A Non-Intrusive Approach to Gaze Estimation

Stephen Bialkowski, Zeb Johnson, Kathryn Morgan, Xiaojun Qi, and Donald H. Cooley

Computer Science Department, Utah State University, Logan, UT 84322-4205
stepbial@cc.usu.edu; zebgjohnson@hotmail.com; kjmorgan@cc.usu.edu; xqi@cc.usu.edu; don.cooley@usu.edu

Abstract

Human eye tracking and gaze estimation is a new technology which acts as an interface between the user and the computer. Its main goal is to determine where a user is looking on the computer screen. Numerous methods have been developed to accomplish this task. However, they need expensive and proprietary hardware, or require the user to wear a device on his/her head, or place limitations on the user's head movement. In this paper, we propose an inexpensive and non-intrusive algorithm to estimate the eye gaze on the computer screen. The proposed algorithm can quickly determine the glints and pupils of the user under normal lighting conditions and estimate the gaze of the user by a pretrained scaled conjugate gradient based neural network. Extensive experimental results demonstrate the promising robustness of the algorithm on estimating the gaze of users under normal lighting conditions and with moderate head movement, regardless of whether users wear glasses or contacts.

Keywords: Eye tracking, gaze estimation, and neural network.

1. Introduction

Human eye tracking and gaze estimation is a rapidly growing research area with many useful human-computer interaction applications, such as video conferencing [1] and eye-typing [2]. Gaze tracking can also be used in monitoring driver alertness [3] by observing the appearance of the eyes for drowsiness or fatigue. The improvement of eye tracking and gaze estimation systems also provides a substitute means for physically impaired persons who find it difficult to, or are completely unable to, control a mouse. In this research, we focus on developing an inexpensive and non-intrusive gaze estimation system. The proposed system requires little calibration for new or return users to the system and is robust against normal lighting conditions and moderate head movements.

The existing non-intrusive research in eye detection, tracking, and gaze estimation can be classified into three categories: passive image-based approaches, active infrared-based approaches, and hybrid approaches.

The passive image-based approaches [4-6] detect and track eyes by exploiting the unique intensity distribution (e.g., dark pupil and white sclera) or shape (e.g., circular iris and eye corners) of the eyes to distinguish the human eye from other objects. Three typical methods are template-based, appearance-based, and feature-based. These traditional passive image-based methods achieve decent eye tracking results for images with good contrast when faces are in the frontal orientations and eyes are without closure and occlusion. However, they cannot work well for different subjects under different illuminations.

Active infrared-based approaches [7-9] exploit the spectral (reflective) properties of pupils under near infrared illumination to produce the bright/dark pupil effect for eye detection and tracking. However, these methods require distinct bright/dark pupil effect to work well and heavily depend on the brightness and size of the pupils, which are often affected by eye closure and occlusion due to face rotation, external illumination interferences, and the distances of the subjects to the camera. In addition, they require a sophisticated control system and an expensive camera capable of generating interlaced images using even and odd fields.

Several systems combine both passive image-based and active infrared-based approaches to address the aforementioned problems. Haro et al. [10] propose to perform pupil tracking using the conventional appearance-based matching method, the bright pupil effect, and the motion characteristics. However, this method cannot track the closed or occluded eyes or eyes with weak pupil intensity due to external illumination interference. Ji et al. [11] perform pupil verification based on the shape and size of pupil blobs to eliminate spurious pupil blobs after eliminating the external light interferences using the real time subtraction and a special filter.

In this paper, we propose a non-intrusive eye gaze estimation system with the aid of inexpensive equipment. The remainder of the paper is organized as follows. Section 2 presents the proposed algorithm. Section 3 shows the experimental results. Section 4

concludes the paper and summarizes the directions for future work.

2. Proposed Algorithm

The configuration of the proposed system is shown in Fig. 1. We built an inexpensive hardware to facilitate the process of finding glints and pupils. This hardware consists of a digital video camera, an infrared filter, several infrared lights, and an infrared light controller board. The infrared lights are arranged in two sets: close to and away from the camera axis. The on status of the lights close to the camera axis brightens the pupils and the on status of the lights away from the camera axis brightens the glints. The off status of both lights darkens the pupils and glints, respectively. In our research, we turn on the infrared lights away from the camera axis to capture a pattern of double bright glints with a dark spot (the pupil) above them. We take a sequence of images at the rate of 1 frame per 2 seconds from the camera positioned in front of the user. Three major steps of the proposed algorithm, namely, finding the glint centers, finding the pupil centers, and extracting the features for the scaled conjugate gradient based neural network, are summarized in the following subsections.

Fig. 1: The system configuration

2.1. Find the glint centers

The algorithmic view of finding the glint centers are:

Step 1: Generate a binary image ***S*** using the saturation image in HSV color space. The threshold is empirically set to be half of the maximum intensity (i.e., any intensity smaller than the threshold is set 1).

Step 2: Generate a binary image ***D*** by converting the color image to the grayscale image and setting the threshold as 60 (i.e., any intensity smaller than the threshold is set 1). Apply the opening operation on ***D*** with a disk of one pixel radius to remove small objects. Then, apply the close operation with the same structuring element to remove noise and fill holes.

Step 3: Dilate the intersection of ***S*** and morphologically processed ***D*** with a disk of 15 pixel radii to obtain an image ***SD***. If two relatively similar regions are found in ***SD*** (the similarity measure is summarized later in this section), go to Step 6. Otherwise, go to Step 4.

Step 4: Generate a binary image ***L*** by converting the color image to the grayscale image and setting the threshold as 130 (i.e., any intensity larger than the threshold is set 1). Remove all regions with an area greater than 70 pixels. Apply the dilation operation with a disk of 10 pixel radii, a maximum distance between a glint and a pupil in all our experiments, to get a bigger candidate region.

Step 5: Intersect the morphologically processed ***D*** and ***L*** to obtain an image ***DL***. If two relatively similar regions are found in ***DL***, go to Step 6. Otherwise, intersect ***SD*** and morphologically processed ***L*** to obtain an image ***SDL***. If two relatively similar regions are found in ***SDL***, go to Step 6. Otherwise, the glint information in the previous frames is used to update the missing glint center(s).

Step 6: Generate a binary image, which contains all pixels of the relatively bright glints, for each of two similar regions using a base threshold ***BT***. This threshold is computed as follows: 1) Start with the initial local threshold T_1 in each of two similar regions, where T_1 equals (max+mean)/focusMeasure with max and mean respectively being the maximum and average grayscale intensity of the region, and focusMeasure being 1.7, which is experimentally determined to be optimal for a well-focused image. In our system, any intensity value smaller than T_1 is set 0. 2) If the number of 1's in the binary image obtained in 1) reaches a reasonable lower bound ***Num***, the base threshold is set as T_1. 3) Otherwise, compute a series of new threshold T_2's by increasing T_1 with an accumulative increment proportional to the natural log of the number of 0's in each binary image obtained by using the previously un-incremented threshold until the number of 1's in the current binary image has reached ***Num***. The last threshold in the series is ***BT***.

Step 7: Generate the final glint region for each of two binary images of the similar regions using a second threshold ***ST*** to avoid excessive reflections. This threshold is initially set to ***BT*** and is simply incremented by a constant value until there are less than three small regions containing 1's or until the new threshold achieves 95% of the maximum intensity in each region. The final glint region is determined by overlapping both regions obtained by using ***BT*** and ***ST***. If more than two glint regions are yielded due to excess reflections, two most horizontally aligned regions are chosen.

The similarity measures for comparing two regions are computed as the sum of the following:

1. The normalized differences in region area, orientation, eccentricity, and width.

2. The slope from one region to another. If the slope exceeds 0.2, the slope value is added.
3. The ratio of the distance between two regions and the expected distance if the distance is smaller than the expected.

The expected distance computed from previous frames may be used to further penalize each region pair that does not comply. Smaller similarity values indicate more similarity.

2.2. Find the pupil centers

The algorithmic view of finding the pupil centers are:

Step 1: Remove the glints from the input image and fill the holes by region filling. This step ensures the proposed system functions well when the glints overlap the pupils.

Step 2: Find two subimages $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$, whose centers are the glint centers reflected from the left and right eyes and whose size is 65×121.

Step 3: Apply the canny edge detector to find the edges in $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$.

Step 4: Find the darkest region $\boldsymbol{P}$ in each small box (34×61) near each glint center. Specifically, the small box covers the region of 3 pixels above and 30 pixels below the glint center. This choice is mainly because the pupil center is always above the glint pair center when the user is looking at the screen.

Step 5: Find the pupil boundary by working outwards from the darkest pixel in each small box until an edge is found. Close the region and calculate the pupil's center in each region.

2.3. Extract features for training

Several features are extracted and fed into the scaled conjugate gradient based neural network to learn the gaze coordinates on the computer screen. These features are: two glint pair center coordinates within the subimage that contains the eye, two pupil pair center coordinates within the subimage that contains the eye, average position between glint pair centers, angle and distance from each glint center to its corresponding pupil center, and angle and distance from the left glint pair center to the other. To simplify the training process, we exclusively train the gaze coordinates on 9 positions even laid out by a 3×3 grid shown in Fig. 2. A user friendly graphical interface is designed to instruct the users to look at each of the 9 coordinates shown as a blinking icon for approximately 40 seconds.

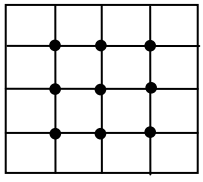


Fig. 2: Nine gaze coordinates for training and testing

The same features are extracted for a new image and fed to the pretrained network to estimate any of the nine gaze coordinates on the screen.

A neural network with a conjugate gradient based training algorithm is used to alleviate the assumptions made in the strictly geometric approach. The scaled conjugate gradient based neural network is chosen in our system because of its efficiency advantage over other conjugate gradient based neural networks.

3. Experimental Results

To date, we have collected training data from 5 persons with or without glasses (contacts) under normal lighting conditions and with moderate head movement. We then estimate the gaze coordinates for each person to evaluate the performance of the proposed system. Specifically, each person look at any of the 9 coordinates shown in Fig. 2 at his/her will. The ground truth of each gaze coordinate is recorded to compare with the gaze estimation computed from the pretrained neural network on the same person.

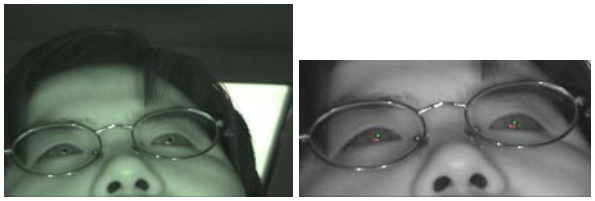
Fig. 3 demonstrates the intermediate results of several important steps on a person wearing a glass. It clearly shows that the proposed system can correctly extract the glint and pupil centers even with the reflected lights along the frame of the glasses.

Table 1 summarizes the average accuracy of the pupil and glint detection and the gaze estimation for each person. The first column lists the characteristics of each person. The second column lists the average accuracy in detecting pupils and glints for each image. The third column lists the average accuracy in detecting pupils and glints using the glint information from the previous frame when one glint is missed. The last column lists the linear regression correlation coefficients for *x*- and *y*-coordinates between the estimation and the ground truth, the mean squared error between the estimation and the ground truth, and the standard deviation for the error in inches, respectively. The experimental results demonstrate that the proposed approach achieves 72.2% accuracy in detecting pupil and glint without using the previous frame, 92.2% accuracy in detecting pupil and glint using information from the previous frame, and 0.0017 mean squared error in gaze coordination estimation in average. The average error for the gaze estimation is within 0.4742 inches. It also shows that the proposed system performs well regardless of whether a person wears glasses or contacts. That is, the reflection caused by the glasses or contacts can be filtered out by our image processing techniques.

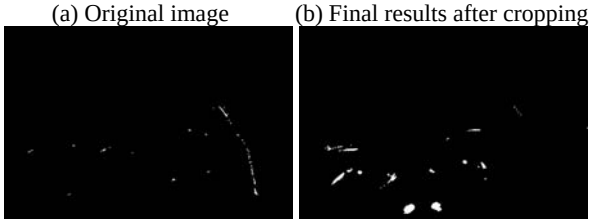
4. Conclusions and Future Work

In this paper, we propose an inexpensive and non-intrusive algorithm to estimate the eye gaze coordinate on the computer screen. It can quickly determine the glints and pupils of the user under normal lighting conditions and estimate the gaze of the user by a pretrained network. Extensive experimental results demonstrate the robustness of the algorithm under normal lighting conditions and moderate head movement, regardless of whether the user wears glasses or contacts.

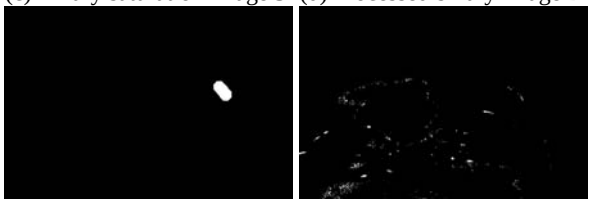
The current system takes about 1 second for locating the glints and pupils and extracting the features using Matlab 7.0.1 on Dell Precision 530 with Intel Xeon Processor at 3.06 GHz and 1 GB of RAM. We will further improve our prototype to make it a real-time system. We will also involve different races of subjects in our experiments and test our system under more lighting conditions and find the range of the allowable head movement.



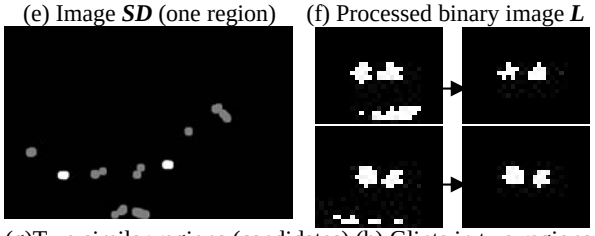
(a) Original image (b) Final results after cropping



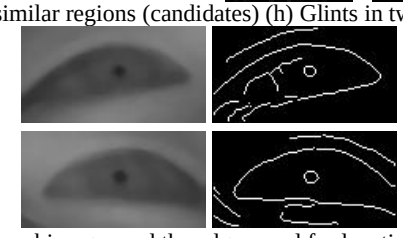
(c) Binary saturation image ***S*** (d) Processed binary image ***D***



(e) Image ***SD*** (one region) (f) Processed binary image ***L***



(g)Two similar regions (candidates) (h) Glints in two regions



(i) Two subimages and the edges used for locating pupils

Fig. 3: Intermediate results of some important steps

Table 1: Accuracy of the proposed system

Subjects	Initial Results	Refined Results	Gaze Estimation
W/o Glasses (American)	73%	98%	0.999; 0.997; 0.000638406; 0.3411
W/o Glasses (Asian)	66%	90%	0.999; 0.999; 0.000120807; 0.1484
W/o Glasses (American)	81%	93%	0.996; 0.99; 0.00152432; 0.5271
W Glasses (Asian)	71%	85%	0.986; 0.979; 0.00531655; 0.9844
W Contacts (Asian)	70%	95%	0.998; 0.996; 0.000750916; 0.36994
Average	**72.2%**	**92.2%**	**0.9956; 0.9922; 0.0017; 0.4742**

5. References

[1] D. Machin, L. Q. Xu, and P. Sheppard, "A Novel Approach to Real-time Non-intrusive Gaze Finding," *Proc. of British Machine Vision*, pp. 428-437, 1998.

[2] A. S. Johansen, D. W. Hansen, J. P. Hansen, and M. Nielsen, "Eye Typing Using Markov and Active Appearance Models," *Proc. of IEEE Workshop on Applications of Computer Vision*, pp. 132-136, 2002.

[3] Q. Ji and X. Yang, "Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance," *Real-Time Imaging*, Vol. 8, pp. 1077-2014, 2002.

[4] K. M. Lam and H. Yan, "Locating and Extracting the Eye in Human Face Images," *Pattern Recognition*, Vol. 29, pp. 771-779, 1996.

[5] W. M. Huang and R. Mariani, "Face Detection and Precise Eyes Location," *Proc. of Int. Conf. on Pattern Recognition*, pp. 722-727, 2000.

[6] G. C. Feng and P. C. Yuen, "Multi-Cues Eye Detection on Gray Intensity Image," *Pattern Recognition*, Vol. 34, pp. 1033-1046, 2001.

[7] Y. Ebisawa, "Improved Video-Based Eye-Gaze Detection Method," *IEEE Trans. on Instrumentation and Measurement*, Vol. 47, No. 2, pp. 948-955, 1998.

[8] C. Morimoto, D. Koons, A. Amir, and M. Flickner, "Pupil Detection and Tracking Using Multiple Light Sources," IBM Technical Report, 1998.

[9] Q. Ji and X. Yang, "Real Time Visual Cues Extraction for Monitoring Driver Vigilance," *Proc. of Int. Workshop on CVS*, pp. 107-124, 2001.

[10] A. Haro, M. Flickner, and I. Essa, "Detecting and Tracking Eyes by Using Their Physiological Properties, Dynamics, and Appearance," *Proc. of Int. Conf. on CVPR*, pp. 163-168, 2000.

[11] Q. Ji and Z. Zhu, "Eye and Gaze Tracking for Interactive Graphic Display," *J. of Machine Vision and Applications*, Vol. 15, No. 3, pp. 139-148, 2004.