# Discrimination of Gaze Directions Using Low-Level Eye Image Features

**Yanxia Zhang**
Lancaster University
United Kingdom
yazhang@lancaster.ac.uk

**Andreas Bulling**
University of Cambridge
& Lancaster University
United Kingdom
andreas.bulling@acm.org

**Hans Gellersen**
Lancaster University
United Kingdom
hwg@comp.lancs.ac.uk

## ABSTRACT

In mobile daily life settings, video-based gaze tracking faces challenges associated with changes in lighting conditions and artefacts in the video images caused by head and body movements. These challenges call for the development of new methods that are robust to such influences. In this paper we investigate the problem of gaze estimation, more specifically how to discriminate different gaze directions from eye images. In a 17 participant user study we record eye images for 13 different gaze directions from a standard webcam. We extract a total of 50 features from these images that encode information on color, intensity and orientations. Using mRMR feature selection and a k-nearest neighbor (kNN) classifier we show that we can estimate these gaze directions with a mean recognition performance of 86%.

### Author Keywords

Wearable eye tracking, Gaze estimation, Appearance-based, Low-level image features, Machine learning

### ACM Classification Keywords

I.5.4 Pattern Recognition: Applications; I.4.7 Image processing and computer vision: Feature Measurement

### General Terms

Algorithms, Experimentation

## INTRODUCTION

Eye tracking has a long history in human-computer interaction (HCI) and has contributed considerably to our understanding of visual perception and attention. In HCI, eye trackers have successfully been used for applications such as gaze communication, gaze-based typing or usability studies. Despite considerable advances in tracking accuracy and speed, most video-based eye trackers are still stationary and restrict free movements of the user's head and body. In addition, these systems require specialized hardware including high resolution video cameras and infrared illumination.

The advent of mobile eye trackers promises new applications for continuous eye tracking and analysis 24/7 [3]. Daily life settings require compact eye trackers that are easy to setup and use and adaptable to a particular user. For some of these applications, e.g. for eye-based activity and context recognition [4], accurate gaze tracking may not be necessary. For the majority of applications, however, mobile eye trackers need to provide robust methods for pupil tracking and gaze estimation in the presence of artifacts caused by head and body movement, ever-changing lighting conditions, as well as occlusion from eye lashes, eyelid and glasses.

### Goal and Contributions

In this paper, we propose a new method for gaze estimation that relies on low-level image features and machine learning techniques. This approach is potentially more robust to varying lighting conditions, head movements and requires less computation than current approaches. The specific contributions of the work are 1) the development of a set of 50 low-level image features used in computer vision research suitable for gaze estimation, 2) a data set of eye images recorded for different gaze positions simultaneously using a video-based eye tracker and a standard webcam, and 3) the evaluation of a learning method to map these image features to different gaze directions.

### Outline

We first describe related work and the experiment that we conducted to collect the data set. We then describe the low-level features we extracted from the raw eye images and evaluate a machine learning approach to map these features to different gaze directions. We present the first results of this evaluation and conclude with a summary and an outlook to future work.

## RELATED WORK

### Model-based approaches

Model-based approaches use an explicit geometric model of the eye to estimate gaze direction. Pupil and iris parameters, orientation and ratio of the major and minor axes of the pupil ellipse, as well as pupil-glint displacement are examples of very popular geometric eye features [8]. An example of such system can be found in [17] which is based on the Pupil Center Corneal Refection (PCCR) techniques. Model-based methods can be very accurate but they typically require specialized hardware and infra-red (IR) illumination.

*PETMEI'11*, September 18, 2011, Beijing, China.

These systems have performance issues in outdoors or under strong ambient light. In addition, the accuracy of gaze estimation decreases when accurate iris and pupil features are not available, e.g. due to occlusions from eye lashes or eyelid or due to blinking. These limitations make it hard to apply model-based approach for gaze estimation in mobile eye tracking.

### Appearance-based approaches

Appearance-based methods can make the system less restrictive with cheaper and easier hardware equipment. Calibration of the cameras is typically not required as gaze mapping is learned directly from raw image data [8]. Baluja *et al.* [1] suggested a method using a regression neural network in which intensities of 15x30 eye images were used to map the screen coordinates using 2000 training samples. In their work, the eye was located by searching for the specular reflection of a stationary light in the image of the user's face. Similar neural network-based approaches can be found in [11, 16]. The main limitation of these approaches is the fact that they require a large set of training data. In addition, they are computationally intensive for high resolution images as increasing amount of hidden nodes are required in the network. Hansen *et al.* [7] developed an eye typing system using Markov and active appearance models. They adopted a Gaussian process interpolation method for gaze estimation. Williams [15] proposed a sparse, semi-supervised Gaussian process regression method to map input image to the gaze coordinates with partially labeled training samples. Basilio *et al.* [9] developed a calibration-free eye gaze detection system which is also based on gaussian processes. Other gaze estimation work using appearance-based approaches based on appearance manifolds can be found in [13, 12]. While all of these papers investigated means to estimate gaze from low resolution images without IR illumination, in real-world settings their approaches face considerable challenges with artefacts caused by head and body movements.

### Eye Image Features

Image intensity provides powerful features that are widely used in appearance-based gaze estimation methods (see [1, 11, 16, 15] for examples). Similar to intensities, the color distribution in the eye region is different for different gaze directions. While RGB histograms are often used to represent image color information, the RGB color system depends on image characteristics. Gevers *et al.* [6] found that while converting to a color invariant system such as normalized *rgb* the color model is less sensitive to illumination and object pose. A survey by Hansen *et al.* [8] showed that applying different filters on images will result in enhancing particular image characteristics while suppressing others. Daugman *et al.* [5] showed that a set of Gabor filters with different frequencies and orientations can be used for iris recognition. Williams *et al.* [15] applied steerable filters to the eye images for gaze estimation.

In computer vision, researchers have investigated a large variety of image features for applications such as object detection and tracking or image segmentation. Common features describe intensity and color of image pixels or are based on filter responses. These feature types have been widely used in computer vision but haven't been extensively explored in eye tracking research.

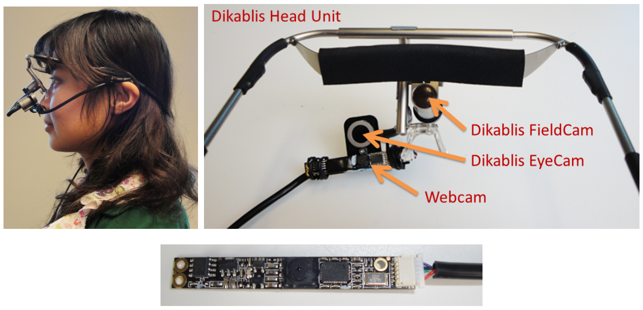


**Figure 1. Participant wearing the Dikablis eye tracker. The eye camera was adjusted to point to the participant's left eye. The webcam is mounted underneath to get close-up eye images.**

## EXPERIMENT

We conducted an experiment to collect a data set of naturalistic eye images. For the sake of a later comparison we simultaneously collected eye images from both a video-based eye tracker with IR illumination as well as from a standard webcam. We collected data from 17 participants, 5 female and 12 male, aged from 18 to 40 years (mean: 26.9 years, std: 6.8 years) of different ethnicities and with different eye colors. None of the participants wore glasses but two wore contact lenses during the experiment.

### Apparatus

The experimental system consisted of a webcam, a monitor and the Dikablis eye tracker from Ergoneers GmbH (see Figure 1). We used the Microdia Sonix USB 2.0 Camera with a maximum frame rate of 30Hz. The webcam was fixed on a plastic frame and mounted under the Dikablis eye camera to capture images of the eye with sufficient resolution. The sampling rate for the Dikablis eye tracking system was 25Hz. The resolution of the captured eye images was 640x480 for the webcam and 384x288 for the Dikablis.

### Setup and Procedure

The experiment was carried out in a real office with normal lighting conditions. Participants wearing the Dikablis headunit were seated at distance of approximately 60cm from a 23" 1680x1050 pixels ( 43° in horizontal and 27.6° in vertical of visual angle) computer screen. Free movements of the head and the upper body were possible, but the participants were instructed to sit as static as possible. The visual stimulus was shown on the screen as a red point with a radius of 20 pixels (0.5° of visual angle) on light grey background.

Images of the participants' left eye were recorded using the webcam and the Dikablis eye camera and labeled according to the current gaze direction on the screen (see Figure 2 for examples). Data synchronization was handled using the Context Recognition Network (CRN) Toolbox [2]. The CRNT streams data (such as the frame index of webcam images, video streams from the Dikablis eye and field cameras,

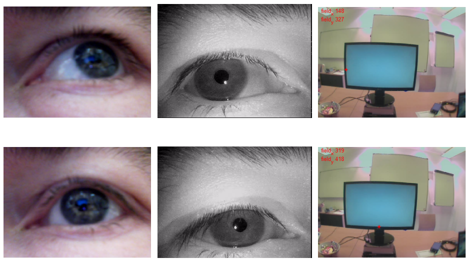
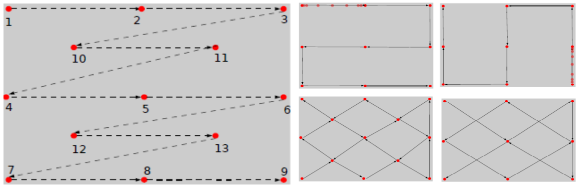


**Figure 2. Examples of recorded eye images. The images to the left are from the webcam. The grey images in the center and the images to the right are from the Dikablis eye and field camera, respectively. The red point on the screen shows the point of gaze.**



(a) (b)

**Figure 3. Screenshot of the experimental stimulus. (a) A red point is displayed in order at 13 different locations on the screen. It stays at each location for 5 seconds and moves to the next location. Location** $1$ **and** $10$ **are spaced 10.75° in horizontal and 6.9° in vertical of visual angle from each other. (b) The point moves horizontally, vertically and diagonally at constant speed.**

gaze coordinates in the scene image, and point labels from the stimulus displayed on the monitor) into a single file.

The experiment consisted of two sessions. In the first session, participants were instructed to fixate at 13 different locations indicated by the red point remaining at these locations for 5 seconds (see Figure 3(a) for an example). In the second session, participants were asked to follow the moving point with their eyes along several predefined paths (see Figure 3(b)). For each path the point moved horizontally, vertically and diagonally at constant speed. Each of these sessions was performed three times, each lasting for about 7 minutes. This resulted in a total data set of about 21 minutes.

## DISCRIMINATION OF GAZE DIRECTIONS

We use an appearance-based approach for gaze estimation. In contrast to model-based approach, we do not use explicit geometric features such as pupil-glint vector, the relative position of the pupil center, or contours and eye corners. Instead, we represent each eye image by a vector of low-level image features and use machine learning techniques to map these features to different gaze directions.

### Feature Extraction and Selection

We followed the work of [14] and calculated three types of low-level features: Color (C), Intensities (I), Orientations (O). All of these features were extracted offline using MATLAB (see [14] for the saliency toolbox we used).

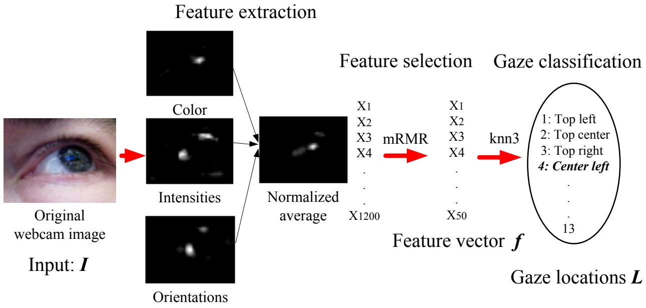


**Figure 4. An example of mapping raw image data to the gaze point. Input image *I* is converted to a feature vector *f*. *f* includes 50 features that carry information on color, intensities and orientations. A subset of these features is selected using mRMR. Finally, the feature vector *f* is classified into different gaze locations *L* using a kNN classifier.**

As illustrated in Figure 4, the raw input image denoted with $I$ is processed in three feature extraction steps: $f_C$ extracts the red-green (RG) and blue-yellow (BY) color opponencies; $f_I$ extracts the grey scale intensities, i.e. intensity is computed as the average of $r$, $g$, $b$ values in the color image for each pixel; $f_O$ obtains local orientations information by convolving the intensity image with a set of Gabor filters in four different orientations $\{0°,45°,90°,135°\}$. The complete feature vector $f_{C,I,O}$ is then calculated by:

$$f_{C,I,O} = \frac{1}{3}(f_C + f_I + f_O) \qquad (1)$$

Input to the saliency toolbox was the 640x480 color image $I$ from the webcam. The toolbox generated three 30x40 feature maps denoting color, intensities and orientations respectively (see [14] for details on how these feature maps were generated). Accordingly, as shown in Figure 4, the image is represented respectively by a 1200-component vector in each individual feature space $f_K$ $(K \in \{C, I, O\})$. The resulting feature vector $f_{C,I,O} = [x_1, x_2, x_3, ..., x_{1200}]$ is obtained by averaging them. To yield a fast and efficient mapping, a feature selection procedure is followed instead of directly using the feature vector by the learning algorithm. Finally, we use mRMR (minimum Redundancy Maximum Relevance) feature selection to reduce the high dimensional image data $I$ into a low dimensional feature vector $f = [x_1, x_2, x_3, ..., x_{50}]$ (see [10] for details on mRMR).

### Classification

We evaluated 13 different gaze locations; each location was assigned a unique label $l \in L = \{1, 2, 3, ..., 13\}$. For this first analysis, we only used the webcam images from the 17 participants. For each participant, approximately 1900 images were collected. We evaluated our system using a person-dependent evaluation scheme. For each participant, 70% of the images were randomly selected and used for training (the training set), the remaining 30% were used for testing on the same participant (the test set).

The training set comprised $N$ observations of raw eye images $I$ and labeled with the corresponding observations of gaze directions $Y$. We can then use a k-nearest neighbor

| Feature | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | P13 | P14 | P15 | P16 | P17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C,I,O | 17.3% | 11.3% | 17.1% | 13.5% | 9.1% | 17.1% | 11.3% | 10.6% | 14.9% | 16% | 10.3% | 26.1% | 21.1% | 9.5% | 12.1% | 17.3% | 17.6% |
| C | 17.0% | 11.6% | 16.0% | 14.7% | 9.5% | 15.7% | 11.3% | 10.7% | 14.6% | 17.5% | 13.8% | 26.7% | 21.8% | 10.6% | 11.3% | 20.5% | 18.6% |
| I | 17.9% | 12.2% | 16.8% | 14.3% | 10.5% | 16.2% | 10.6% | 11.4% | 13.8% | 14.7% | 9.4% | 27.2% | 19.4% | 9.1% | 11.5% | 17.6% | 17.5% |
| O | 17.2% | 11.1% | 17% | 12% | 10.2% | 15.8% | 11.1% | 13.8% | 14.1% | 15% | 9.5% | 25.9% | 19.5% | 9.5% | 10.9% | 16.4% | 15.9% |

**Table 1. Person-dependent evaluation: this table presents the error rates for 17 participants using different features (C: color, I: intensity, O: orientations). The participants consist of 5 females and 12 males, aged from 18 to 40 years (mean: 26.9, std: 6.8) old with different ethnicities and eye colors. None of the participants wore glasses but two wore contact lenses during the experiment.**

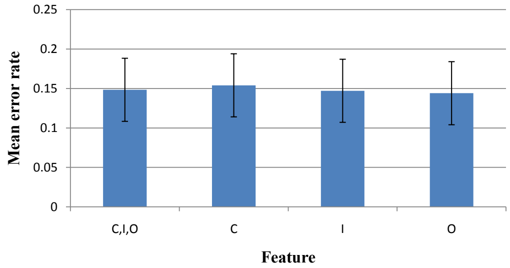
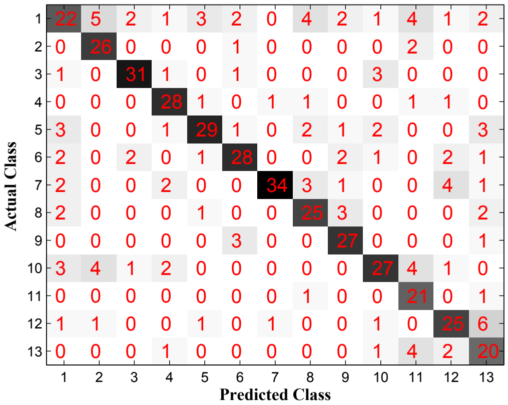


**Figure 5. Mean error rates with standard deviation for 17 participants using different features (C: color, I: intensity, O: orientations)**



|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 5 | 2 | 1 | 3 | 2 | 0 | 4 | 2 | 1 | 4 | 1 | 2 |
| 2 | 0 | 26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 3 | 1 | 0 | 31 | 1 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 28 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 5 | 3 | 0 | 0 | 1 | 29 | 1 | 0 | 2 | 1 | 2 | 0 | 0 | 3 |
| 6 | 2 | 0 | 2 | 0 | 1 | 28 | 0 | 0 | 2 | 1 | 0 | 2 | 1 |
| 7 | 2 | 0 | 0 | 2 | 0 | 0 | 34 | 3 | 1 | 0 | 0 | 4 | 1 |
| 8 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 25 | 3 | 0 | 0 | 0 | 2 |
| 9 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 27 | 0 | 0 | 0 | 1 |
| 10 | 3 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 27 | 4 | 1 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 21 | 0 | 1 |
| 12 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 25 | 6 |
| 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 2 | 20 |

**Figure 6. Confusion matrix of the gaze estimation results for participant 12 using color, intensity and orientation features combined.**

classifier (kNN) with $k = 3$ to learn the mapping $W$ from image features $f$ to gaze direction $y$.

$$W : f \to y \tag{2}$$

## RESULTS

We assessed the classification performance by using different types of image features: Color, Intensity, Orientations and three of them combined. Figure 5 shows the mean error rates with standard deviation averaged across all 17 participants. Using orientations only results in the lowest mean error rate of 14%. Table 1 presents the error rates for each participant. The error rates are between 9.1% to 21.8%, except for participant 12 who showed error rates of up to 27.2%.

We then analyzed the results for participant 12 in more detail. Figure 6 shows the confusion matrix of the gaze estimation results for participant 12 using color, intensity and orientation features combined. As can be seen from the Figure 6, the predicted class does not match the actual class in several cases, particularly for class one (C1), which corresponds to the upper left gaze location.

## Discussion

Overall, we achieved a mean recognition performance of 86% using color, intensity and orientation features. This result is very encouraging given that our data set includes blinks, changes in lighting conditions and subtle head movements. As can be seen from Table 1 and Figure 6 recognition performance for participant 12 was worse compared to the other participants. In the post-experiment analysis of the videos we found that participant 12 blinked very frequently during the data recording, which may have affected the recognition performance. In addition, the left upper gaze location was particularly bad. This suggests that the participant was slow to jump to the first starting gaze location with his eyes after each break (a black cross marker was displayed on screen center during the break) during the experiments.

The comparison of the different features revealed that no single feature performed best. Figure 5 shows that combining color, intensity and orientation information does not improve the overall recognition performance. Table 1 shows performance differs across participants. While combining all three features improves recognition performance in most cases, for some it even results in an increase of the error rate.

The recognition system makes mistakes when two classes are spatially close to each other. The system accuracy and precision could be improved either by adding more screen points in the training process or by interpolating between training points using continuous regression methods. In addition, at the moment the classifier only assigns one class label to each testing instance. Assigning a confidence as well would allow us to use probabilistic models.

In mobile settings, other challenge are changes in the geometry relationship between the actual plane of visual gaze. To address this issue, in future work we plan to include prior geometric information in the learning process. We also plan to include other features such as scale-invariant feature transform (SIFT) features, Haar-like features and different color models for gaze estimation. We also plan to improve the feature selection procedure to derive optimal feature sets for different mobile eye tracking applications.

## CONCLUSION

In this paper we proposed to use low-level features extracted from eye images for gaze estimation. This initial evaluation has shown that 13 different gaze directions can be robustly discriminated from each other. These results are encouraging and suggest that it is feasible to use a machine learning approach and low-level features for the development of more robust gaze estimation techniques. Our work has the potential for developing ideal eye tracker for applications such as mobile health monitoring, eye-based input control and remote gaze communication. For the future work, we need to consider how to obtain 3D Point-of-View when the head moves dynamically.

## REFERENCES

1. S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report cmu-cs-94-102, 1994.

2. D. Bannach, P. Lukowicz, and O. Amft. Rapid prototyping of activity recognition applications. *IEEE Pervasive Computing*, 7(2):22–31, 2008.

3. A. Bulling and H. Gellersen. Toward Mobile Eye-Based Human-Computer Interaction. *IEEE Pervasive Computing*, 9(4):8–12, 2010.

4. A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(4):741–753, Apr. 2011.

5. J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1167–1175, 2007.

6. T. Gevers. Color in image search engines. *Principles of Visual Information Retrieval*, page 35, 2001.

7. D. Hansen, J. Hansen, M. Nielsen, A. Johansen, and M. Stegmann. Eye typing using markov and active appearance models. In *Applications of Computer Vision, 2002. (WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 132–136. IEEE, 2002.

8. 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 (PAMI)*, 32(3):478–500, 2010.

9. B. Noris, K. Benmachiche, and A. Billard. Calibration-free eye gaze direction detection with gaussian processes. In *Proc. Int. Conf. on Computer Vision Theory and Applications*. Citeseer, 2008.

10. H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1226–1238, 2005.

11. R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Proceedings of the Workshop on Perceptual User Interfaces*, pages 98–100, 1997.

12. Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. *European Conference on Computer Vision (ECCV)*, pages 656–667, 2008.

13. K. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 191–195. IEEE, 2002.

14. D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

15. O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the s^ 3gp. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1. IEEE Computer Society, 2006.

16. L. Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. In *British Machine Vision Conference*, pages 428–437. Citeseer, 1998.

17. Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1. IEEE Computer Society, 2005.