Research on Gaze Estimation Method Combined with Head Motion Changes

Xiangyi Zhan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zxyy129@163.com

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

***Abstract*—The line of sight reflects the focus of human attention. Gaze estimation technology has a wide range of application prospects in human-computer interaction, human emotion analysis, commercial advertising, and so on. Gaze estimation needs to be jointly determined by eye movement and head movement because the human gaze is often with the head movement. In this paper, a gaze estimation system with head movement is implemented by a monocular camera, using eye movement features and head movement posture changes to estimate the gaze point. A camera is used as the information acquisition device to extract the eye movement feature information, and the offset of head movement is estimated at the same time. The final fixation point position coordinates are obtained by compensating the fixation point position in the case of head movement, and then the fixation point position coordinates are estimated. For the compensated gaze drop point, the average pixel error on the X-axis is estimated to be 118 pixels, and the average pixel error on the Y-axis is estimated to be 136 pixels.**

Keywords-Gaze Estimation; Head Movement; Pupil Center Detection; Head Eye Movement

I. INTRODUCTION

With the continuous improvement of the level of science and technology, the way of human-computer interaction (HCI) is constantly changing. The eyes are the most important organ for the human body to obtain external information, taking the gaze position from the eyes as a way of human-computer interaction, which has the characteristics of more direct, natural, and human nature. Eye tracking technology captures the eye position information and posture of the camera to obtain the gaze direction and the position of the observation point, track the eye line of the human eye, and make the human-machine interaction through eye movement naturally and conveniently. Therefore, studying gaze estimation under head motion is of great significance.

A. Gaze Estimation

The current gaze estimation methods are mainly divided into two categories: feature-based gaze estimation and epigenetic gaze estimation. Feature-based sight estimation methods can be divided into model-based and regression-based sight estimation methods according to different implementation schemes of eye direction and eye feature mapping relationship [1].

The model-based gaze estimation method first takes the straight line obtained by connecting the pupil or iris center as the optical axis. At the same time, a fixed deviation Angle between the optical axis and the visual axis is obtained based on prior knowledge to compensate for the actual line of sight to improve the accuracy of the gaze estimation. Chen proposed a 3D gaze estimation algorithm based on single-camera face tracking [2]. The algorithm was implemented without the help of an external light source. The solving equations of the 3D eye center, pupil center, and visual axis were derived, and the fixation point was solved by one-time calibration. Regression-based gaze estimation methods usually includes the following two processes: selecting and extracting eye motion and constructing regression function to obtain the mapping relationship between the line of sight and eye movement features. Eye movement is usually obtained by connecting a fixed point unrelated to eye movement with a moving point strongly related to eye movement. The most commonly used eye movement is calculated by the corner

point and the iris center point. Yamazoe proposed a real-time gaze estimation method using specific face feature association. This method calculates the geometric relationship between the eye center and the eye radius in advance and simplifies the sight estimation method based on a 3D sight model by tracking specific facial features [3]. Valenti used the method of Isocenters for iris center location and canthus location under a single webcam and achieved a better detection accuracy in low-resolution images [4].

Appearance-based gaze estimation method extracts features from the whole image and then obtains the mapping relationship model between features and line of sight direction through training. In the early stage, scholars used manual features such as HOG or LBP to extract features and then used K-nearest neighbor [5], random regression forest [6] and support vector machine (SVM) [7], and other models to achieve line-of-sight estimation. Subsequently, with the development of neural network technology, neural network models based on deep learning have been widely applied to gaze estimation. To improve the generalization performance of appearance-based models, Wang proposed a method combining adversarial learning with Bayesian inference for the overfitting of appearance, head pose, and point estimation [8]. To solve the interference of face images with free head movement on the line of sight model mapping, Zhou proposed to input different areas of the face into the 3D line of sight estimator with adaptive weighting during head movement, which greatly improved the efficiency and accuracy of the regression model [9].

B. Head Pose Estimation

Head pose estimation is intrinsically related to visual gaze estimation, namely, the head pose can represent the direction and focus of the human eyes. Physiological studies have shown that a person's gaze prediction comes from a combination of head posture and eye orientation. There are mainly the following methods for obtaining head pose: the method based on standard template matching, the method based on the geometric relationship of key points of the face, the method based on feature regression, and the method based on manifold learning [10].

The method based on standard template matching compares the header image with the model with pose labels in the feature space [11] and calculates the similarity between the input features and the standard template set [12].

The method based on the geometric relationship of the key points of the face determines the corresponding pose according to the position of the eyes, mouth, and nose tip, which can recover the global pose change of the head from the video. Fridman used the random forest machine learning method to classify the pose of the driver's facial 56-point position relationship features [13]. The geometric method is simple, fast, and has low time complexity. However, this method has high requirements for the detection and location of facial feature points.

The multi-layer perceptual neural network is a common regression tool in the method based on feature regression. This method corresponds the output layer of the network to the discrete head posture, and the head posture estimation results are obtained by training the neural network directly. Murphy train regressors by support vector regression (SVR) [14]. Huang introduced the random forest algorithm into head pose estimation [15]. The method based on regression has good real-time performance and high accuracy, but it requires a large amount of computation and is greatly affected by the head detection and positioning results.

Based on manifold learning, Huang improved the accuracy of head pose in non-uniform samples by supervising local subspace learning; and discussed the phenomenon of over fitting [16]. Foytik roughly estimated the head pose through the linear supervision function, and then realized the accurate judgment of the head pose through the linear regression function [17]. The method based on manifold embedding has high time complexity and low accuracy, which is still far from the real practical process.

II. METHODS

The experimental environment of this study is simple, and image data of subjects are collected through a monocular camera and a single screen. The computer display is 31.5 inches with a

resolution of 2560×1440 pixels, and the camera is fixed in the middle of the top of the screen with a camera of 2048 × 2048 pixels. A total of 12 laboratory members, including 8 men and 4 women, participated in the data collection. Aged between 23 and 30 years with good health and good vision. Before the experiment, each subject was familiar with the specific contents and precautions of the experiment, and they were willing to participate in the experiment. In order to eliminate external interference, each experiment was completed by only one subject alone after making experimental preparations. The experiment was carried out in natural light.

The data collection in this study was divided into two parts. The first part was a static experiment, in which subjects were required to move their heads as close to the starting position as possible. The second part is the dynamic experiment, in which subjects' heads can move freely.

A. Static Experiment

Before starting the experiment, the camera and other equipment should be calibrated to ensure that they are in normal operation. Subjects e adjust their seated position on the chair to be approximately 60cm away from the center screen, ensuring that they are within the best focal length of the camera and that the origin of the subject's head is as close to the origin of the camera coordinate system as possible.

At the beginning of the experiment, a red dot with a radius of 25 pixels was displayed on the screen. The red dot represented the object to be captured by the eye. Each time the red circle appeared, the subject was asked to look at the center of the red circle and press the space bar at the same time. At this point, the camera would take an image of the subject's face, and the software would record the coordinates of the target point, which would then show the next origin. This is shown in Figure 1. Through the experiment, a set of data, including a face image and screen coordinates, are obtained.

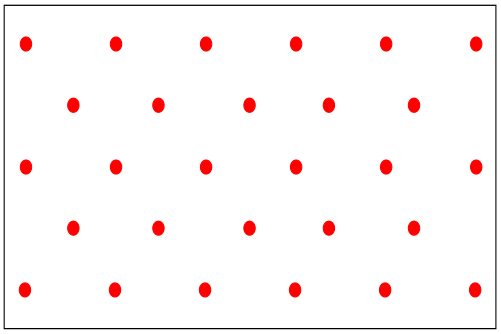

Figure 1. Schematic diagram of gaze target in static experiment

When the head is still, there is only eye movement in the process of gazing at the target, and the images captured are all frontal face images, which are almost the same, except for the difference in eye direction, as shown in Figure 2.

Figure 2. Partial data of static experiment

B. Dynamic Experiment

Experiments in the dynamic line-of-sight estimation stage are called dynamic experiments, and the experimental environment is the same as that of static experiments. Before starting, subjects should adjust their sitting position so that the distance between subjects and the center screen is about 60cm. At the beginning of the experiment, the subjects' eyes were required to face the center of the middle screen, and the initial state of head posture was calibrated by pressing the space bar. Dynamic experiments require subjects to move their heads in a specific way by focusing their eyes on a static point in the center of a computer screen, and moving their heads while still focusing on that point.

In this experiment, the subject can rotate the head freely, rotate around the X, Y, and Z axes, and generate the pitch Angle, yaw Angle, and roll Angle as well as the translation along the X, Y, and Z axes. A 5-second video clip was collected for this experiment. During head movement, the coordinated movement of the eye and the head occurs during the fixation of the target, and the pictures obtained by shooting show various appearances, resulting in the attitude change of the eye region, as shown in Figure 3.

Figure 3. Partial data of dynamic experiment

III. METHODS

In this study, we used a feature-based gaze estimation method to estimate the subject's fixation position by shooting video images with a non-wearable monocular camera. For the estimation of head motion and pose, we establish a rigid body model of the head and estimate it by geometric calculation. In this study, we use image processing and feature extraction methods to fuse head features and eye features to realize the research of gaze estimation. The technical route for the construction phase of this article is shown in Figure 4.

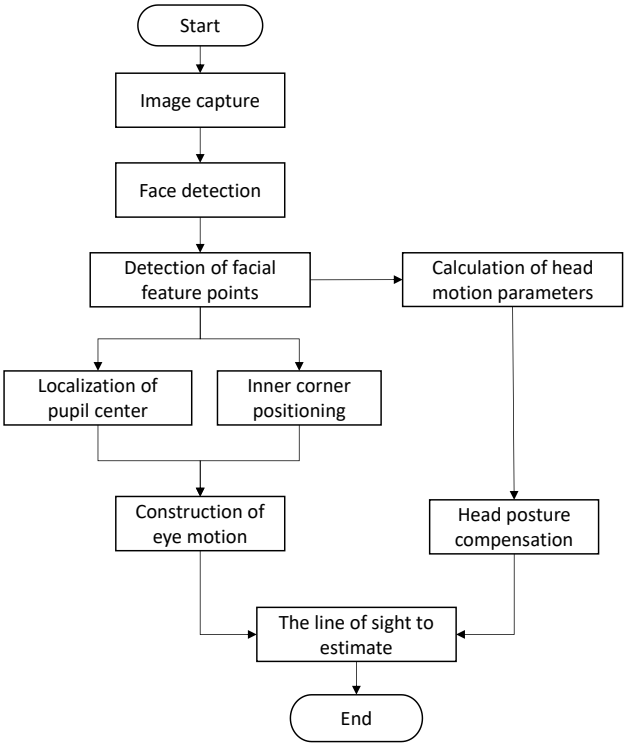

Figure 4. Technology route

A. Eye Feature Extraction

To obtain the eye image, the location of the iris center of the human eye and the inner corner of the human eye are extracted to construct eye motivities. We need to extract the position region of the face and the position coordinates of the key points of the face from the image.

The face detection method based on Adaboost and wavelet features are used to obtain the location region information of the face. Based on Haar-like features, the algorithm added a cascade classifier and integral graph algorithm to improve the speed of image processing, meet the real-time operation requirements of the whole fixation point estimation algorithm, and also to ensure the accuracy of face detection.The theory of integral graph is introduced into the algorithm, which can reduce the problem of excessive computation due to the excessive number of Haar features in the search window. The result of face recognition is shown in Figure 5.

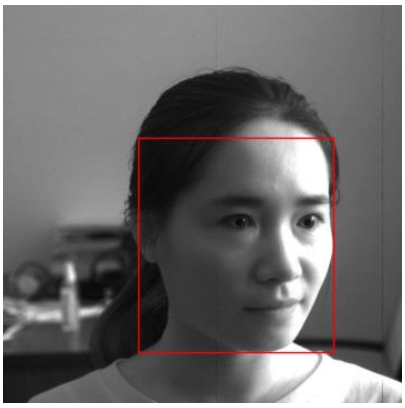

Figure 5. Face detection

Next, face feature point detection, face feature point detection is also known as face alignment, through the algorithm to detect the inner and outer corner of the face, nose tip, corner of the mouth, and another key points of the face position coordinates. Under normal circumstances, we get the position coordinates of 68 feature points of the face. After using the adaboost+haar-like algorithm to obtain the region location of the face, this paper uses the dlib c++ library to detect the feature points of the face. The position coordinates of 68 feature points of the face were obtained. The main feature points of the face included the feature points around the eye area, the coordinate points of the inner corner of the eye needed to constitute the eye movement, and the important feature points of the face needed to calculate the rotation matrix and displacement matrix of the head. The detection results of face feature points are shown in Figure 6.

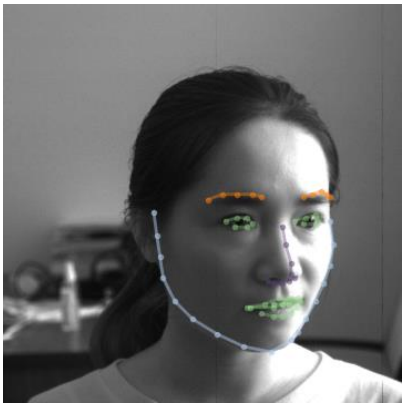

Figure 6. Facial feature points

Based on face feature points, this paper uses gradient analysis combined with star rays to calculate the position coordinates of iris center points. The algorithm can be run in the case of low-quality images, which not only ensures the computational speed but also improves the robustness of the algorithm, and can meet the different head movements and the changes of light and shade.

After obtaining the eye movement information, the mapping relationship between the feature information data of eye movement and the position coordinates of points on the screen should be established. The mapping relationship between the feature data of eye movement data and the coordinate points of the screen is a kind of nonlinear mapping relationship. Due to the real-time requirement of the gaze point estimation system, it is necessary to select a mapping model that can process the high-dimensional data and has a fast training speed. After a comprehensive comparison of various mapping models, random forest is chosen as the mapping model.

B. Head Movement Posture

Through the motion analysis of the head sequence image, the global motion of the whole head is considered as a rigid body motion model, and the head motion parameters are estimated. An object whose 3-D distance between any pair of points on the object does not change with time is defined as a rigid body. The human head can be simplified into a rigid body in the physical sense, that is, the distance between the surface points of the head's rigid body remains unchanged during its movement. Therefore, motion parameters can be estimated based on the property that the distance between points remains unchanged before and after rigid body motion.

Compared with the classical head pose estimation methods, EPNP and POSIT algorithm, it is found that the head pose estimation algorithm based on PNP has smaller advantages in estimation accuracy than the POSIT algorithm, while the head poses estimation algorithm based on POSIT has obvious advantages in calculation speed. Therefore, In this paper, the POSIT algorithm is chosen to solve the head posture, which not only ensures the requirements of solving accuracy but also meets the requirements of computing speed.

The position coordinates of the obtained facial key points and the corresponding key points of the general 3D face model were mapped and projected, and the rotation and displacement matrix (RT) of the final head pose estimation and the Yaw, Pitch, and Roll angles of the final head were calculated.

68 face feature points were obtained, from which several key points with different depths of high reliability and robustness were selected to ensure the accuracy and robustness of head pose estimation. Therefore, 8 key points were selected. That is, the left corner of the left eye, the right corner of the left eye, the right corner of the right eye, the left corner of the right eye, the nose tip, the left corner of the mouth point, the right corner of the mouth point and the lower frontal corner are used to estimate the head pose. The estimation result of the head pose is shown in Figure 7.

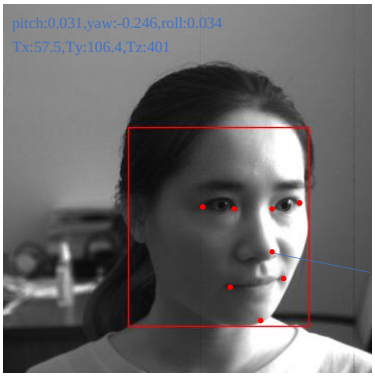

Figure 7. Estimation of head pose

C. Gaze Estimation

Is by the human eye eye eye gaze position information and joint decision movement of the head, in the head and the screen under the condition of relatively static, estimates the gaze of the initial state, to achieve the head moving cases of the fixation point estimate, need to deal with the movement of the head, above calculated the rotation matrix and translation matrix of the head, The initial fixation position was compensated according to the deflection of the tester's head, and the actual fixation position was finally obtained.

We assume that the influence of head movement on the position coordinates of fixation point is $(\Delta u_x, \Delta u_y)$.In the initial case, the head remains stationary between the screen, and the 3D initial coordinate of a point of the head is set as (x_0, y_0, z_0), the pixel coordinates of the projection of this point on the two-dimensional plane is (u_0, v_0), when the head movement occurs, the coordinates of this point are (x_1, y_1, z_1), We obtained the rotation matrix R and translation matrix T in the case of head movement through the improved POSIT algorithm, and the formula can be obtained as follows:

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = R \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + T \tag{1}$$

Where the R rotation matrix and T translation matrix are shown below.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{2}$$

$$T = \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{3}$$

Therefore, the offset caused by the head movement can be calculated by the formula $(\Delta u_x, \Delta u_y)$, and f is the focal length of the camera.

$$\Delta u_x = f\frac{x_1}{z_1} - u_0 \tag{4}$$

$$\Delta u_y = f\frac{y_1}{z_1} - v_0 \tag{5}$$

The position coordinate of the gaze point under the stationary state of the head is (u_x, u_y), and the position coordinate of the gaze changed by the head movement is (s_x, s_y):

$$s_x = u_x + \Delta u_x \tag{6}$$

$$s_y = u_y + \Delta u_y \tag{7}$$

IV. Results and Analysis

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready

to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

Finally, the eye line estimation system of head movement was tested, and 12 subjects who participated in the experimental data collection participated in the test experiment. Each test participant needs to independently complete data collection, calibration training, and result testing. During the test, the test subjects were required to gaze at the 12 test points displayed on the screen in turn. To avoid interference caused by multiple test points at the same time, only one fixation point was displayed on the screen at the same time, and the calculated position coordinates of the fixation points were represented by red crosses. In the case of head movement, the data obtained from the test are sorted out, and the form of visual display of the experimental results is presented, as shown in Figure 8.

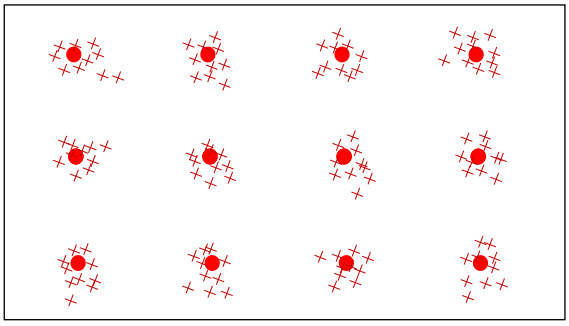

Figure 8. Schematic diagram of the test effect

To investigate the influence of the compensation effect of head motion and posture on the experimental results, the same group of testers was asked to carry out the test experiment with the same test equipment and test points under the same lighting conditions when the head was still and the head was moving. The test results are shown in the following table:

TABLE I. COMPARISON TABLE OF EXPERIMENTAL RESULTS UNDER DIFFERENT CONDITIONS

State of the head	average pixel error		average Angle error	
	X-axis (pixel)	*Y-axis (pixel)*	*X-axis (degree)*	*Y-axis (degree)*
The head rest	83	125	1.74	2.32
The head movement (No compensation)	265	380	7.23	8.15
The head movement (A compensation)	118	136	2.88	3.23

V. CONCLUSION

Gaze estimation is of great significance in the field of human-computer interaction, which is widely used in psychology, graphics, medicine, advertising psychology, military science, and other fields. At present, model-based line-of-sight estimation methods are complicated, and appearance-based line-of-sight estimation methods need a large amount of training data. In this paper, a simple gaze estimation method is adopted. Firstly, the face image is taken as the input, and the gaze estimation is performed in the initial state by extracting the eye movement information. Then, by extracting face feature points, the head rigid body motion model was established, and the head motion posture was calculated to estimate the head posture. Finally, the head movement posture was used as compensation for the eye movement gaze estimation, and the final free head movement gaze estimation was obtained.The experimental results show that the rigid motion model of the head can be used as compensation for the gaze estimation, and the head can be accurately estimated under the condition of free motion. For the compensated gaze drop point, the average pixel error of the estimated axis is 118 pixels, and the average pixel error of the Y-axis is 136 pixels.

ACKNOWLEDGMENT

This work is supported by The National Natural Science Foundation of China (No. 52072293).

REFERENCES

[1] Wood E, Bulling A. Eyetab: Model-based gaze estimation on unmodified tablet computers[C]//Proceedings of the Symposium on Eye Tracking Research and Applications. 2014: 207-210.

[2] Chen J, Ji Q. 3D gaze estimation with a single camera without IR illumination[C]//2008 19th International Conference on Pattern Recognition. IEEE, 2008: 1-4.

[3] Yamazoe H, Utsumi A, Yonezawa T, et al. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions[C]//Proceedings of the 2008 symposium on Eye tracking research & applications. 2008: 245-250.

[4] Valenti R, Staiano J, Sebe N, et al. Webcam-based visual gaze estimation[C]//International Conference on Image Analysis and Processing. Springer, Berlin, Heidelberg, 2009: 662-671.

[5] Zhang Y, Bulling A, Gellersen H. Discrimination of gaze directions using low-level eye image features[C]//Proceedings of the 1st international

workshop on pervasive eye tracking & mobile eye-based interaction. 2011: 9-14.

[6] Wang Y, Shen T, Yuan G, et al. Appearance-based gaze estimation using deep features and random forest regression[J]. Knowledge-Based Systems, 2016, 110: 293-301.

[7] Huang Q, Veeraraghavan A, Sabharwal A. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets[J]. Machine Vision and Applications, 2017, 28(5): 445-461.

[8] Wang K, Zhao R, Su H, et al. Generalizing eye tracking with bayesian adversarial learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11907-11916.

[9] Zhou X, Jiang J, Liu Q, et al. Learning a 3D gaze estimator with adaptive weighted strategy[J]. IEEE Access, 2020, 8: 82142-82152.

[10] Murphy-Chutorian E , Trivedi M M . Head Pose Estimation in Computer Vision: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(4):607-626.

[11] HE S, LIANG A, LIN L, et al. A Continuously Adaptive Template Matching Algorithm for Human Tracking[C]//In 2017 First IEEE International Conference on Robotic Computing (IRC). Taichung: IEEE, 2017:303-309.

[12] HASSNER T, MASI I, KIM J, et al. Pooling Faces: Template Based Face Recognition with Pooled Face Images[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Las Vegas: IEEE, 2016:59-67.

[13] FRIDMAN L, LANGHANS P, LEE J, et al. Driver Gaze Estimation Without Using Eye Movement[J]. IEEE Intelligent Systems, 2015, 31(3):49-56.

[14] MURPHY C, ERIK T, MOHAN M. Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(2):300-311.

[15] HUANG C, DING X, FANG C. Head Pose Estimation Based on Random Forests for Multiclass Classification[C]// International Conference on Pattern Recognition (ICPR). Istanbul: IEEE, 2010:934-937.

[16] HUANG D, STOREY M, TORRE F D L, et al. Supervised Local Subspace Learning for Continuous Head Pose Estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs: IEEE, 2011:2921-2928.

[17] FOYTIK J, ASARI V K. A Two-Layer Framework for Piecewise Linear Manifold- Based Head Pose Estimation[J]. International Journal of Computer Vision, 2013, 101(2):270-287.