Real-Time Gaze Estimation Using Monocular Vision

Zhizhi Guo[1(✉)], Qianxiang Zhou[1], Zhongqi Liu[1], Xin Zhang[2], Zhaofang Xu[1], and Yan Lv[1]

[1] School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China
lol6759797@qq.com
[2] China National Institute of Standardization, Beijing 100191, China
zhangx@cnis.gov.cn

Abstract. Improving the accuracy of gaze estimation and the tolerance of head motion is a common task in the field of gaze estimation. The core problem of gaze estimation is how to accurately build up the mapping relationship between image features and gaze position. To this end, we propose a method to reconstruct input features in the optimized subset as the key to our solution. The HOG feature is considered as the input feature. First, we found the closest calibration point to gaze position and constituted the optimized subset. Then, we get a set of weights that can linear reconstruct test samples in the optimized subset. And this set of weights is used to express the mapping relationship. At last, a linear equation is fitted to solve head motion problem. In this paper, the experiment results demonstrate that our system can achieve high accuracy gaze estimation with one camera.

Keywords: Gaze estimation · Feature reconstruction · Head move compensation · Optimized subset

1 Introduction

Eyes as one of the most important organs, is regarded as an important information input source in the human-interactive, and gaze estimation is considered as an important new type of human-interactive method. Because of its convenience and rapidity, the gaze estimation has been widely researched in recent years. And with the development of image and video processing technology, the high precision gaze estimation on monocular data has been achieved.

In general, gaze estimation method can be roughly divided into model-based method and interpolation-based method. The former method uses the eyeball geometric model, image features and hardware parameters to calculate the gaze position. Although this kind of method has been achieved in the literature [1–7], these systems tend to take at least 2 cameras and some hardware parameters. Even if the hardware cost and the complexity of calibration do not be considered, the deviation of the gaze direction calculated through the model is above 4 degrees.

M. Kurosu (Ed.): HCI 2016, Part II, LNCS 9732, pp. 61–70, 2016.
DOI: 10.1007/978-3-319-39516-6_6

Unlike the model-based method which needs accurate mathematical model as the input information, the interpolation-based methods do not require a calibrated hardware setup or extensive information about the user. This kind of method using the calibration process to construct the mapping relationship between the high-dimensional image features and the low-dimensional gazing space, and the mapping relationship is used to calculate the gaze position of test image. Sugano et al. [8] introduced a method that through Gaussian process regression establishes the mapping relationship between the eye image and the gaze point. They use visual saliency map as the input feature and achieve the accuracy of 3.5 degrees. Villanueva et al. [9] established the mapping relationship using the vector from pupil center to two corneal reflection centers, and the system accuracy reached less than 4 degrees. Cerrolaza et al. [10], Ramanauskas et al. [11] used a similar method to establish the different types of mapping relationship, which reached a similar gaze estimation accuracy.

On the other hand, the interpolation-based method also has its own disadvantages. The accuracy of gaze estimation is closely related to the number of training samples. Xu et al. [12] and Tan et al. [13] used more than 200 training samples to build the mapping relationship between input features and gaze points. Obviously, such a long time calibration process makes the user feels fatigue and disgust, so it can't be spread to commercial use or other applications.

In this paper, we propose a novel interpolation-based gaze estimation method. It utilizes the PCA + HOG feature as input feature. The core idea of the method is to found the optimized subset among all the training samples and used the ℓ^1-minimize to reconstruct test feature vector in optimized subset. The linear combination of the optimized subset is the initial gaze estimation result. Then we construct a gaze compensation equation to compensate the initial gaze estimation result, it can compensate the effect of head movement on initial gaze estimation. Eventually, the gaze estimation result gains good accuracy in the case of only using 33 calibration points.

2 Gaze Estimation Method

This paper obtains the mapping relationship between input feature and gaze position based on the input features reconstruction. The selection of input features is very important, it has decisive effect on the accuracy of features reconstruction. HOG feature has strong robustness to illumination changes and image geometric deformation. Therefore this paper uses the HOG feature as the system input feature.

2.1 Feature Extraction

When gaze position changes, the most intuitive feeling from the 2-D image is that the pupil position in the eye changed. We use the face alignment method proposed by Ren et al. [14] to find the left eye region and the right eye region as the interest area Fig. 1. In the process of HOG feature extraction, the whole eyes image is regarded as a block, and each block is divided into 3*6 cells Fig. 1. In this way, in each eye image, we get a 162-D (3*6*9) feature vector.

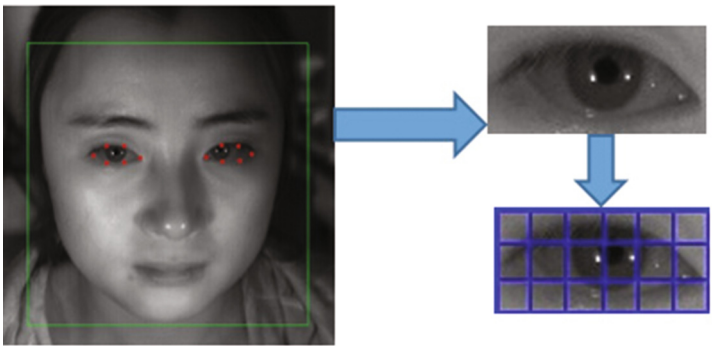

Fig. 1. Feature vector extraction.

Such a big feature vector not only affects the speed of feature reconstruction but also contains a lot of useless feature dimension which will become noise in the process of reconstruction. The main factor that reflects the essence of image changes can be analyzed from high dimensional feature vectors by PCA. It can be seen from Fig. 2, the former 10-D features contains 90 % information of the feature space. Therefore, we use PCA to reduce feature vector dimensions from 162 to 10, and make it to be the system input feature.

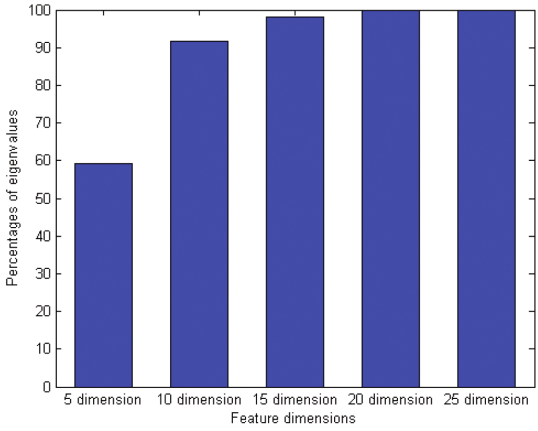

Fig. 2. The percentages of eigenvalues in different feature dimensions

The proposed gaze tracking system is a real-time system, the sizes of two consecutive frames of eye image do not appear a big change. To each test frame, we use the former frame to judge if the interest area is right or not. If the height or weight of interest area has a large change, we skip the frame to test next frame.

2.2 Initial Gaze Estimation

The feature vector reflects the changes of 2-D eye image when gazing at different positions. Assuming a training set of eye images consist of all the eye features matrices $E = [e_1, e_2, \ldots, e_n] \in \mathbb{R}^{m*n}$ and corresponding gaze position matrices $P = [p_1, p_2, \ldots, p_n] \in \mathbb{R}^{1*n}$. We hope to find a mapping from E to P:

$$P = AE \tag{1}$$

where $A \in \mathbb{R}^{1*m}$ is the projection matrix. Obviously, if n > m, the system of equation is overdetermined. We cannot find a mapping matrix that is accurate for all the training samples. However, if n < m, the certain A can be found. So we need to choose the optimized subset $E' = [e'_1, e'_2, \ldots, e'_{n'}]$ and $P' = [p'_1, p'_2, \ldots, p'_{n'}]$ in all training sets. The new mapping can be found:

$$\mathrm{P}' = \mathrm{A}'\mathrm{E}' \tag{2}$$

where A' is the new projection matrix. We hope any test sample ($\hat{e}$, $\hat{p}$) can find the corresponding A'. So in the process of the mapping relationship constructing, only a few e'_i weights is allowed to be different than zero. We can make the problem of solving the mapping matrix A transformed into selecting the fewest e_i to construct optimized subset in all image training samples. The optimized subset E' should be closely to the test image feature $\hat{e}$, so that they can have a same mapping relationship, and a existed set of reconstruction weight $\{w_i\}$ can linear reconstruct $\hat{e}$.

$$\hat{e} = \sum\nolimits_i w_i e'_i \tag{3}$$

Ideally optimized subset would contain the samples which are close to the test image in the gaze space. A training feature vector which has the minimum Euclidean distance to test feature vector can be found by Eq. 4. The calibration point corresponding to this vector is marked for the main point, which is the closest to the real gaze position.

$$\min_i d_i = argmin \ \ || \hat{e} - c * e_i ||_1 \tag{4}$$

where $\hat{e}$ is the feature vector of test image, e_i is the ith feature vector of training samples, c is a coefficient, d_i is the minimum Euclidean distance. The feature vectors, corresponding to the main point and six other calibration points around it, constitute the optimized subset $\hat{E} = [e_{main}, e^1_{main}, \ldots, e^6_{main}]$. If the main point is on the edge, it is used to constitute the optimized subset with the existed calibration points around it.

Reconstructing the weight w in optimized subset is formulated as a sparse reconstruction problem, which can be solved by minimizing the ℓ^1 norm of w [15, 16]. Due to the existence of real noise, it may not be possible to represent the test sample exactly as a linear combination of the optimized subset. A small constant ε was introduced to

express the maximum allowed Euclidean distance from $\hat{E}w$ to the ground true $\hat{e}$. The reconstruction weight w can be get by:

$$\hat{w} = \text{argmin} \parallel \text{w} \parallel_1 \quad s.t. \quad \parallel \hat{E}\text{w} - \hat{e} \parallel_1 < \varepsilon \tag{5}$$

Lu et al. [17] has demonstrated that use of the same weights to estimate the gaze parameters is justified by locality, as the linear combinations in the subspaces spanned by $\{e_j^{'}\}$ and $\{p_j^{'}\}$ are equal. Finally, the test gaze position $\hat{p}$ can be calculated by:

$$\hat{p} = \sum\nolimits_i w_i p_i \tag{6}$$

2.3 Gaze Position Compensation

The vertical height changes of the ground truth gaze point on the influence of the initial gaze results are shown in Fig. 3. In Fig. 3, the vertical direction of gaze estimation result has the trend to close the center of the screen. The vertical error of test points in the screen edge are bigger than in the screen central (point height between 300 and 800). In addition, the error of gaze estimation results in vertical direction is larger than in horizontal direction (Sect. 4). The main reasons for the above phenomenon are: 1. Due to the people vision area on horizontal is wider than it on vertical, people move head largely when changing the gazing point on vertical. 2. When the gaze point changes in the horizontal direction, the eye image has significant changes; when the gaze point changes in the vertical direction, the eye image has little changes.

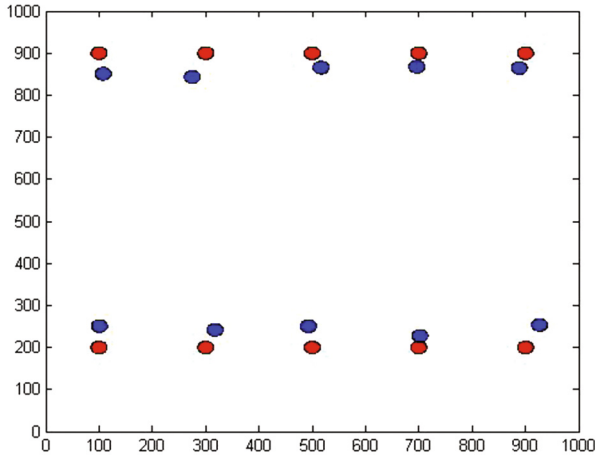

Fig. 3. Initial gaze estimation result under different gaze positon.

The curves of mean errors and eye open sizes have almost the same linear relation. Therefore, after the initial gaze position is got, a linear equation is used to compensate the vertical deviation. The final gaze estimation result in vertical direction is calculated by:

$$p_{yf} = \hat{p}_y + (S_h - S_l)/(H_h - H_l) * (H/2 - \hat{p}_y) \tag{7}$$

where $\hat{p}_y$ is the estimated value of initial gaze estimation in vertical direction, H is the total pixels in the screen vertical direction, S_h is the eye open size when gazing the highest training point on the screen, S_l is the eye open size when gazing the lowest training point on the screen, H_l is the vertical coordinate of the lowest training point on the screen, H_h is the vertical coordinate of the highest training point on the screen.

3 Experiments

In this section, the experiments performed to evaluate the proposed gaze estimation system. 6 male and 4 female subjects are chosen to do the experiment under a condition of one camera with a resolution of 1280*720 and 2 infrared light sources with 850 nm wavelength. We implemented our system with a 24-inch computer screen, and the resolution is 1920*1080 pixels. The subject was asked to sit at a distance of 600 mm from the computer screen. In the experimental process, subject's head tries to aim at the screen center as much as possible, and the subject is allowed to have slight head motion.

The whole experiment process is divided into calibration stage and test stage. In the process of calibration, the subject focused on each calibration point shown on the screen and allowed the camera to capture frontal appearance. In the process of test, the subject watched the test points shown on screen. There are 30 test points distributed in each position of the screen, and were shown in random order.

3.1 Evaluation and Comparison

For each input image, the gaze positions of the left eye and the right eye are calculated respectively, the average value of the two gaze positions is regard as the double-eye gazing position. In order to show the experimental result directly and compare with other state-of-the-art methods, the angular of estimation error will be calculated by:

$$\text{error} \approx \arctan(||\hat{p} - p_0||_2/\text{D}) \tag{8}$$

where $||\hat{\text{p}} - \text{p}_0||_2$ denotes the Euclidean distance between a real 2D gaze position p_0 and the estimated 2D gaze position $\hat{\text{p}}$, D indicates the distance between the subject's eye and screen.

Table 1 shows the mean errors of the gaze estimation system for each subject. It shows the highest estimation accuracy in all subjects. In general, the gaze estimator of one eye achieves a mean error of 83 pixels, corresponding to an angle error of 2.15°; while the gaze estimator of double eyes has a mean error of 69 pixels, corresponding to an angle error of 1.79°. In some subjects, the left eye's gaze estimation accuracy is

higher than the right eye's, while other subjects are opposite. The left and right eye's overall average gaze estimation precisions are basically the same. It demonstrates that which eye's gaze estimation result is more accurate depends on the subjects' own individual differences, and every double eye gazing estimation accuracy is higher than one eye gazing estimation accuracy.

Table 1. Mean pixel error and mean angel error

Subject	Left eye		Right eye		Double eye	
	Pixels	Angle	Pixels	Angle	Pixels	Angle
1	75	1.94°	86	2.23°	68	1.76°
2	85	2.20°	98	2.54°	76	1.97°
3	98	2.54°	86	2.23°	83	2.15°
4	64	1.66°	80	2.07°	52	1.35°
5	73	1.89°	87	2.25°	65	1.68°
6	77	1.99°	73	1.89°	61	1.58°
7	79	2.05°	94	2.44°	73	1.89°
8	92	2.38°	84	2.18°	76	1.97°
9	81	2.10°	88	2.28°	70	1.81°
10	80	2.07°	85	2.20°	69	1.79°
All avg	**80**	**2.07°**	**86**	**2.23°**	**69**	**1.79°**

In addition, we compare our system with other gaze estimation systems which without head fixed device. Comparison results are shown in Table 2, compared with other excellent methods in recent years, our method has better positioning accuracy no matter in one eye or double eye gazing estimation.

Table 2. Comparison with the state-of-the-art method

Method	Error
Our Method(Double eye)	**1.79°**
Our Method(One eye)	**2.15°**
Feng et al. [17]	2.3°
Valenti et al. [18]	(1.9°, 2.2°)
Sugano et al. [8]	3.5°

3.2 Compensation Equation Evaluation

To compensate the error caused by slight head motion during gaze estimation, a gaze compensation algorithm is proposed in Sect. 2.3. This section is mainly to evaluate the effect of the compensation algorithm for the final result of gaze estimation.

It can be seen from the Fig. 4, the mean pixel error of initial double-eye gazing estimation on the y direction (mean error 61) is obviously larger than it on the x direction (mean error 37), and the reason has been explained in Sect. 2.3. After the use of gaze compensation method, the mean pixel error on y direction is descended from 61 to 45, decreased by 35 %.

3.3 Distance Change

In the above experiment, the distance between subjects and screen is set to be 600 mm. This section evaluates the robustness of the proposed method in test with distance changes between subjects and screen. We choose three subjects to do the experiment, the test distances were respectively set as 500 mm, 600 mm, 700 mm and 800 mm. The experiment processes are exactly as described in Sect. 3.1.

It can be seen from the experimental results in Table 3, the overall gaze accuracy under different distances are basically equal. This proves the gaze estimation method mentioned in this paper has a good robustness to distance changes.

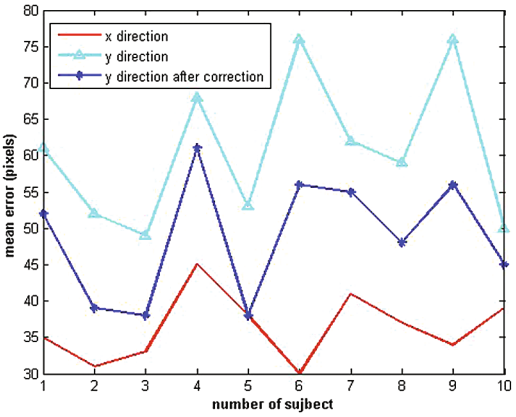

Fig. 4. Mean error at x direction and y direction

Table 3. The mean error of gaze estimation under different distance

Subject	500 mm		600 mm		700 mm		800 mm	
	Pixel	Angle	Pixel	Angle	Pixel	Angle	Pixel	Angle
1	61	1.90°	76	1.97°	84	1.86°	100	1.26°
2	60	1.87°	70	1.81°	77	1.71°	94	1.14°
3	58	1.81°	69	1.79°	79	1.75°	89	1.05°
avg	**59**	**1.84°**	**72**	**1.86°**	**80**	**1.78°**	**94**	**1.14°**

4 Conclusion and Future Work

In this paper, we have proposed an accurate gaze estimation method under a little calibration points. First, the main point is found by using the minimum Euclidean distance among the all calibration feature vectors. The main point and the calibration points around it constitute the optimized subset. In the optimized subset, a set of sparse reconstruction weights is solved by using the ℓ^1-minimum and utilized to linear express the initial gaze estimation result. Based on the result of initial gaze estimation, we use

gazing compensation equation to get the final gaze estimation result. Experiment shows that our method can achieve an accuracy of 1.79° under monocular vision.

However, limitations still exist. First, although our method can achieve high precision gaze estimation with slight head motion, it is still powerless with free head motion. Second, the calibration process is necessary to each subject. The mapping relationship between input feature and gaze position cannot be found without calibration. Overall, handling large head motion and completely removing training stage are currently impossible in this work. For further research, the method could be improved on basis of this paper, 3D head pose estimation can be added, in the meanwhile, reconstructing the head motion compensation equation, and then the free head motion gaze estimator would be achieved.

Acknowledgement. This research was funded by National science and technology support plan "User evaluation technology and standard research of display and control interface ergonomics"(2014BAK01B04).

References

1. Valenti, R., Gevers, T.: Accurate eye center location through invariant isocentric patterns. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1785–1798 (2012)
2. Markuš, N., Frljak, M., Pandžić, I.S., et al.: Eye pupil localization with an ensemble of randomized trees. Pattern Recogn. **47**(2), 578–587 (2014)
3. Morimoto, C.H., Mimica, M.R.M.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (2005)
4. Zhu, Z., Ji, Q.: Novel eye gaze tracking techniques under natural head movement. IEEE Trans. Biomed. Eng. **54**(12), 2246–2260 (2007)
5. Villanueva, A., Cabeza, R.: A novel gaze estimation system with one calibration point. IEEE Trans. Syst. Man Cybern. Part B Cybern. **38**(4), 1123–1138 (2008)
6. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. In: VISAPP, pp. 125–130 (2011)
7. Coutinho, F.L., Morimoto, C.H.: Improving head movement tolerance of cross-ratio based eye trackers. Int. J. Comput. Vis. **101**(3), 1–23 (2012)
8. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2012)
9. Sesma-Sanchez, L., Villanueva, A., et al.: Gaze estimation interpolation methods based on binocular data. IEEE Trans. Biomed. Eng. **59**(8), 2235–2243 (2012)
10. Cerrolaza, J., Villanueva, A., Cabeza, R.: Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems. In: Proceedings of the Symposium on Eye Tracking Research & Applications, pp. 259–266. ACM (2008)
11. Ramanauskas, N., Daunys, G., Dervinis, D.: Investigation of calibration techniques in video based eye tracking system. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 1208–1215. Springer, Heidelberg (2008)
12. Xu, L.Q., Machin, D., Sheppard, P.: A novel approach to realtime non-intrusive gaze finding. In: BMVC, pp. 428–437 (1998)
13. Tan, K., Kriegman, D., Ahuja, N.: Appearance-based eye gaze estimation. In: WACV, pp. 191–195 (2002)

14. Ren, S., Cao, X., Wei, Y., et al.: Face alignment at 3000 FPS via regressing local binary features. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1685–1692. IEEE (2014)
15. Wright, J., Ganesh, A., Zhou, Z., et al.: Demo: Robust face recognition via sparse representation. In: IEEE International Conference on Automatic Face and Gesture Recognition, Fg 08, pp. 1–2 (2008)
16. Donoho, D.L., Tsaig, Y.: Fast solution of l1-norm minimization problems when the solution may be sparse. IEEE Trans. Inf. Theory **54**(11), 4789–4812 (2008)
17. Lu, F., Sugano, Y., Okabe, T., et al.: Adaptive linear regression for appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 2033–2046 (2014)
18. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**(2), 802–815 (2012)