HHS Public Access

Author manuscript

*Pattern Anal Appl*. Author manuscript; available in PMC 2018 January 02.

Published in final edited form as:

*Pattern Anal Appl*. 2016 August ; 19(3): 611–620. doi:10.1007/s10044-014-0407-5.

# An auxiliary gaze point estimation method based on facial normal

**Wei Sun**,
School of Aerospace Science and Technology, Xidian University, Xi'an, Shaanxi, China

**Nan Sun**,
School of Aerospace Science and Technology, Xidian University, Xi'an, Shaanxi, China

**Baolong Guo**,
School of Aerospace Science and Technology, Xidian University, Xi'an, Shaanxi, China

**Wenyan Jia**, and
Departments of Neurosurgery, Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA

**Mingui Sun**
Departments of Neurosurgery, Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA

## Abstract

Considering the main disadvantage of the existing gaze point estimation methods which restrict user's head movement and have potential injury on eyes, we propose a gaze point estimation method based on facial normal and binocular vision. Firstly, we calibrate stereo cameras to determine the extrinsic and intrinsic parameters of the cameras; Secondly, face is quickly detected by Viola–Jones framework and the center position of the two irises can be located based on integro-differential operators; The two nostrils and mouth are detected based on the saturation difference and their 2D coordinates can be calculated; Thirdly, the 3D coordinates of these five points are obtained by stereo matching and 3D reconstruction; After that, a plane fitting algorithm based on least squares is adopted to get the approximate facial plane, then, the normal via the midpoint of the two pupils can be figured out; Finally, the point-of-gaze can be obtained by getting the intersection point of the facial normal and the computer screen. Experimental results confirm the accuracy and robustness of the proposed method.

### Keywords

Binocular vision; Viola-Jones framework; Integro-differential operators; Plane fitting; Facial normal

Correspondence to: Wei Sun.

## 1 Introduction

The point-of-gaze is defined as the intersection of the gaze direction with the surface of the object being viewed (such as the screen of a computer). Gaze tracking device is a video system which estimates the gaze direction or the point-of-gaze by monitoring the eyes' movement, and the ultimate goal is to get the gaze point on the computer screen or something else. How to estimate the gaze direction and the point-of-gaze quickly and accurately are the main issues of the gaze point estimation technology. As an interaction method to determine the user's regions of interest for human–machine interface, gaze point estimation technology has already been available for various purposes and has been applied to some other fields. There are a variety of gaze estimation methods proposed by many researchers [1]. In this paper, we would like to discuss some typical methods in the following section.

The first one is a methodology for determining point-of-gaze based on head-mounted system, which need fixed eyes camera and screen camera on helmet or frames, although the method has good accuracy and robustness, it increases the complexity of the system. Another one is a gaze point estimation method based on pupil center cornea reflection (PCCR) technique [2], based on the principle of bright and dark pupil, the method extracts pupils from the image captured by a static camera, then corrects the relative position of camera and eyeball based on the theory of corneal reflection, and regards the corneal reflection point as the base point. Thereby, the pupils' center coordinates are defined as the gaze position. Now, this method has been extensively applied in eye trackers, such as Tobii system developed by the Swedish company. Although its real-time performance, it increases position error as well. In addition, the infrared light within a certain wavelength range causes significant damage to eyes, such as retinal burn and cataracts. Therefore, based on hereinbefore problems, the ideal eye tracking system in the future should be a flexible head free gaze tracking system [3], without limitation of LED light source, low cost, high reliability and multi-channel system.

According to the discussion above, we want to estimate the gaze direction in respect to the face position to provide a better prediction. So, in this study, a gaze estimation method based on face normal is presented, this method employs spatial geometric relationships of facial feature points to calculate face normal, and gaze point can be figured out by the intersection with the computer screen. This system can be used in moving object tracking system. This proposed method not only avoids the damage of the infrared light to human eyes and restrictions of helmet gaze tracking system, but also releases head's movement, and reduces the system cost and simplifies the structure of the equipment.

The rest of the paper is organized as follows: in Sect. 2, the structure of the proposed system and flow chart of gaze point estimation method based on face normal are presented. In Sect. 3, we discuss the proposed algorithm in detail. In Sect. 4, several comparison experiments are provided, and test results illustrate the pros. and cons. of the proposed algorithm. Finally, we summarize our approach and discuss its limitations in Sect. 5.

## 2 Structure of the proposed system

The proposed system is composed by a binocular camera and a computer screen (see Fig. 1). When the user focuses his eyes on the screen, we firstly detect the human face and extract the center points' 2D coordinates of feature area such as two pupils, two nostrils and the mouth, and then the stereo vision is used to calculate their three-dimensional coordinates in the world coordinate system and we can get a fitting plane according to these points. Next, the normal of the fitting plane and the intersection point of the normal and screen can be calculated. Finally, the intersection point is the point-of-gaze we are looking for. The flow of the proposed algorithm is shown in Fig. 2:

## 3 Description of the proposed algorithm

### 3.1 Binocular calibration

The relative positions of camera coordinate system, world coordinate system, screen coordinate system and image coordinate system are shown in Fig. 3.

The relationship of point $P$ in the world coordinate system $O_wX_wY_wZ_w$ and its projection point $P(u, v)$ in the computer coordinate system $uv$($P(x, y)$ in Fig. 3 is in image coordinate system, the same point is given in Eq. (1).

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} 1/d_x & 0 & u_0 \\ 0 & 1/d_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} \alpha_x & 0 & u_0 & 0 \\ 0 & \alpha_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{X}_{\boldsymbol{w}} = \boldsymbol{M}\boldsymbol{X}_{\boldsymbol{w}} \quad (1)$$

where $\alpha_x$, $\alpha_y$, $u_0$, $v_0$ are the intrinsic parameters of a camera, and $\alpha_x = f/d_x$, $\alpha_y = f/d_y$, respectively, $\boldsymbol{M}_1$ is the matrix of intrinsic parameter of a camera, $t$ and $R$ are the external parameters of a camera, $t$ is translation matrix, $R$ is rotation matrix, and $\boldsymbol{M}_2$ is the matrix of camera external parameter.

The process determining the intrinsic and external parameters of a camera is called camera calibration. $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ can be calculated after the calibration.

A flat panel calibration method [4] is used to calibrate binocular cameras in this paper, including three major steps as given below.

*Step 1* Capture 15–20 pairs of calibration plate images at different angles by moving the stereo cameras or the calibration plates.

*Step 2* Detect feature points and solve intrinsic and external parameters of the left and the right camera.

*Step 3* Use the least square method to solve all of the parameters accurately, calculate the matrixes of intrinsic and external parameter of each camera and their relative position can be calculated.

### 3.2 Face detection

Face detection technology [5] has many important applications in many fields. For example: face recognition based on video surveillance system at airport, security area, etc. in the past 10 years, face detection [6, 7] has been a very challenging problem at image processing area and many algorithms have been proposed.

In 2001, Viola and Jones proposed an important algorithm which is called fast face detection based on Ada- Boost. AdaBoost provides an effective learning algorithm and strong bounds on generalization performance. So, in this study, we use this method for face detection.

**3.2.1 Features definition—**In object detecting procedure, we classify images based on some simple features. It is far superior to use features rather than the pixels directly. The most common reason is that features can encode ad-hoc knowledge which is difficult to learn by a finite training data. For this system, there is also another critical motivation for using these features: feature-based system sometimes operates much faster than pixel-based system.

The simple features used are Haar [8] basis functions which have been used by Papageorgiou et al. [9]. More specifically, we use three kinds of features. As given in Fig. 4, the value of a two-rectangle feature is the difference between the sums of the pixels within two rectangular regions, which have the same size and shape and are horizontally or vertically adjacent. The sum of the pixels in the white rectangles is subtracted from the sum of pixels in the gray rectangles. Two-rectangle features are shown in Fig. 4a, b. Figure 4c shows a three-rectangle feature, and Fig. 4d is a four-rectangle feature. A three-rectangle feature computes the sum within two outside rectangles and subtracted from the sum of the center rectangle. Finally, a four-rectangle feature computes the difference between diagonal pairs of rectangles.

**3.2.2 Integral image—**Rectangle features can be computed very quickly using an intermediate representation of the image which we call integral image. The integral image in Fig. 5 at location *x, y* defines the sum of the pixels above and left of *x, y*, respectively:

$$ii(x,y)=\sum_{x'\le x,y'\le y} i(x',y') \quad (2)$$

where *ii*(*x, y*) is the integral image and *i*(*x, y*)is the original image.

Using the following pair of recurrences:

$$s(x,y)=s(x,y-1)+i(x,y) \quad (3)$$

$$ii(x,y)=ii(x,y-1)+s(x,y) \quad (4)$$

where $s(x, y)$ is the accumulation of row, $s(x,-1) = 0$ and $ii(-1, y) = 0$. So, integral image can be calculated in one pass through the original image.

In Fig. 5, the sum of the pixels in rectangle $D$ can be calculated with four references. The value of the integral image at location 1 is the sum of the pixels in rectangle $A$ and we define it as $ii(1) = A$. The value at location 2 is A + B, at location 3 is A + C, and at location 4 is A + B + C + D. So, the sum in area D can be calculated as the integral image values in these locations and can be defined as $ii(4) + ii(1) - (ii(2) + ii(3))$.

**3.2.3 Adaboost algorithm—**Given a feature descriptor set and a training set consist of positive and negative sample images, many machine learning-based approaches can be used to train a classification function. In some papers, AdaBoost algorithm [10] is used to simplify features set and train the classifier. The original purpose of AdaBoost learning algorithm is used to boost the strong classifier by weak learning algorithm and some scheme in AdaBoost learning procedure guarantee its high performance. Freund and Schapire [9] proved that the training error of strong classifier approaches zero exponentially according to the number of rounds. Some results proved its generalization performance, and AdaBoost algorithm can achieve large margins.

The weak classification algorithm is used to figure out the rectangle feature which can separate the positive and negative examples by the greatest extent. For each feature, the weak classifier determines the optimal classification function to minimize the number of misclassified examples. A weak classifier $h_j(x)$ consists of a feature $f_j$, a threshold $\theta_j$ and a parity $p_j$ indicating the direction of the inequality sign:

$$h_j(x)=\left\{ \begin{array}{cc} 1 & if\ p_j f_j(x)<p_j\theta_j \\ 0 & otherwise \end{array} \right\} \quad (5)$$

Here, $x$ is a 24 × 24 pixel sub-window of an image. A summary of the boosting process [8] is given as follows.

- Given images $(x_1, y_1), \cdots (x_n, y_n)$ where $y_i = 0, 1$ for negative and positive examples, respectively.
- Initialize weights $\omega_{1,i}=\frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$, respectively, where $m$ and $l$ are the number of negatives and positives, respectively.
- For $t = 1, \ldots, T$
  - **a.** Normalize the weights

$$\omega_{t,i} \leftarrow \frac{\omega_{t,i}}{\sum_{j=1}^{n}\omega_{t,j}}$$

where $\omega_t$ is a probability distribution.

**b.** For each feature $j$, train a classifier $h_j$ which is restricted to using a single feature. The error is evaluated with respect to $\omega_t$, $\varepsilon_j = \sum_i \omega_i |h_j(x_i) - y_i|$

**c.** Choose the classifier $h_t$, with the lowest error $\varepsilon_t$

**d.** Update the weights:

$$\omega_{t+1,i} = \omega_{t,i} \beta_t^{1-e_i}$$

where $e_i = 0$ if example $x_i$ is classified.

correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$.

- The final strong classifier is:

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$$

where $\alpha_t = \log \frac{1}{\beta_t}$.

As discussed above, Viola–Jones algorithm [11] improves the speed of eigen value calculation and reduce the risk of wrong recognition, experimental results in Sect. 4 prove its efficiency and feasibility.

### 3.3 Feature point extraction

**3.3.1 Nostrils detection—**The color of nostrils has a significant saturation. Depending on its black color, the threshold must be defined and two clustering centers of saturation can be found. Extraction of nostrils involves four steps.

- Divide images into right and left sub-images to get the two distinct nostrils.
- Convert RGB images to HSV color space, and get hue and saturation of two sub-images.
- Segment the pixels which have a saturation above a certain thresh (0.8 times of the maximum saturation).
- Calculate positions which represent the centroid of the nostrils area.

Sometimes, finding nostrils in an area given by face's geometry area depends on the angle between camera and face; we should keep the nostrils visible in the images.

**3.3.2 Mouth detection—**Detecting the middle position of the mouth area is an important task in the proposed method. There are a lot of methods focusing on this issue, such as gradient horizontal, vertical decent, hue or saturation. In this study, it is implemented based on the distinct hue of lips and this area can be segmented by a predefined hue threshold value.

- Convert RGB images to HSV color space, and get hue and saturation of the image.
- Erode and dilate Hue component of the image, then convert image to binary image.
- Multiply found values with saturation and only use pixel which have a value beyond a certain thresh (0.5 times of the maximum saturation).
- Segment non-zero pixels and calculate the centroid, which is the center coordinates of mouth area.

Comparing with other methods, this method is scene illumination independent [12–15], thus intensity and direction of the light source cannot influence the results.

**3.3.3 Pupils detection—**Pupils detection begins with whether an iris is visible in the captured image, and then precisely locating its inner and outer boundaries [16] (pupil and limbus). Because of the circular geometry of the iris, these tasks can be accomplished from the input image $I(x, y)$ by integro-differential operators which search over the image domain $(x, y)$ to get the maximum in the blurred partial derivative, with respect to increasing radius $r$, of the normalized contour integral $I(x, y)$ along a circular arc $\mathrm{d}s$ with radius $r$ and center coordinates $x_0, y_0$:

$$\max_{(r,x_0,y_0)} \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint_{r,x_0,y_0} \frac{I(x,y)}{2\pi r} \mathrm{d}s \right| \quad (6)$$

and $G_\sigma(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(r-r_0)^2}{2\sigma^2}}$

where *denotes convolution and $G_\sigma(r)$ is a smoothing operator such as Gaussian function with scale $\sigma$. The proposed operator behaves as a circle edge detector with several different scales, which searches iteratively for a maximum contour integral derivative with increasing radius and scales in the parameters space and radius $(r, x_0, y_0)$ which defines the path of contour integration.

Firstly, the scale factor $\sigma$ is set for coarse scale analysis so that only the very pronounced circular transition from iris to (white) sclera can be detected and this strong circular boundary can be detected precisely; Secondly, we search within the confined central interior of iris area to get the fainter papillary boundary, using a finer convolution scale $\sigma$ and a smaller search range which define the paths $(r, x_0, y_0)$ of contour integration. In the initial search for the outer bounds of the iris, the angular arc of contour integration $\mathrm{d}s$ is restricted in range of two opposing 90° cones centered on the horizontal meridian, since eyelids generally obscure the upper and lower limbus of the iris. Then, in the subsequent interior search for the papillary boundary, the arc of contour integration $\mathrm{d}s$ in Eq. (6) is restricted to upper 270°. To avoid the corneal specular reflection that is usually superimposed in the lower 90° cone of the iris from the illuminator located below the video camera. Taking the absolute value arithmetic in Eq. (6) is not necessary when the operator is used to locate the

outer boundary of the iris, since the intensity of sclera is always lighter than the iris and so the smoothed partial derivative with increasing radius near the limbus is always positive. However, the intensity of pupil is not always darker than the iris, such as persons who have normal early cataract or significant back-scattered light from the lens and vitreous humor; applying the absolute value in Eq. (6) makes the operator a better circular edge-finder regardless of such polarity-reversing conditions. With $\sigma$ automatically adopting to the stage of search for both the pupil and limbus and being correspondingly finer in successive iterations, the operator defined in Eq. (6) has been proven to be virtually infallible in locating the visible inner and outer annular boundaries of irises.

As discussed above, the integro-differential operators proposed by Daugman [17] is one of the most widely used algorithm, some conditions should be met to guarantee it works well, such as, high-quality image, and environment should be stable.

### 3.4 Stereo matching with epipolar constraint

Epipolar geometry [18] of two perspective cameras, as given in Fig. 6, assigns each point $p_1$ in one image to an epipolar line $l_2$ in the second image. All epipolar lines in each image intersect in the epipoles $e_1$ and $e_2$.

After calibrating the stereo cameras, the epipolar constraint equation can be calculated by the projection matrix of $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$.

The epipolar constraint equation [19] is given in Eq. (7).

$$p_2^T F p_1 = 0 \quad (7)$$

where $p_1 = (x, y)^T$ is the feature point in left image, and $p_2 = (x', y')^T$ is the matching point in right image, $\boldsymbol{F} = [\boldsymbol{m}]_\times \boldsymbol{M}_{21} \boldsymbol{M}_{11}^{-1}$ denotes the fundamental matrix.

In the proposed method, the scene structure is modeled by a set of planar disparity planes; disparity planes can be reduced by extracting a set of disparity planes that is sufficient to represent the scene structure. This is done by applying local matching [20, 21] in the pixel domain followed by disparity plane estimation. Local matching requires defining a matching score and a search window. The principle of local matching is shown in Fig. 7.

In Fig. (7), the size of the matching window is $(2n + 1) \times (2m + 1)$, $n = 0, 1, 2\ldots$, $m = 0, 1, 2\ldots$, $[-D, +D]$ is the search scope, and $d$ is the disparity values.

In our approach, we use a self-adapting dissimilarity measure called sum of absolute difference (SAD).

$$C_{SAD}(p, d) = \sum_{(x,y) \in W_p} |I_l(x, y) - I_r(x + d, y)| \quad (8)$$

where $W_p$ is the matching window.

### 3.5 Three-dimensional reconstruction

The model of 3D reconstruction [22] based on pinhole camera model is illustrated by Fig. 8. $P_l$, $p_r$ are projection points in the left and right image of $P$.

According to the projective theory, the relationship of $P$'s 3D coordinate and its projective point in camera coordinate system is shown as Eq. (9).

$$Z_{lc}\begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = \boldsymbol{M}_l \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad z_{rc}\begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = \boldsymbol{M}_r \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{9}$$

where $z_{lc}$ and $z_{rc}$ are the $z$ axis coordinates of point $P$ in left and right camera coordinate system, $\boldsymbol{M}_l$ and $\boldsymbol{M}_r$ are projection matrices.

From Eq. (9), with the 3 × 4 matrix of $\boldsymbol{M}_l$ and $\boldsymbol{M}_r$, we can get six equations; then we can shrink them into four linear independent equations by eliminating $z_{lc}$ and $z_{rc}$ in Eq. (9) and rewrite it in matrix form as given in Eq. (10).

$$\boldsymbol{K} \cdot \boldsymbol{P}_w = \boldsymbol{U} \tag{10}$$

$\boldsymbol{K}$ and $\boldsymbol{U}$ are the coefficient matrices of simultaneous equations, world coordinate of $\boldsymbol{P}_w$ can be calculated by the least square method, as shown in Eq. (11).

$$\boldsymbol{P}_w = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{U} \tag{11}$$

The other feature points' 3D coordinates can be calculated by repeating steps Eqs. (9)–(11).

### 3.6 Fit the face plane and normal

According to the least square method, approximate face plane will be fitted with the five feature points.

Space plane equation is $Ax + By + Cz + 1 = 0$, given $n$ points; the equation of fitting plane can be written in matrix form as Eq. (12):

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} \tag{12}$$

Multiply both sides of Eq. (12) by $\begin{bmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{bmatrix}^T$.

then, we can get the following equation:

$$\begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i z_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i z_i \\ \sum x_i z_i & \sum y_i z_i & \sum z_i^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} -\sum x_i \\ -\sum y_i \\ -\sum z_i \end{bmatrix}$$

and we can solve Eq. (12) by the following Eq. (13).

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \left[ \begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i z_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i z_i \\ \sum x_i z_i & \sum y_i z_i & \sum z_i^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum \mathrm{x}_i \\ -\sum \mathrm{y}_i \\ -\sum \mathrm{z}_i \end{bmatrix} \right] \quad (13)$$

As given in Eq. (13), the parameters A, B and C can be figured out, the face plane equation can be calculated as well.

The point-of-gaze is the intersection of left and right sight line; it is dependent with face and eyeballs movement. In this study, we propose a new method to define and calculate the gaze point. The gaze point can be defined as the intersection of the face normal via midpoint between two eyes and the screen. The steps to determine normal are given as follows.

If the coordinates of two pupils are $P_1(x_1, y_1, z_1)$and $P_2(x_2, y_2, z_2)$, the midpoint is.

$$P(x_0, y_0, z_0) = \frac{P_1(x_1, y_1, z_1) + P_2(x_2, y_2, z_2)}{2} \quad (14)$$

So, the normal equation on this point is.

$$\frac{x - x_0}{A} = \frac{y - y_0}{B} = \frac{z - z_0}{C} \quad (15)$$

### 3.7 Determine the point-of-gaze

In this binocular vision system, we suppose that the optical axis of the left camera is the same as the *Z* axis of the world coordinate system, and the positive direction is pointing to the front of the camera. *XOY* plane is the imaging plane of left camera. The direction from left to right camera is considered as the minus *X* direction, and the positive direction of the Y-axis is perpendicular upwards to *XOZ* plane.

In this study, we assume that $XY$ plane of the camera coordinate system overlaps with computer screen, therefore, with the face normal equation the position of gaze is the intersection of screen plane and the normal. Hence, the 2D coordinate of the point-of-gaze on the screen plane can be figured out by setting the normal equation $Z = 0$.

## 4 Experimental results

Some experiments based on the proposed method with binocular system have been done. The good performance obtained confirms the accuracy and robustness of the proposed method. These experiments using images of real scene are now given below.

First, we develop a graphical user interface(GUI) at Matlab.R2009a platform, read video from binocular camera, and then detect face in video frames. Figure 9 shows the effects of the interface when the video is loaded.

In this study, we roll the planar calibration board from different angles in front of the cameras to calibrate the binocular vision system, each of the cameras take 20 photo pairs in different poses, which are given in Fig. 10.

Zhang's flat panel calibration method has been used in our experiments, intrinsic and external parameters of the binocular cameras are given in Table 1.

As the world coordinate system overlaps the left camera coordinate system, so all of the external parameters of left camera are zero, and Table 2 shows external parameters of the two cameras.

When we get the camera external parameters, the rotation matrix R and translation vector t are available; and $\boldsymbol{M_2}$ can be calculated. $\boldsymbol{M_1}$ also can be obtained via intrinsic parameters. The projection matrix $\boldsymbol{M}$ and fundamental matrix $\boldsymbol{F}$ can be calculated by stereo matching.

$$F=[6.33597152044979e-070.000188166666910164\ 0.417706172171736$$
$$-0.000365603513857672-0.000244126563910449\ 36.8888330846424$$
$$-0.950295359678450-37.0542275562950\ 2452.72993678888]$$

After the cameras capture the stereo video, the results of left image with detected face and facial organ are shown in Figs. 11 and 12.

Extracted feature points are given in Fig. 13.

Then, we use epipolar constraint and stereo matching for calculating matching points. Figure 14 shows the results that extract the matching points of mouth in right image.

The experiment results based on stereo matching and 3D reconstruction for facial organ are listed in Table 3 as follows.

The approximate face plane by the least-square fitting planar method is shown in the Fig. 15.

The fitting plane equation is:

$$-0.0356 \times x - 0.0684 \times y - z + 0.5802 = 0$$

And the normal equation via the midpoint of two pupils is:

$$\frac{\mathrm{x}-(-0.0222)}{-0.0356} = \frac{y-(-0.0466)}{-0.0684} = \frac{z-0.5851}{-1}$$

Let $z = 0$, so the 3D coordinate of the point-of-gaze is (−0.0430, −0.0866, 0) (m).

The comparable results verify that the calculated gaze point is approximate with true position what we gaze on the screen according to the world coordinate system. Furthermore, we make a comparison between two traditional methods and the method we proposed, given the position of five points from upper left to lower right of the screen, and their actual positions are a (−0.1625, −0.0582, 0), b (−0.0996, −0.1160, 0), c (−0.0301, −0.1715, 0), d (0.0826, −0.2213, 0) and e (0.1899, −0.2703, 0). and the measurement of the proposed method for each point is a′ (−0.1631, −0.0564, 0), b′ (−0.1023, −0.1169, 0), c′ (−0.0326, −0.1699, 0), d′ (0.0815, −0.2174, 0), e′ (0.1854, −0.2694, 0). Figure 16a shows the comparable results and Fig. 16b gives more detail of the results of the proposed method.

Using the three-dimensional function in solid geometry, the distance between any two points $X(x_x, y_x, z_x)$ and $Y(x_y, y_y, z_y)$ in space can be determined by Eq. (16):

$$L_{XY} = \sqrt{(x_x - x_y)^2 + (y_x - y_y)^2 + (z_x - z_y)^2} \quad (16)$$

Therefore, we can confirm the better performance of the proposed method by calculating the distance between the measurement data of different methods and their actual location. According to Eq. (16), the errors are given in the Table 4.

## 5 Conclusions

A new methodology to determine the point-of-gaze based on face normal is presented in this paper. The method can be used for interacting with an application such as auxiliary locating moving object in object tracking system. According to the location based on the world coordinate system and direction of the coordinate axis, test results show that all of the measurement what we got are correct. In addition, errors of the proposed method are far less than that of the other two methods. And the measurement precision is higher than 5 mm. Therefore, the method proposed in the paper is more valuable in practical application.

In summary, the primary contribution of the proposed method is that the system does not need an infrared light, which avoids damaging the eyes of human. This methodology does not require a special head tracking system, only needs a pair of video grabbing device. Furthermore, there are no restrictions to head movement, so, the system cost is reduced and the equipment is simplified as well. However, the method still has some disadvantages and

limitations. We consider it essential to continue the theoretical study of the method to build a perfect system for gaze estimation technology.

## Acknowledgments

This work was supported by Fundamental Research Funds for the Central Universities (Grant JB141307); National Nature Science Foundation of China (NSFC) (Grants 61201290), and NSFC Grants 61105066, 61305041, 61305040; the China Scholarship Council (CSC) and the National Institutes of Health (Grant R01CA165255) of the United States.

## References

1. Villanueva, Arantxa, Cabeza, Rafael. A novel gaze estimation system with one calibration point. IEEE Transactions on Systems, Man, and Cybernetics. 2008; 38(6):1123–1138.
2. Shao, G., Che, M., Zhang, B., Cen, K., Gao, W. A novel simple 2D model of eye gaze estimation. IEEE 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC); Nanjing, Jiangsu. 2010. p. 300-304.
3. Zhu ZW, Ji Q. Novel eye gaze tracking techniques under natural head movement. IEEE Trans Biomed Eng. 2007; 54(12):2246–2260. [PubMed: 18075041]
4. Zhang, Ning, Chang, Lei, Xu, Xiping. Research on the technology of three-dimensional reconstruction based on machine vision. Laser Optoelectron Prog. 2012; doi: 10.3788/LOP49.051001
5. JONESM. Face recognition: where we are and where to go from here. IEEE TEIS. 2009; 129:770–777.
6. Wen G, Bo C, Shan Shiguang, et al. The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Trans Syst Man Cybern. 2008; 38(1):149–161.
7. Yamazoe, H., Utsumi, A., Yonezawa, T., Abe, S. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. Proceedings of the 2008 symposium on eye tracking research & applications; ACM; 2008. p. 245-250.
8. Yanhui S, Wenyong W, Xiaochun C. The amelioration to the face recognition algorithm based on the Viola-Jones frame. J Northeast Norm Univ. 2005; 37(3):24–27.
9. Yoav, Freund, Schapire, RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997; 55(1):119–139.
10. Luan B, Sörös P, Sejdić E. A study of brain networks associated with swallowing using graph-theoretical approaches. PLoS ONE. 2013; 8(8):e73577. [PubMed: 24009758]
11. Daugman J. How iris recognition works. Circuit Syst Video Technol IEEE Trans. 2004; 14(1):21–30.
12. Sun W, Guo BL, Li DJ, Jia W. Fast single-image dehazing method for visible-light systems. Opt Eng. 2013; 52(9):093103.
13. Wei S. A new single image fog removal algorithm based on physical model. Int J Light Electron Opt. 2013; 124(21):4770–4775.
14. Sun W, Han L, Guo B, Jia W, Sun M. A fast color image enhancement algorithm based on max intensity channel. J Mod Opt. 2014; 61(6):466–477. [PubMed: 25110395]
15. Paul, Viola, Jones, Michael J. Robust real-time face detection. Int J Comp Vis. 2004; 57(2):137–154.
16. Tan TN, He ZF, Sun ZN. Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition. Image Vis Comput. 2010; doi: 10.1016/j.imavis.2009.05.008
17. Constantine, Papageorgiou, Tomaso, Poggio. A trainable system for object detection. Int J Comput Vis. 2000; 38(1):15–33.
18. Castillo Carlos D, Jacobs DW. Using stereo matching with general epipolar geometry for 2D face recognition across pose. IEEE Trans Pattern Anal Mach Intell. 2009; 31(12):1198–2304.

19. Lu, Jiangbo, Hua, Cai, Jian-Guang, Lou, Jiang, Li. An epipolar geometry-based fast disparity estimation algorithm for multiview image and video coding. IEEE Trans Circuit Syst Video Technol. 2007; 17(6):737–750.

20. Zhao Y, Liu HX, Wang ZY, et al. An improved nearest neighbor searching method for classification problems. J Nanjing Univ (Nat Sci). 2009; 45:455–462.

21. Hirschmuller, Heiko, Scharstein, Daniel. Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans Pattern Anal Mach Intell. 2009; 31(9):1582–1599. [PubMed: 19574620]

22. Geiger, Andreas, Ziegler, Julius, Stiller, Christoph. StereoScan: dense 3D reconstruction in real-time. IEEE Intelligent vehicles symposium (IV); Baden-Baden. 2011. p. 963-968.

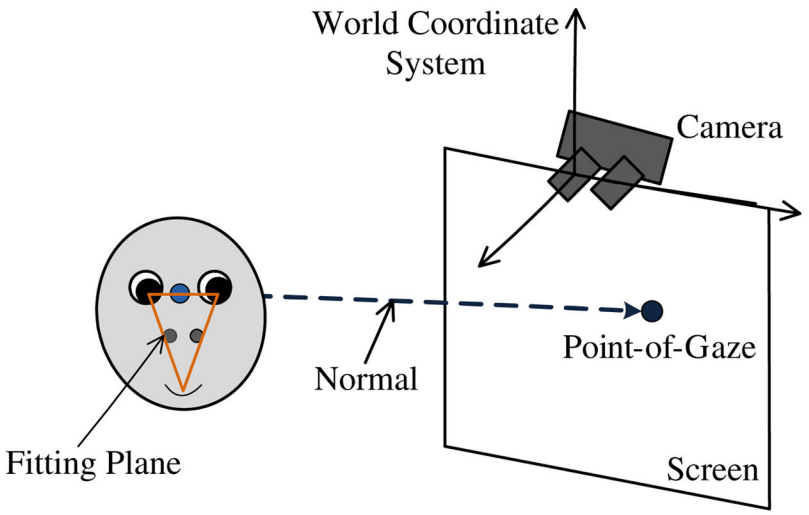


**Fig. 1.**
The scheme of the proposed system

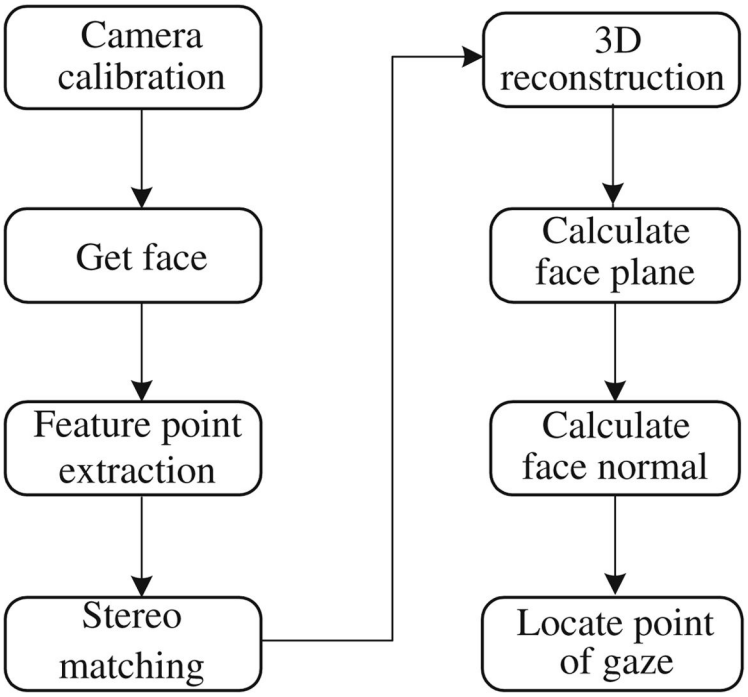


**Fig. 2.**
Flow chart of gaze estimation method based on face normal

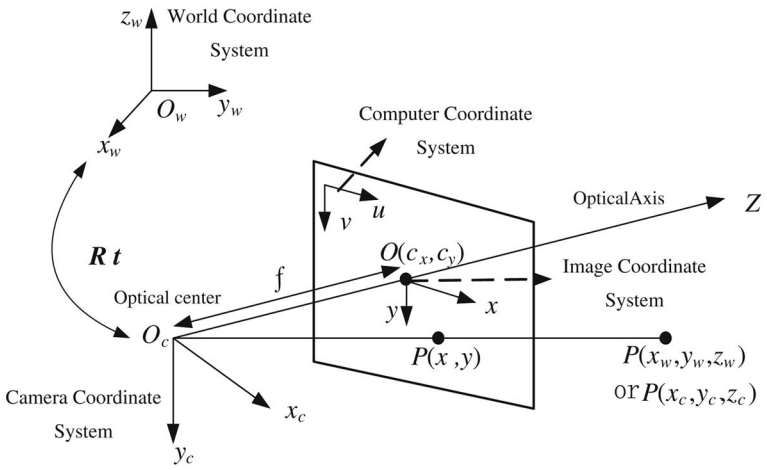


**Fig. 3.**
The relationship of four coordinate systems

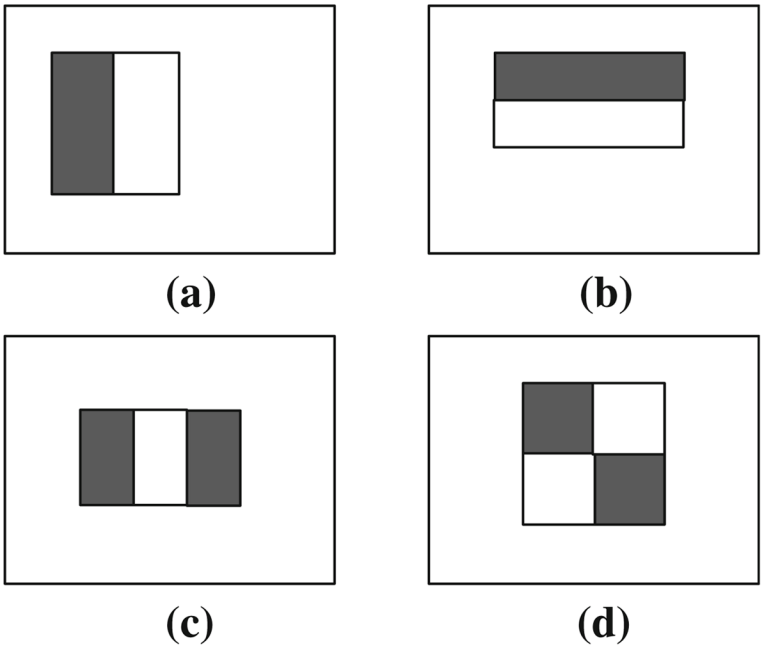


**Fig. 4.**
Example of rectangle features

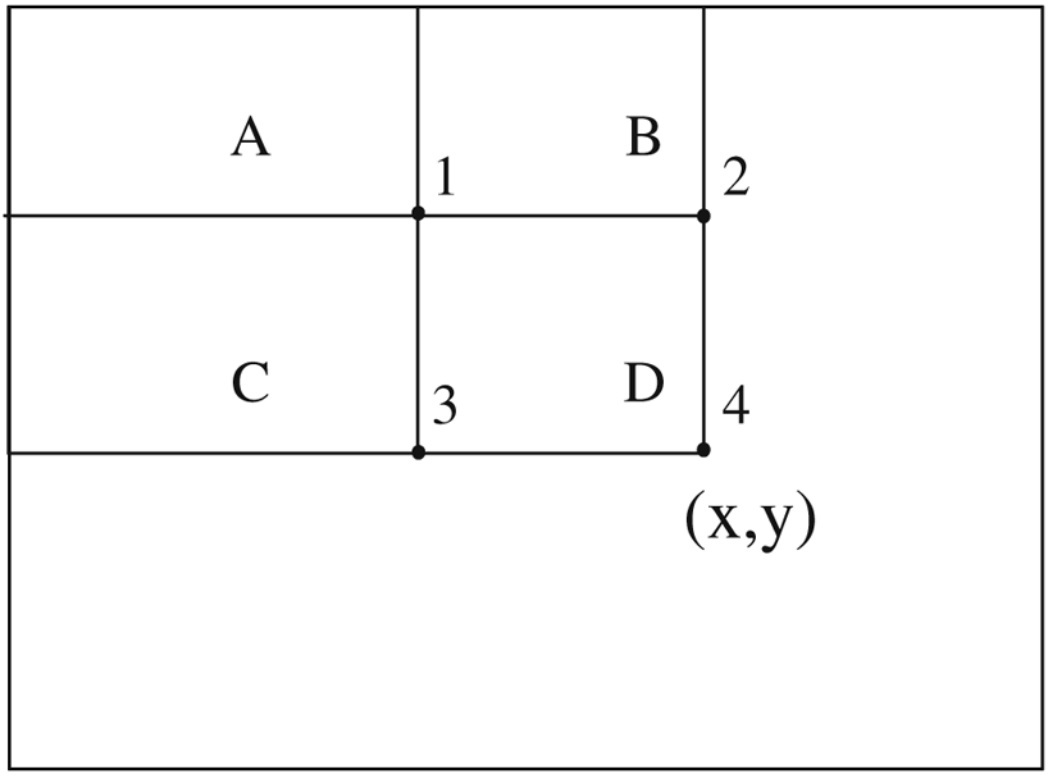


**Fig. 5.**
The integral image calculation

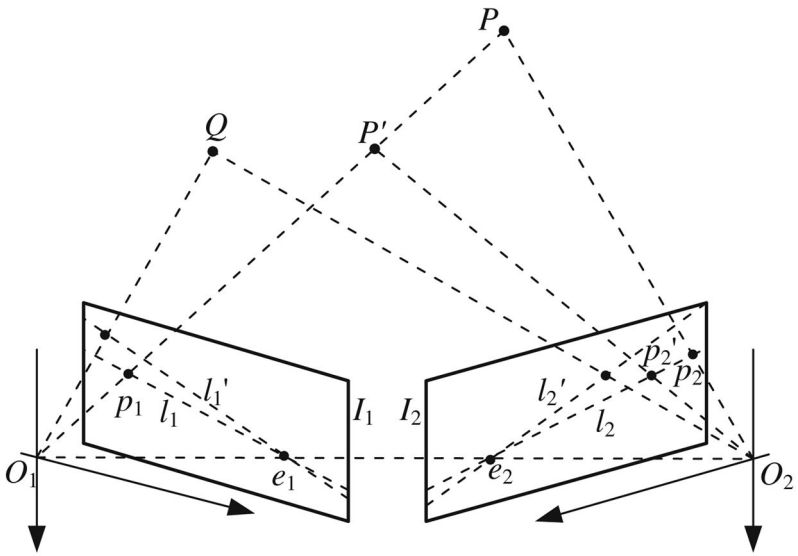


**Fig. 6.**
The epipolar geometry of two perspective cameras

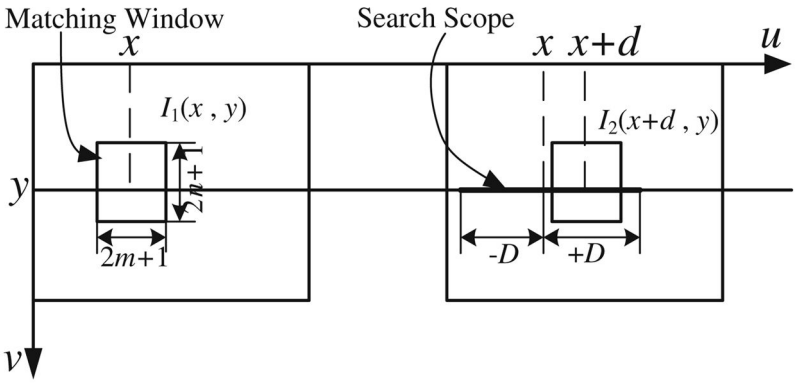


**Fig. 7.**
The principle of local matching

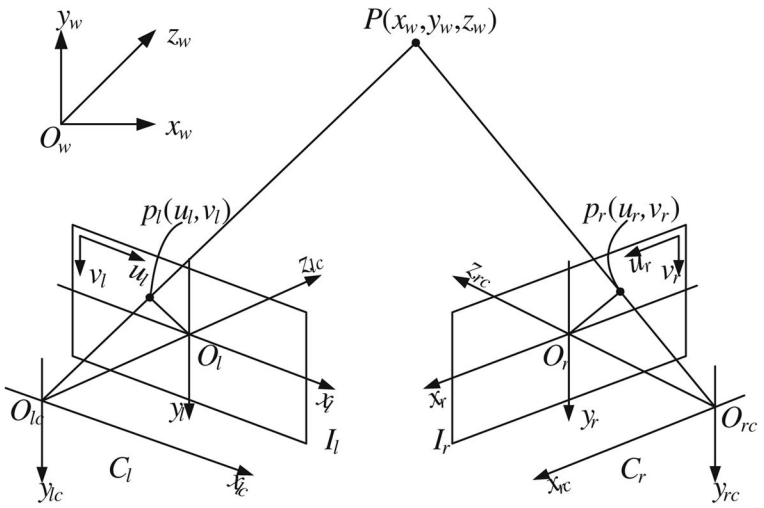


**Fig. 8.**
The principle of 3D reconstruction


**Fig. 9.**
The GUI of the experiment platform

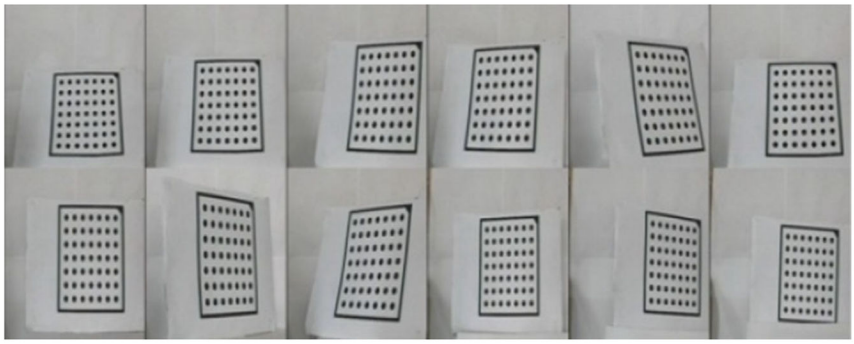

**Fig. 10.**
The sample of calibration plate images

**Fig. 11.**
Feature extraction

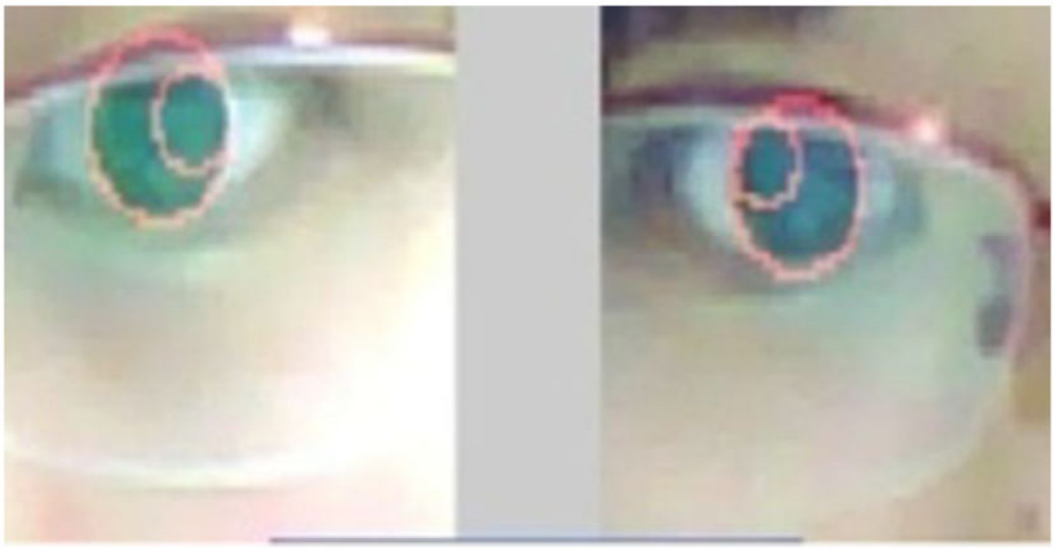

**Fig. 12.**
Eyes detection

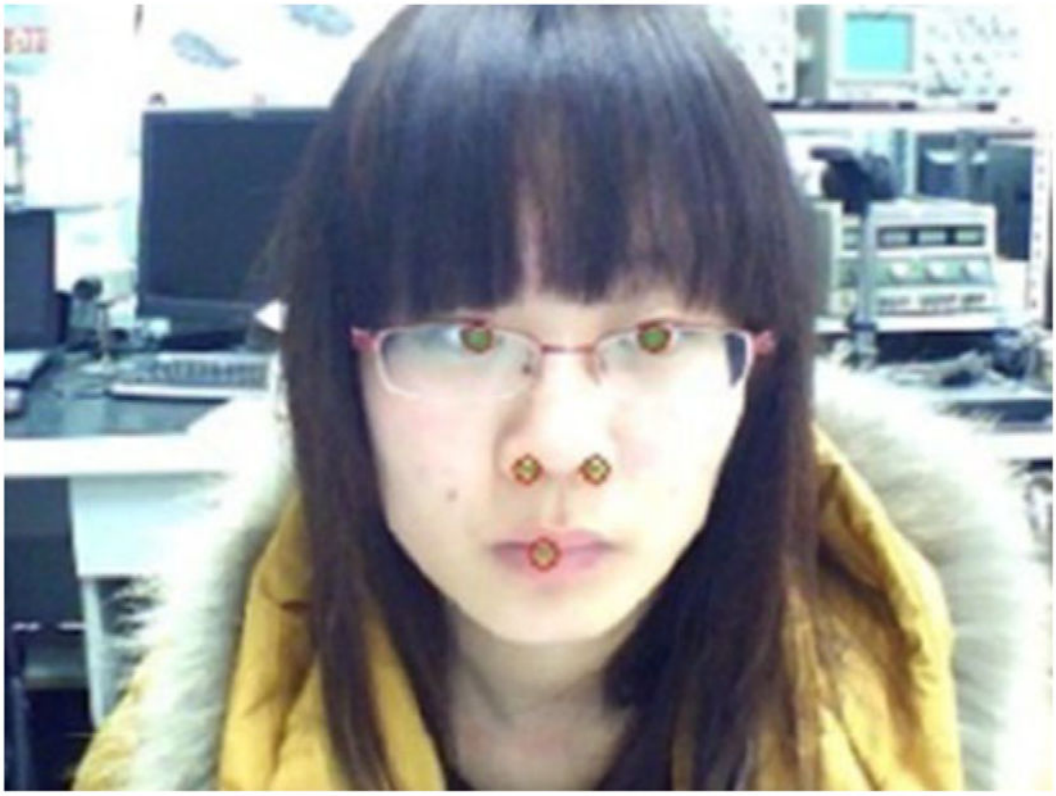

**Fig. 13.**
Feature points extraction

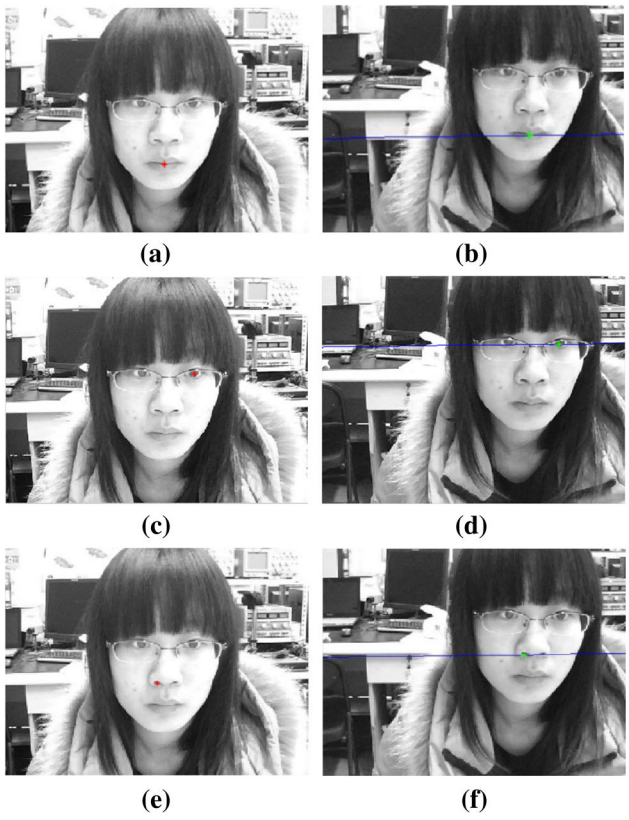


**Fig. 14.**

Epipolar constraint and stereo matching for mouth and left eye and nostril **a** Feature of mouth **b** Epilpolar and matching point in the right image **c** Feature of left eye **d** Epilpolar and matching point in the right image **e** Feature of left nostril **f** Epilpolar and matching point in the right image

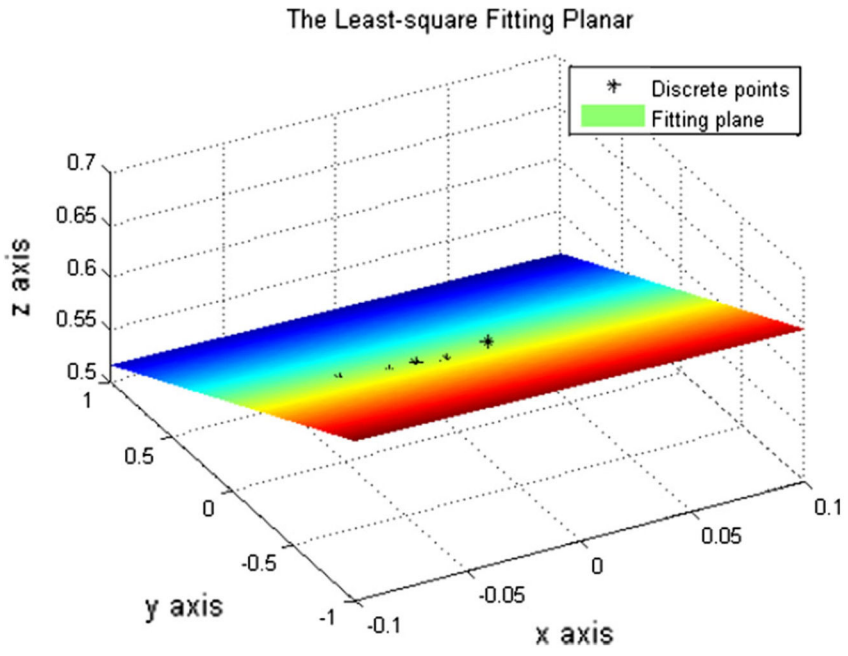


**Fig. 15.**
The least-square fitting planar for the five features

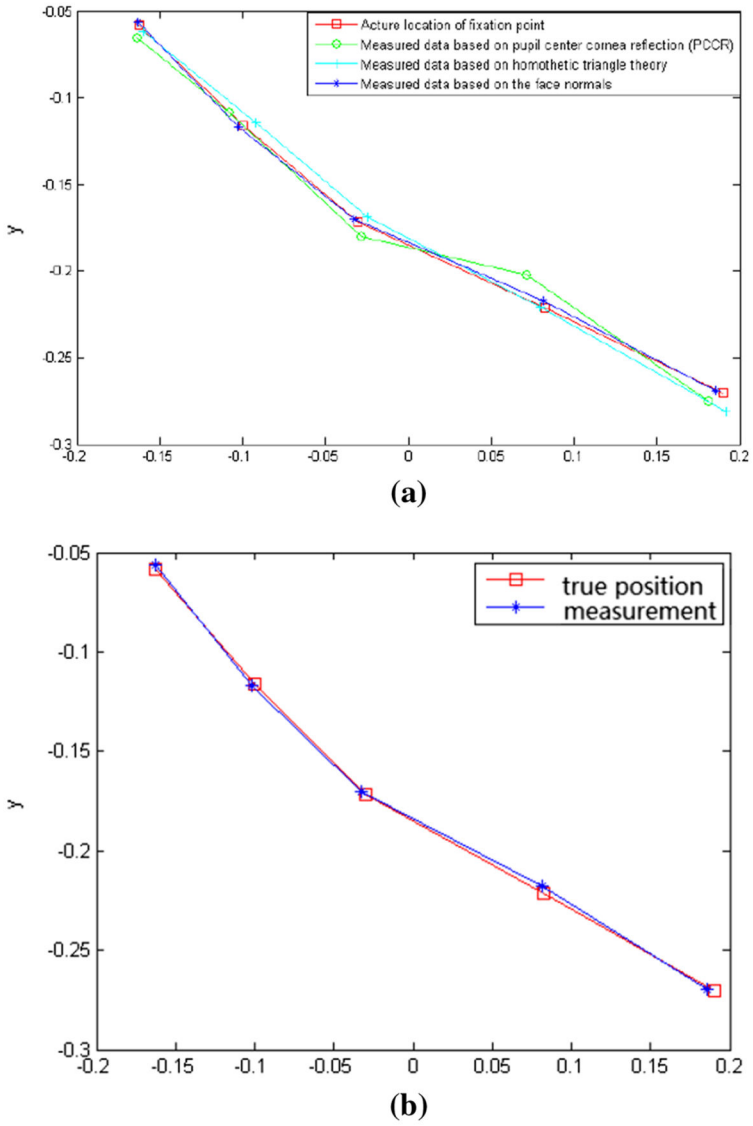


**Fig. 16.**

Gaze point test results. **a** The comparable results of gaze point obtained by several methods **b** The detail results by the proposed method

**Table 1**

Camera intrinsic parameters

| Camera intrinsic parameters | Left camera | Right camera |
|---|---|---|
| f (m) | 0.0133757 | 0.0134837 |
| dx (m) | 1.48364e–005 | 1.48334e–005 |
| dy (m) | 1.48e–005 | 1.48e–005 |
| cx (pixel) | 267.464 | 300.884 |
| cy (pixel) | 198.802 | 131.287 |

**Table 2**

Camera external parameters

| Camera external parameters | $\alpha$(°) | $\beta$(°) | $\gamma$(°) | $t_x(m)$ | $t_y(m)$ | $t_z(m)$ |
|---|---|---|---|---|---|---|
| Left camera | 0 | 0 | 0 | 0 | 0 | 0 |
| Right camera | 0.351491 | 359.76 | 359.165 | −0.0404223 | 0.000498579 | −0.000189711 |

**Table 3**

The results of stereo matching and 3D reconstruction

| Feature points | Coordinates in left image (pixel) | Coordinates of matching point in right image (pixel) | Three-dimensional coordinate (m) |
|---|---|---|---|
| Mouth | (229.33, 385.84) | (325.00, 324.00) | (−0.0253, −0.1236, 0.5925) |
| Right nostril | (213.00, 340.63) | (310.00, 279.00) | (−0.0356, −0.0927, 0.5866) |
| Left nostril | (252.33, 343.17) | (350.00, 281.00) | (−0.0098, −0.0935, 0.5832) |
| Right pupil | (183.00, 271.00) | (281.00, 208.00) | (−0.0550, −0.0470, 0.5857) |
| Left pupil | (284.00, 270.00) | (383.00, 208.00) | (0.0107, −0.0461, 0.5845) |

**Table 4**

The distance between the measurement and their actual location of different methods

| Algorithm\points | a | b | c | d | e |
|---|---|---|---|---|---|
| Pupil center cornea reflection | 0.0070 | 0.0114 | 0.0085 | 0.0222 | 0.0100 |
| Homothetic triangle theory | 0.0043 | 0.0081 | 0.0061 | 0.0035 | 0.0110 |
| Face normal | 0.0019 | 0.0028 | 0.0030 | 0.0041 | 0.0046 |