Detecting a Gazing Region by Visual Direction and Stereo Cameras

Akihiro Sugimoto* Akihiro Nakayama Takashi Matsuyama
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Kyoto 606–8501, Japan
sugimoto@nii.ac.jp

Abstract

We develop a wearable vision system that consists of a user's visual direction sensor and stereo cameras. First, we establish a method for calibrating the system so that it can detect user's blink points even in a real situation such that the depth of blink points changes. Next, we propose a method for detecting a gazing region of a user in terms of the planar convex polygon. In our method, the system first identifies the fixation point of a user, and then applies a stereo algorithm and robust statistics to detect his gazing region. Now the system can detect the gazing region of a user and provide him with its 3D position.

1. Introduction

With the rapid progress of computer facility, computer usage in every aspect of our daily life has become more and more popular; wearing the computer in our everyday life is becoming tangible to reality. Thus, a tremendous amount of effort has been made to establish technologies for realizing the wearable computer (see [1, 3, 4] for example).

The current approach for human computer interaction such as graphical user interface is, on the other hand, based on the concept that the computer is used as a terminal of the computer network. The user, therefore, has to explicitly manipulate objects on a computer monitor to interact with the computer. In usage of the wearable computer, however, the computer itself should first understand the user's situation, intention as well as activities, and then provide in good time the user with useful information at that time. Namely, context-aware human computer interaction is required.

A device sensing information in the scene nearby a user is indispensable to the computer for understanding the user's situation. In particular, the camera is most promising because of two reasons. One is the amount of acquired information and the other is sharing the field of view with the user. The visual direction of the user results in strongly reflecting his interest or attention regardless of his consciousness.

The above observation motivated us to develop a wearable vision system that consists of a user's visual direction sensor and stereo cameras. Our system detects the region and its depth information gazed by a user. More precisely, the system detects the visual direction of a user, and then recognizes whether or not the user is gazing at a region. Once the system recognizes that the user is, it detects the region in terms of the planar convex polygon and, at the same time, reconstructs the depth map of the region. Now the system can detect the gazing region of a user and provide him with its 3D position. Note that we assume that the scene nearby a user is static in this paper.

2. Overview of our system

2.1. System configuration

Our wearable vision system consists of the head part and the computer (Fig. 1). The head part (Fig. 2) has a user's visual direction sensor and two cameras. The projection centers of the two cameras are designed to be aligned with the centers of the user's eyeballs. Fig. 3 shows a user with the head part of our system. Eye-mark recorder EMR-8 from NAC Image Technology is employed as the user's visual direction sensor. EMR-8 uses the pupil-corneal reflection method in eye tracking and overlays user's blink points, i.e., the points in 3D at which the user looks while blinking, onto the image captured by the right camera. This overlaid point is called an "eye mark". The sampling rate of eye marks by EMR-8 is 60Hz (about 17ms). We also employ the off-the-shelf camera, EVI-G20, produced by Sony and a PC with Pentium III 750MHz.

The eye mark detected by EMR-8 and two images captured by the stereo cameras are all put into the computer. The computer then identifies the region at which the user's right eye is gazing, and reconstructs the depth map of the region.

*Presently with National Institute of Informatics, Chiyoda-ku, Tokyo 101-8430, Japan.

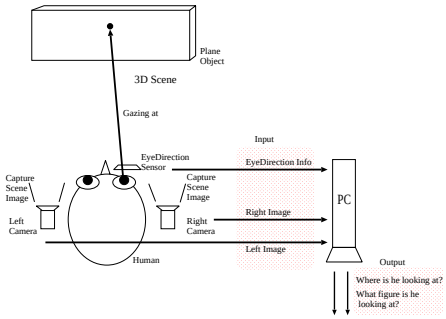


Figure 1: Wearable vision system we developed.

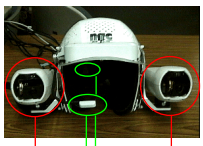


Figure 2: Head part.

Figure 3: A user with the head part.

2.2. System calibration

We have to set some coordinate systems for analysing the visual direction of a user. They are the right-camera coordinates, the left-camera coordinates, the user-centered coordinates and the visual direction angle coordinates with respect to the user's right eyeball (Fig. 4).

We introduce the camera coordinate system to each camera where the projection center of the camera is identical with the origin. For reconstructing the 3D position of a point of interest, we calibrate the intrinsic and extrinsic camera parameters in advance and then employ stereo vision technique. In this paper, we employ the method proposed by Zhang [8] to calibrate the camera parameters. We certified that the optical axes of the two cameras are almost parallel with each other and that the poses are identical.

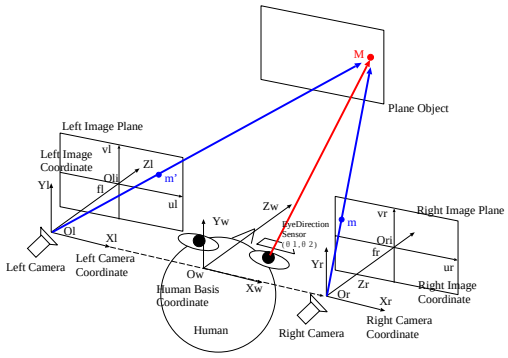


Figure 4: Introduced coordinate systems.

The origin of the user-centered coordinates is set to the middle point between the origins of the two camera coordinate systems. As a result, the origin is almost identical with the middle point between the centers of the user's two eyeballs. The pose of the user-centered coordinates, on the other hand, is set to be identical with that of the two camera coordinate systems.

We set the rotation center of the user's right eyeball as the origin of the visual direction angle coordinates. In this coordinate system, the coordinates represent rotation angles, pan and tilt angles, with respect to the optical axis of the right-camera coordinate system. EMR-8 measures the rotation angles of the user's right eyeball in terms of the coordinates in this visual direction angle coordinate system. In this measurement, the pupil-corneal reflection method is employed where the cornea is illuminated by an infrared light and the light reflected back from the cornea is then captured to estimate the direction of the cornea.

To observe user's blink points we overlay them onto the image captured by the right-camera. For this purpose, we have to calibrate the relative position and pose between the right-camera coordinates and the visual direction angle coordinates. The algorithm for this calibration is provided with EMR-8. Namely, a user gazes at nine points on a plane (called a calibration plane) one by one in a given order, and then the nine pairs of visual direction angles and the images of the points are used to calibrate the two coordinates. This algorithm allows the system to overlay the user's blink point onto the image captured by the right-camera.

Unfortunately, however, EMR-8 assumes in its usage that the distance between a calibration plane and a user is not large and that the user always keeps his blink points on the plane. These assumptions cause the problem that the overlaid eye marks do not accurately reflect the user's blink points in the image for the case where the distance between the user and his blink point dynamically changes; this case always occurs in our daily life.

For example, we consider the case where user's blink points are farther than a calibration plane (Fig. 5). Let M be the blink point of a user. EMR-8 then identifies M' on the calibration plane as the blink point of the user and overlays its image m' onto the right-camera image as the user's eye mark at that time. As we see, this overlay is incorrect because the image m of M should be overlaid. The horizontal $(x-)$ component δ of the residual of m' from m follows from Fig. 5:

$$\delta = wf\left(\frac{1}{D_c} - \frac{1}{D_o}\right), \tag{2.1}$$

where f and w respectively denote the focal length of the right camera and the horizontal component of the distance

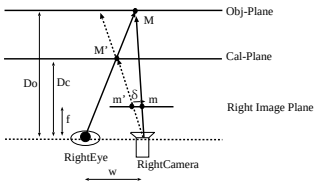


Figure 5: Blink point farther than the calibration plane.

between the rotation center of the user's right eyeball and the projection center of the right camera. D_{c} and D_{o} are the distance of the calibration plane and the blink point from the projection center of the right camera, respectively. We remark that the vertical $(y-)$ components of the residual can be also derived in the same way. For the case where user's blink points are nearer than the calibration plane, the residual is represented in the similar equation as (2.1).

To correct the point to be overlaid, we have to know w, f, D_{c} and D_{o} in advance, and then compute δ. We can measure w, f and D_{c} since we can calibrate them beforehand. D_{o}, on the other hand, can be computed by a stereo algorithm since two calibrated cameras are mounted on our system. Accordingly, we can correct the residual δ, and this correction enables the system to correctly overlay the blink point onto the image even though the distance between the user and his blink point dynamically changes.

3. Detecting the gazing region of a user

For detecting the gazing region of a user, i.e., the region at which a user is gazing while he is fixating his blink points, the system requires two functions. Firstly, the system has to identify whether or not a user is fixating his blink points. Secondly, once the system identifies the user's fixation point, it has to identify the region at which the user is gazing.

3.1. Fixation detection

Our eyeballs are reported to always repeat two kinds of motion: fixation (or gaze-holding) and saccade (or gaze-shifting) [2]. Blink points are fixated during fixation to keep the retinal image of an gazing object still. During saccade, on the other hand, blink points are shifted to capture the image of an object in the center of the retina. It is also reported that fixation remains for about 250ms on the average while saccade does for about 10ms.

During fixation, the retinal image is clearly formed, and, therefore, we regard that the user is gazing at an object at that time. We remark that the visual angle during saccade has the following properties [2]:

- 99% of the angles are less than 15 degrees.
- Most of the angles are between 3 degrees and 6 degrees.
- Saccade occurs with about 2 degrees even while we are gazing, and it is about 7 or 8 degrees while we are looking around.

We thus have to identify the time interval, 250ms on the average, when fixation occurs from obtained time-series images of user's blink points. We focus on the displacement of images of blink points and then identify whether or not the user's blink points stably remain still. If the time interval when blink points remain still is more than 250ms, we identify that the user is fixating his blink points at that time: fixation occurs. The fixation point is computed as the average of the blink points over the time interval.

The displacement from which we identify whether fixation occurs is evaluated in terms of the angular field of view with respect to the camera, which corresponds to the visual angle of human beings. To be more precise, if the displacement of images of blink points is less than 0.5 degree[1] in terms of the angular field of view with respect to the camera, we regard that his blink points remain still.

The above procedures enable the system to identify whether or not a user is fixating his blink points from time-series images of user's blink points and to compute the fixation point of the user if fixation occurs. The fixation point is of course overlaid onto the right-camera image.

3.2. Gazing-region detection

The functional field of view is used to distinguish cognitive measures of the visual field sensitivity from the more clinical sensory measures. It is defined as the spatial area or visual field extent that is needed for a specific task [7]. The visual angle of the functional field of view has the following properties:

- At least about 10 degrees are required for reading a book.
- The area for which we can process information in detail is less than 2.5 degrees.

We regard that the functional field of view is closely related with the gazing region of a user. Though the functional field of view depends on a task in general, we set it in this paper to be 10 degrees in terms of the visual angle. The gazing region of a user is, therefore, set to be 10 degrees in terms of the angular field of view with respect to the camera. The size of the gazing region itself changes depending on the depth from the user as shown in Fig. 6.

[1] We settle down this value due to two reasons. One is that saccade is at least about 2 degrees in terms of the visual angle. The other is that we observed that about ± 0.5-degree errors are involved in the detection of eye marks by EMR-8 even when the assumptions described above are satisfied in its usage.

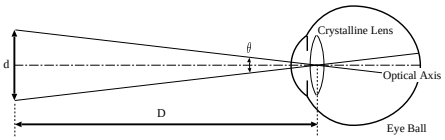


Figure 6: Size and depth of a gazing region.

We consider the case where a user is gazing at some regions one by one in the scene. If we allows user's motion in that case, two movements, the movement of the fixation point and the movement of the user's head, are merged, and discriminating one from the other is a hard problem, which results in difficulty in accurately detecting gazing regions of the user. In our daily life, on the other hand, it is popular that meaningful information is involved in planar convex polygons such as figures, signposts or guideposts. We therefore introduce two assumptions below in this paper.

- Gazing regions of a user are planar and convex polygons.
- A user does not move his head while the system measures his visual directions.

Under these assumptions, we apply the methods developed in the computer vision literature to detect the 3D shapes of user's gazing regions. To be more concrete, we detect a planar convex polygon within the functional field of view whose center is a user's fixation point.

In the previous section, we have obtained the image of a user's fixation point in the right-camera image. We now apply the template matching to reconstruct the depth of each pixel inside the functional field of view whose center is the image of the fixation point. As a result, we obtain the depth map of points within the functional field of view during fixation.

Since we assume that a gazing region is planar, we apply the plane fitting to the reconstructed points. In the plane fitting, we employ the robust statistics rather than the method of least squares. This is because the reconstructed points do not necessarily exist on a plane. We remark that the gazing region exists as a part in the functional field of view and therefore some points at which the user is not gazing may also exist in the functional field of view. For the set of the reconstructed point (X_i, Y_i, Z_i), we apply the method of least median of squares [6] to obtain plane parameters (a_1, a_2, a_3) satisfying

$$\min_{a_1,a_2,a_3} \operatorname{med} (1 - a_1 X_i - a_2 Y_i - a_3 Z_i)^2, \qquad (3.2)$$

from which we identify the plane representing the gazing region.

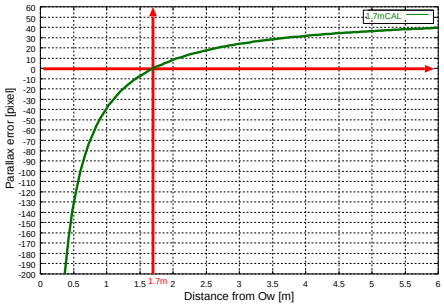


Figure 7: Calibration curve ($w = 114$mm).

To obtain the gazing region, we select points on the plane among the reconstructed points, and then apply the incremental method [5] to obtain the convex hull of the selected points. Finally, we project the hull onto the right-camera image to represent the user's gazing region in terms of the convex polygon.

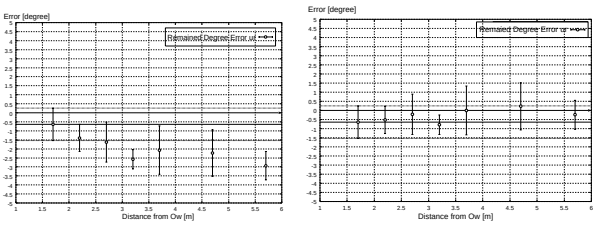
4. Experiments

4.1. Precision evaluation of detected blink points

We evaluated the effect of our correction to the detected eye marks by EMR-8 described in Section 2.2 under the condition that the depth of user's blinks point changes.

We first set a calibration plane whose distance is 1.7m from a user, and then made the user gaze at a set of nine points on the plane one by one to calibrate EMR-8. Next we moved the plane so that the distance from the user changes by 0.5m from 1.7m to 5.7m in turn. At each distance, we made the user gaze at another set of nine points and obtained the coordinates in the right-camera image of the eye marks detected by EMR-8. In fact, we used the average of the coordinates of the stably detected eye marks. We then applied our correction described in (2.1) to the eye marks to obtain the corrected coordinates of images of the blink points. We remark here that we carefully measured $D_{\rm c}$ and w to obtain $D_{\rm c} = 1.7$m and $w = 114$mm. We therefore had the calibration curve shown in Fig. 7.

For the cases with/without our correction, we computed the average and the variance of the residuals over the given set of nine points, and compared the two cases (Fig. 8). Note that error bars in Fig. 8 represent the standard deviation. Fig. 8 shows that the residuals of corrected coordinates almost stably remain small independent of the change in distance of the plane from the user. In fact, they are within perturbation of the standard deviation from the average for the distance 1.7m at which EMR-8 was calibrated. This observation indicates that our correction is valid and effective.



(a) not corrected (b) corrected

Figure 8: Errors of detected blink points.

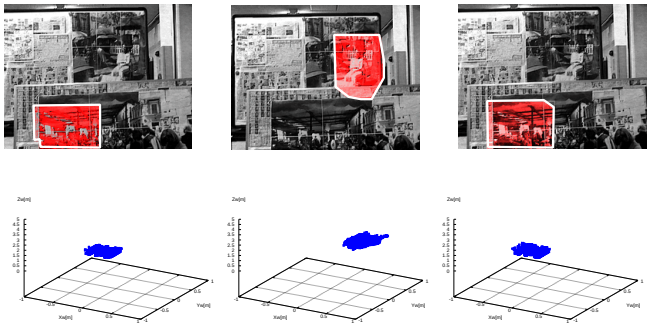


Figure 9: Detected gazing regions and their depth.

4.2. Detection of gazing regions

We also experimented on detecting user's gazing regions in the scene. To be more concrete, we set two objects at different distances from a user: one was 1.7m far from the user and the other was 2.2m. The user freely looked around at the scene for a while but had intention that he sometimes gazes at some regions of the objects within the time. The system then identified the time interval in which he was gazing and estimated his gazing regions in the scene and reconstructed their depth.

The size of images captured by the two cameras was 640×480 pixels. To obtain the correspondence of pixels between the right-camera image and the left-camera image, we applied the template matching in which similarity was evaluated by the sum of the intensity difference between the pixels over the template. Note that the size of the employed template was 10×10 pixels. When fixation was recognized, the computational time in detecting and visualizing the estimated gazing region was about 1.2s that includes the time for correcting the coordinates of the images of blink points.

Figure 9 shows a few example images obtained in this experiment. We obtained the response from the users that the gazing regions are almost precisely detected. We see that the estimated gazing region has a nonplanar part in the middle case of Fig. 9. This is because the false of the template matching caused wrong reconstruction of the depth. Fig. 9 also shows the depth of the gazing regions are almost precisely reconstructed depending on their depth.

5. Concluding remarks

We presented, in this paper, the wearable vision system we developed for context-aware human computer interaction in usage of the wearable computer. The presented system consists of a user's visual direction sensor and stereo cameras. Our system recognizes whether or not a user is fixating his blink points, and once the system recognizes that the user is, it detects his gazing region in terms of the planar convex polygon and, at the same time, reconstructs the depth map of the region. Our system can be brought into use for understanding human intention and activities in our daily life.

Reducing the computational time in detection and visualization of the gazing region is a next step. We also plan to eliminate the assumption that the user's head does not move and to estimate user's positions and fixation points while the user freely moves around the 3D space.

Acknowledgements

This work was supported by Grants-in Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contraction of 13224051 and 14380161.

References

[1] H. Aoki, B. Schiele and A. Pentland: Realtime Personal Positioning System for Wearable Computers, Vision and Modeling Technical Report, TR-520, Media Lab. MIT, 2000.

[2] R. Carpenter: *Movements of the Eyes*, 2nd ed., Pion, London, 1988.

[3] B. Clarkson, K. Mase and A. Pentland: *Recognizing User's Context from Wearable Sensors: Baseline System*, Vision and Modeling Technical Report, TR-519, Media Lab. MIT, 2000.

[4] M. Kourogi, T. Kurata and K. Sakaue: A Panorama-Based Method of Personal Positioning And Orientation And Its Real-Time Applications for Wearable Computers, *Proc. of Int. Symposium on Wearable Computers*, Switzerland, pp.107-114, 2001.

[5] F. P. Preparata and M. I. Shamos: *Computational Geometry: An Introduction*, Springer-Verlag, 1985.

[6] P. J. Rousseeuw: Least Median of Squares Regression, *J. American Stat. Assoc.*, Vol. 79, pp. 871–880 (1984).

[7] A. F. Sanders: *The Selective Progress in the Functional Field of View*, Van Gorcum & Comp., N. V., Amsterdam, 1964.

[8] Z. Zhang: A Flexible New Technique for Camera Calibration, *IEEE Transactions on PAMI*, Vol. 22, No. 11, pp. 1330–1334 (2000).