubic foot of space. *IEEE Transactions on Biomedical Engineering*, BME-21(4):309–317, July 1974.

[16] T. Ohno, N. Mukawa, and A. Yoshikawa. FreeGaze: a gaze tracking system for everyday gaze interaction. In *Symposium on Eye tracking research & applications (ETRA 2002)*, pages 125–132, New Orleans, LA, USA, March 2002.

[17] Y. Ono, T. Okabe, and Y. Sato. Gaze estimation from low resolution images. In L.-W. Chang, W.-N. Lie, and R. Chiang, editors, *Advances in Image and Video Technology*, volume 4319/2006 of *Lecture Notes in Computer Science*, pages 178–188. Springer-Verlag Berlin Heidelberg, Hsinchu, Taiwan, December 2006.

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

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

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

[21] 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, November-December 2011.

[22] R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Workshop on Perceptual User Interfaces (PUI-1997)*, pages 98–100, Banff, AB, Canada, October 1997.

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

[24] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, Winter 1991.

[25] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 511–518, Kauai, HI, USA, December 2001.

[26] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the S^3GP. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, volume 1, pages 230–237, New York, NY, USA, June 2006.