Real-Time Eye-Gaze Estimation by Using a Virtual Reference Point

Tetsuo Miyake, Taro Iwami, Satoshi Horihata, and Zhong Zhang
Department of Production Systems Engineering
Toyohashi University of Technology
Hibarigaoka 1-1, Tenpaku-cho, Toyohashi 441-8580, Japan
{

Abstract: - A human interface that uses a line of sight has been studied. In previous studies, some methods use a special light source or a hardware device and others use a software program for image processing for detecting the line. In this paper we propose a new eye-gaze input method. The method uses only a single facial image acquired under ordinary lighting condition, and it requires no laborious camera calibration. A state of eye-gaze is estimated by a virtual reference point, which is determined based on a motion of feature points on a face. We demonstrate the validity of the new method by simulation and experiments.

Key-Words: - Gaze, Human interface, Real time, Single facial image, Image processing.

1 Introduction

A human interface that uses a line of sight has been studied. If a machine can detect that a person gazes at its eye, which is usually a video camera, the machine can make a certain response to the person. In previous studies, some methods use a special light source or a hardware device [1], and others use a software program for image processing [2], [3], [4] for detecting the line. Gaze direction can be detected with high spatial resolution by using the former methods, but these are not user friendly because they require the user to restrict user's motion. On the other hand, many useful applications can be realized by using the latter. But much calculation time is needed to determine gaze direction precisely, and laborious camera calibration is inevitable.

The method we have proposed [5], [6], based on image processing, can determine whether a person gazes at a camera lens or not. We have called it ”eye-gaze estimation”. The method uses only a single facial image acquired by a video camera under ordinary lighting condition without using any instrument. Previously we used a pair of small marks that were attached on a user's face. The marks however were undesirable for the users of the system.

In this paper we propose a new approach for eye-gaze estimation. A generated virtual point is substituted to the real marks. A position of the virtual point is determined based on a motion of feature points on a face between two successive images. We evaluate the new approach by simulation and experiments, and demonstrate the validity of the new method.

2 Principle of Eye-Gaze Estimation

Fig. 1 shows the image of eye region. A state of eye-gaze is estimated by relative displacement of irises from marks. The marks are attached on a facial surface where a straight line that passes through the center of the right and left eyeball intersects. We call this line an axis of eyeball. A point P in the figure is the midpoint of the center of the right and left marks, and a point Q is the midpoint of the center of right and left irises. The displacement of irises is defined by a distance of these two points. For normalization, the amount of displacement is divided by a distance L.

Because neither a feature of a human face nor lighting condition at image acquisition are symmetrical, a displacement of a front facial image with the eyes gazing at a front camera lens is not ordinarily zero. Hence the displacement on this condition is regarded as reference, and it is subtracted from that of a facial image in an arbitrary state. Let $\boldsymbol{P}$ represent a position vector of P, and $\boldsymbol{Q}$ a position vector of Q. The eye-gaze estimation is done by using a displacement d that is defined by the following equation:

$$d = \frac{||(\boldsymbol{Q}-\boldsymbol{Q_0}) - (\boldsymbol{P}-\boldsymbol{P_0})||}{L}, \tag{1}$$

where $\boldsymbol{P}_0$ and $\boldsymbol{Q}_0$ are the position vectors of corresponding points in the front facial image, which we call the reference image. If d is less than a predetermined threshold, the system estimates that the user gazes at a camera lens.

It was proved geometrically that if a camera lens or other visual targets are optically far enough from a person, the displacement d depends only on where the person gazes at and does not on its head direction [5]. In the actual system, the marks are substituted by a virtual point that is generated in a facial image based on some feature points on a face.

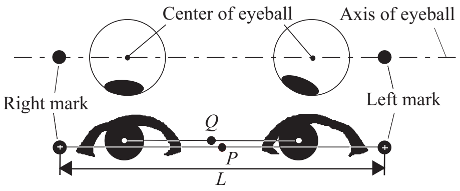


Fig. 1 Displacement of irises.

3 Image Processing for Eye-Gaze Estimation

Image processing for the eye-gaze estimation has two parts as shown in Fig. 2. The one is image processing for generating reference data, whose flow is in a left part of the figure. The processing consists of feature points extraction, detection of eye regions from the obtained image, estimation of iris center and generation of a reference point for the eye-gaze estimation. The other is real-time eye-gaze estimation, whose flow is in a right part of the figure. Image acquisition and a few steps of image processing are carried out by a special hardware entirely in parallel with others carried out by a PC. Each image has 512×480 pixels and 256 gray levels, and it is acquired at intervals of 30 msec on the average.

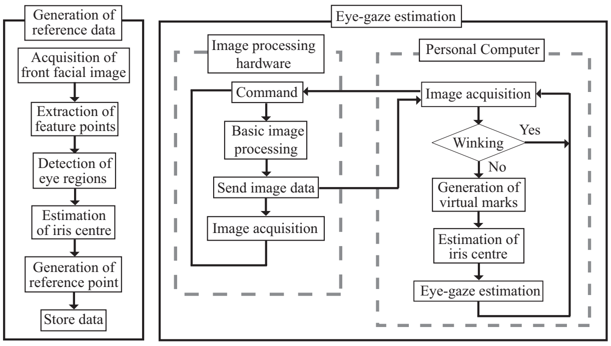

Generation of reference data
Acquisition of front facial image
Extraction of feature points
Detection of eye regions
Estimation of iris centre
Generation of reference point
Store data
Eye-gaze estimation
Image processing hardware
Command
Basic image processing
Send image data
Image acquisition
Personal Computer
Image acquisition
Winking
Yes
No
Generation of virtual marks
Estimation of iris centre
Eye-gaze estimation

Fig. 2 Flow chart of eye-gaze estimation.

3.1 Generation of Reference Data

Image processing for generating reference data is conducted as follows. The first step is acquisition of the reference image. The second step is extraction of feature points. Ten points are selected as them: outer and inner corners of both eyes, an inner edge of both eyebrows, both sides of a nose, and both sides of a mouth. The third step is detection of eye regions based on the feature points positions. After the eye regions are determined, edge points of the irises are extracted from its binarized image. By assuming that the projected image of the iris is a circle, its edge points are fit to the circle with the least squares method. A center of the circle is considered as a center of the iris. The estimated iris center is one of the reference data. Finally, the midpoint of the right and left marks are estimated by using the feature points. This point is used as a virtual reference point (VRP), which corresponds to the point P_0 in Fig. 1.

The result of image processing is shown in Fig. 3. A white circle at the middle indicates the VRP. The cross symbols indicates the center of irises and some feature points. The real marks in the image are also extracted. They are used for comparison in experiments.

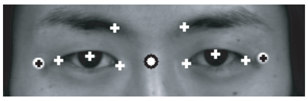
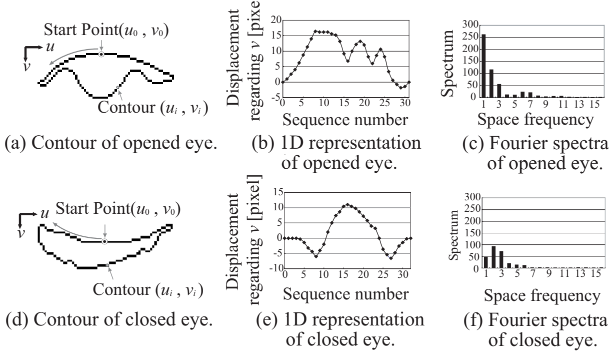


Fig. 3 Result of image processing.

3.2 Detection of Winking

The eye-gaze estimation is failed when an eye is winked. Winking can be detected by recognizing a shape of an eyelid. Fig. 4(a) shows a typical contour line of an opened eye in a binary image. The 2D contour line is represented as a function of one variable by neglecting a u coordinate. Fig. 4(b) illustrates a graph of the function. The horizontal axis of the graph indicates a sequence number of points along the contour line and the vertical axis a displacement with regard to a v coordinate. A series of points are extracted at equal distance and a number of points are reduced to 32. Discrete Fourier transform of this series yields spectra as shown in Fig. 4(c). Fig. 4(d), 4(e), and 4(f) show a contour line of a closed eye, its graph representation, and Fourier spectra respectively. The shape of an eyelid can be recognized by calculating a cross correlation between its Fourier spectra and those of a typical shape of an opend eye.



(a) Contour of opened eye.
(b) 1D representation of opened eye.
(c) Fourier spectra of opened eye.
(d) Contour of closed eye.
(e) 1D representation of closed eye.
(f) Fourier spectra of closed eye.

Fig. 4 Characteristics of opened eyes and closed eyes.

4 Eye-Gaze Estimation with Virtual Reference Point

Theoretically the video camera of our system can be placed optically far enough from a person. It is an outstanding point of our method compared to the other proposed ones. Since it is assumed that some feature points on a face lie on a plane under this condition, the position of a VRP in the facial image is estimated based on affine transformation.

4.1 Generation of Virtual Reference Point

A VRP is regarded as one of the feature points on a face. All feature points are rotated and translated in one in a 3D space as a head rotates, and are projected to a 2D image plane. By assuming that movement of the feature points between successive video images can be expressed by affine transformation, the positions of the VRP is determined.

Let $\hat{\boldsymbol{s}}$ denote a position vector of a point in the reference image, which is represented by homogeneous coordinates, and let $\hat{\boldsymbol{s}}'$ denote that of the same point in an arbitrary facial image whose state of eye-gaze is to be estimated. $\hat{\boldsymbol{s}}'$ is calculated using affine matrix defined by;

$$\hat{\boldsymbol{s}}' = \boldsymbol{A}\hat{\boldsymbol{s}}, \quad \boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & t_u \\ a_{21} & a_{22} & t_v \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

where $\hat{\boldsymbol{s}} = (u, v, 1)^T$ and $\hat{\boldsymbol{s}}' = (u', v', 1)^T$. Since the matrix $\boldsymbol{A}$ has six degrees of freedom, all elements of $\boldsymbol{A}$ can be determined from the coordinates of more than three points. By using coordinates of the ten points described in 3.1, $\boldsymbol{A}$ is determined based on the least square method. The processing flow is shown in Fig. 5.

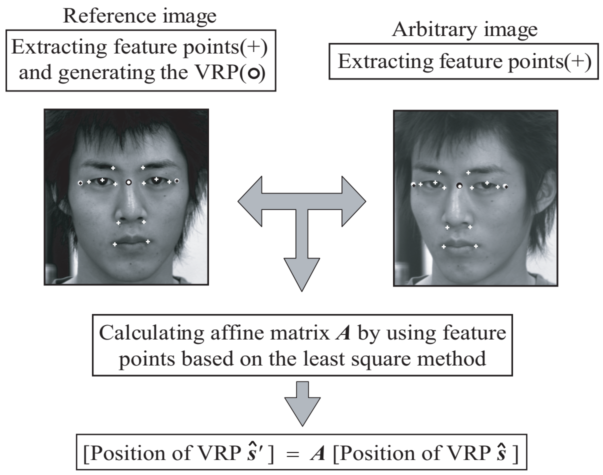

Reference image
Extracting feature points(+) and generating the VRP(○)
Arbitrary image
Extracting feature points(+)
Calculating affine matrix A by using feature points based on the least square method
[Position of VRP $\hat{s}'$] = A [Position of VRP $\hat{s}$]

Fig. 5 Determination of the position of the VRP in an arbitrary image.

Fig. 6 shows the geometrical relation between the real marks and the VRP. The x, y and z axis correspond to the axis of eyeball, the vertical axis, and the optical axis of the camera respectively. θ_x and θ_y represent the angle of head rotation around the x and y axis respectively.

Since the VRP is assumed to lie on a plane where feature points lie, position error of the VRP in the image occurs and it changes as the head rotates due to the distance of D_z. When D_z is much less than C_z, which is satisfied in our system, the errors in horizontal and vertical directions denoted by Δ_u and Δ_v can be evaluated by the following equation:

$$\begin{aligned} \Delta_u &= f\frac{D_z}{C_z}\sin\theta_y \\ \Delta_v &= f\frac{D_z}{C_z}\sin\theta_x\cos\theta_y \end{aligned} \tag{3}$$

where f indicates a focal length of the camera.

The absolute values of θ_x and θ_y are determined by a length of a major and a minor axis of an ellipse that is transformed from a circle with the matrix $\boldsymbol{A}$. The sign of them is determined by the translation elements of $\boldsymbol{A}$. The amount of the errors can be calculated in this manner and the errors are modified.

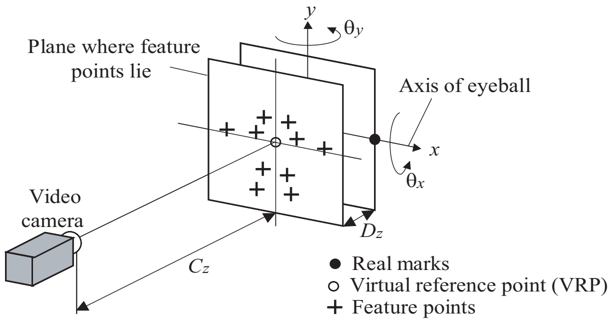


Fig. 6 Relation between the real marks and the VRP.

4.2 Tracking of Feature Points

Since the proposed eye-gaze estimation for human interface does not restrict user's motion, it is not easy to extract feature points from each video image within short time. If an affine matrix that depends on motion of a face is estimated, total time for extraction can be reduced.

In this paper we apply Condensation algorithm [7], or particle filters [8], to estimate the affine matrix. The algorithm has three steps: prediction, observation, and estimation. The first step predicts an affine matrix that expresses transformation between two successive video images. The second step observes the difference in position between actual feature points and predicted ones. The third step estimates optimal affine matrix.

Let $\boldsymbol{A}_i$ denote an affine matrix for the i-th video image, and $\boldsymbol{A}^{(j)}(j = 1, \cdots, N)$ means that this matrix is the j-th affine matrix among N number of matrices generated randomly at time i. A subscript i and a superscript j associated with other symbols have the same meaning.

Firstly $\boldsymbol{A}_i$ is roughly predicted based on the previous two affine matrices.

$$\boldsymbol{A}_i = \boldsymbol{A}_{i-1} + \Delta t_{i-1} \times \Delta \boldsymbol{A}, \quad \Delta \boldsymbol{A} = \frac{\boldsymbol{A}_{i-1} - \boldsymbol{A}_{i-2}}{\Delta t_{i-2}} \tag{4}$$

By adding a system noise $T^{(j)}$ to the matrix, N number of candidates for optimal $\boldsymbol{A}_i$ are generated.

$$\boldsymbol{A}^{(j)} = \boldsymbol{A}_i + \boldsymbol{T}^{(j)} \tag{5}$$

Here $T^{(j)}$ is a translation matrix in vertical and horizontal directions.

In the next step, positions of feature points in the i-th image are calculated using each of N number of affine matrices,

$$\hat{\boldsymbol{s}}_k{}'^{(j)} = \boldsymbol{A}^{(j)} \hat{\boldsymbol{s}}_k, \quad (k = 1, \cdots, K) \tag{6}$$

where $\hat{\boldsymbol{s}}_k$ represents the k-th feature point position in the $(i-1)$-th image, and K represents a number of feature points. After that, a weight score $W^{(j)}$ for each matrix is evaluated by comparing pixel values around $\hat{\boldsymbol{s}}_k$ and $\hat{\boldsymbol{s}}_k{}'^{(j)}$.

$$W^{(j)} = \sum_{k=1}^{K} \mathscr{N}(\,\hat{\boldsymbol{s}}_k,\, \hat{\boldsymbol{s}}_k{}'^{(j)}) \tag{7}$$

where the operator $\mathscr{N}(\ ,\ )$ represents normalized correlation having a value from -1 to $+1$.

Finally probability weight distribution, $\pi^{(j)}$, corresponding to each $\boldsymbol{A}^{(j)}$ is calculated by assuming that the distribution follows Gaussian distribution,

$$\pi^{(j)} = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(K - W^{(j)})^2}{2\sigma^2}\Big) \tag{8}$$

and $\boldsymbol{A}_i$ is estimated by the following equation.

$$\boldsymbol{A}_i = \frac{\sum_{j=1}^{N} \pi^{(j)} \boldsymbol{A}^{(j)}}{\sum_{j=1}^{N} \pi^{(j)}} \tag{9}$$

After estimating the affine matrix $\boldsymbol{A}_i$, it is modified again using a pattern matching method, so that $\boldsymbol{A}_i$ expresses transformation between the reference image and the i-th image accurately. The modification proceeds along the following steps. At first, $\hat{\boldsymbol{s}}_k$ in the reference image is transformed to $\hat{\boldsymbol{s}}_k{}'$ by $\boldsymbol{A}_i$ calculated by (9). Next, patterns around $\hat{\boldsymbol{s}}_k$ and $\hat{\boldsymbol{s}}_k{}'$ are compared with subpixel accuracy, and the feature points in the i-th image are extracted again. Finally $\boldsymbol{A}_i$ is determined with the least square method.

5 Results and Discussion

5.1 Simulation Results

A distribution of displacement of irises was examined by simulation on condition that a gaze line is fixed at a camera and a head is directed to a circumference around the camera. Fig. 7 shows the results of simulation. The distance between the camera and the head is assumed to be 2.4 meters. The angle between the head direction and an optical axis of the camera is set at 10° and 20°. Since a value of 0.01 is used as the threshold value for the eye-gaze estimation, the maximum amount of displacement using the VRP is sufficiently small.

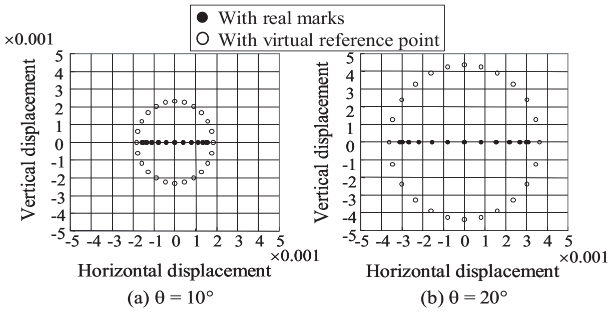


Fig. 7 Distribution of the displacement of irises by simulation.

5.2 Experimental Results

The experiment was conducted attaching real marks on a subject's face so as to compare the amount of displacement of irises with the VRP to that with the real marks. The distance from the subject to the camera was 2.4 meters. The subject continued to gaze at the camera lens during the experiment, but it directed its face in every direction.

Fig. 8 shows the change of displacement in successive images. A horizontal axis of the figure indicates the frame number of the image. The sampling interval was about 30 msec. The motion patterns of the face during the first and second period were the same. Elapsed time of the period was about 3 sec and 1.5 sec respectively.

Fig. 8 demonstrates that the position of the VRP was well modified, and the result of the eye-gaze estimation with the VRP was almost the same as with the real marks in the first period. Since the face was directed downward during the frame no.70 to 95, the system failed in extraction of outer corners of the eye. Because rotational speed of the head was too high in the second period, each image was blurred slightly, and pattern matching for extracting feature points was not done correctly. The reference image and the video images picked out every 30 frames are shown in Fig. 9.

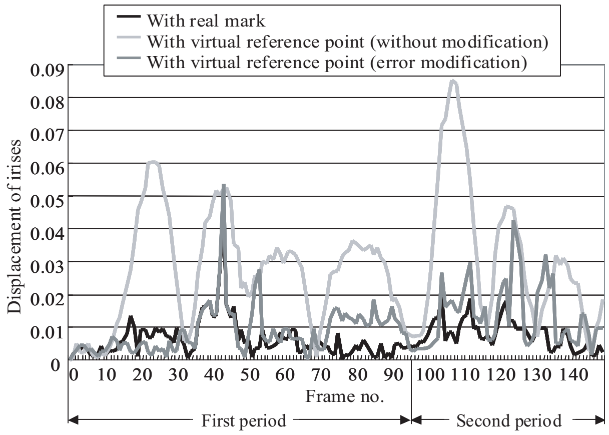


Fig. 8 Experimental results of real-time eye-gaze estimation.

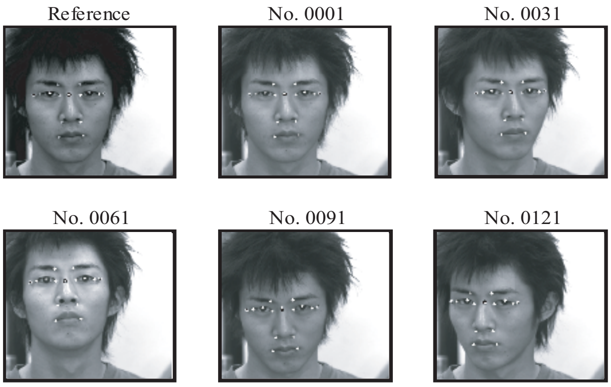


Fig. 9 The reference image and the video images picked out every 30 frames.

6 Conclusion

The real-time eye-gaze estimation by using the virtual reference point(VRP) was proposed, and the validity of the method was demonstrated. When a person gazes at an object under ordinary condition, its head stays still and an angle between a gaze direction and a face direction is not big. In such a case, the percentage of correct estimation reached almost 100% by our previous study.

Since the method neither restricts user's motion, nor uses any special light source or a hardware device, and laborious camera calibration is needless, it is applicable in various situations and at various places. For example, eye-gaze input can turn on and off household electric appliances. It can be useful for any one including handicapped persons.

Future study will include the eye-gaze estimation with multiple cameras so as to detect an arbitrary view direction.

Acknowledgements

This work was partially supported by The 21st Century COE Program ”Intelligent Human Sensing.

References

[1] T. H. Hutchinson, K. P. White, JR., W. N. Martin, K. C. Reichert, L. A. Frey. ”Human-Computer Interaction Using Eye-Gaze Input,” In: *IEEE Trans. on System Man and Cybernetics*, vol. 19, no. 6, pp.1527-1533, 1989.

[2] S. Pastoor, J. Liu, S. Renault, ”An Experimental Multimedia system Allowing 3-D Visualization and Eye-Controlled Interaction Without User-Worn Device,” *IEEE Transactions on Multimedia*, vol. 1, no. 1, pp.41-52, 1999.

[3] Y. Matsumoto and A. Zelinsky, ”An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement,” *Proc. of IEEE Fourth International Conference on Face and Gesture Recognition (FG'2000)*, pp.499-505, March, 2000.

[4] J. G. Wang, E. Sung, V. Ronda, ”On Eye gaze determination via Iris contour,” *Proc. of MVA 2000, IAPR Workshop on Machine Vision Application*, pp.363-366, November, 2000.

[5] T. Miyake, S. Haruta, S. Horihata, ”Image Based Eye-Gaze Estimation Irrespective of Head Direction,” *ISIE2002 IEEE International Symposium on Industrial Electronics*, pp.332-336, 2002.

[6] T. Miyake, S. Haruta, S. Horihata, ”Eye-Gaze Estimation by Using Features Irrespective of Face Direction,” *Journal of Systems and Computers in Japan*, vol. 36, Issue 3, pp.18-23, 2005.

[7] M. Isard and A. Blake, ”Condensation - Conditional Density Propagation for Visual Tracking”, *International Journal of Computer Vision*, vol. 29, no. 1, pp.5-28, 1998.

[8] Kenji Oka, Yoichi Sato, Yasuto Nakanishi and Hideki Koike, ”Real-time Head Pose Estimation Based on Particle Filtering”, in Japanese, *Proc. of the Symposium on 21st Century COE Program, the global dependable information infrastructure project, The information science and technology Strategic Core, February, 2004.*