Proceedings of the 2003 IEEE/ASME
International Conference on
Advanced Intelligent Mechatronics (AIM 2003)

Active Wearable Vision Sensor
—Detecting Person's Blink Points and Estimating Human Motion Trajectory—

Akihiro Sugimoto[†] and Takashi Matsuyama[‡]
[†]National Institute of Informatics, Tokyo 101-8430, Japan
[‡]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
sugimoto@nii.ac.jp

Abstract

To realize versatile real-time man-machine interactions based on understanding human intention and activities, we develop an active wearable vision sensor. The sensor consists of the detector of person's viewing lines and two active cameras. First, we establish a method for calibrating the sensor so that it can detect person's blink points accurately even in a real situation such that the depth of blink points changes. Secondly, we propose a method, the binocular independent fixation control, for incrementally estimating the motion trajectory of a person wearing the sensor.

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 efforts has been made to establish technologies for realizing the wearable computer (see [1, 2, 6, 7, 8] for example).

The current approach for the interactions between human beings and computers such as GUI (Graphical User Interface) is, on the other hand, based on the concept that the computer is a tool to enhance our capabilities or activities. This kind of our interactions with the computer can be regarded as so-called a *master-servant interaction model*. The user there has to explicitly manipulate objects on a computer monitor to interact with the computer, and the computer is just a tool that gives us no response without any order from us. Though multi-modal interface [11] and PUI (Perceptual User Interface) [12, 15] have been proposed for usage of the wearable computer, such interfaces are also based on the master-servant interaction model. In fact, in their context more flexible and simpler interfaces for us to "use" the computer are being studied.

This current concept of the relationship between human beings and computers should change for the next generation way of getting along with computers in our every day life. We should introduce a new interaction model, so-called a *man-machine symbiotic interaction model*, between human beings and computers. In this interaction model, the computer has its own identity and exists as a partner of us. That is, not only the computer gives us responses based on our orders but also it itself autonomously understands our situation, intention or activities, and then provides us in good time with useful information at that time.

The above observation motivated us to develop a wearable vision sensor [14]. Our sensor consists of the detector of person's viewing lines and two active cameras. With the cameras that have the common field of view with a person wearing this sensor, the computer can detect the viewing lines of the person. First, we establish a method for calibrating the sensor so that it can detect person's blink points accurately even in a real situation such that the depth of blink points changes. We formulate errors of the viewing-line detector in terms of the depth in blink points from the person and employ the stereo algorithm to correct the errors. Secondly, we propose a method for incrementally estimating the motion trajectory of a person wearing the sensor. In our method, we propose the binocular independent fixation control. That is, while the person moves we control the two active cameras independently so that each automatically fixates its optical axis to its own fixation point.

2 ACTIVE WEARABLE VISION SENSOR

A device sensing information in the scene nearby a person is indispensable to the computer for understanding his/her situation, intention and activities. In particular, the camera is most promising because of two reasons. One is the amount of acquired information and the other is the capability of having the common field of view with a person. It is also quite natural to regard that the viewing lines of a person result in strongly reflecting his/her interest or attention regardless of his/her consciousness [3, 13].

Why wearable: We may take an alternative approach to understand person's activities where we embed in the surrounding environment multiple sensors such as cameras or magnetic sensors, and process information acquired by

0-7803-7759-1/03/$17.00 © 2003 IEEE

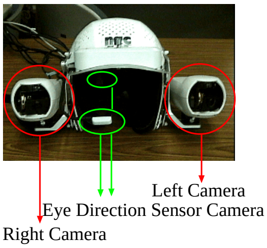


Figure 1: Head part.

Figure 2: A person with the head part.

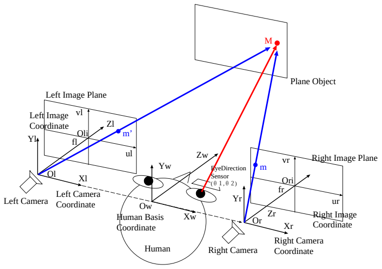


Figure 3: Introduced coordinate systems.

them. Information acquired by the sensors embedded in the surrounding environment, i.e., information through an *objective* point of view, however, is not satisfactory from the point of view that we capture the intension and interest of the person moving in the environment. Information through the person's viewpoint, i.e., information through a *subjective* point of view, is necessary for such tasks. This can be supported by our experience that we often feel difficulty in communicating our intention to a person who is in a spatially different place. From the point of view that we understand human intention and activities, sharing the common field of view with a person and sharing common inputs with the person are required. The wearable vision sensor satisfies these two requirements.

Why active: If the camera is active, namely, we can control the optical axis of the camera through a computer, the function of acquiring information is highly enhanced: the computer can control the camera to autonomously acquire information of the environment independent of person's viewing lines. In other words, depending on the situation, the computer can switch two kinds of functions: (1) acquiring *subjective information* by sharing the common field of view with a person and (2) acquiring *objective information*, i.e., autonomously acquiring information of the environment independent of person's viewing lines. This is the great advantage of acquiring information that cannot be realized with a camera whose optical axis is fixed.

3 SENSOR CONFIGURATION

Our active wearable vision sensor consists of the head part and the computer. The head part (Fig. 1) has two active cameras and a detector of person's viewing lines. The projection centers of the two cameras are designed to be aligned with the centers of the person's eyeballs. The computer, on the other hand, is a PC with Pentium III 750MHz and 1GB memory. Fig. 2 shows a person with the head part of our sensor.

Eye-mark recorder EMR-8 from NAC Image Technology is employed as the detector of person's viewing lines. EMR-8 uses the pupil-corneal reflection method in eye tracking and overlays *person's blink points*, i.e., the points in 3D at which a person 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). As an active camera, we employed the off-the-shelf camera, EVI-G20, produced by Sony. EVI-G20 accepts commands from a computer to rotate its optical axis by the pan (within $\pm 30^\circ$) and the tilt (within $\pm 15^\circ$). It is also designed so that the projection center of the camera is identical with the rotation center of the camera body.

The viewing line of a person detected by EMR-8 and two images captured by the two cameras are all put into the computer. The blink point of the person's right eye is superimposed as the eye mark on the right-camera image.

4 DETECTION OF PERSON'S BLINK POINTS

4.1 Introduction of coordinate systems

We have to set some coordinate systems for analyzing the relationship between the viewing line of a person and his/her blink point. They are the right-camera coordinates, the left-camera coordinates, and the viewing-line angle coordinates with respect to the person's right eyeball (Fig. 3).

We introduce the camera coordinate system to each camera where the projection center of the camera is identical with the origin. To reconstruct the depth 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 [16] to calibrate the camera parameters. We verified that the optical axes of the two cameras are almost parallel with each other and that the poses are identical.

We set the rotation center of the person's right eyeball as the origin of the viewing-line 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 person's right eyeball in terms of the coordinates in this viewing-line angle coordinate system.

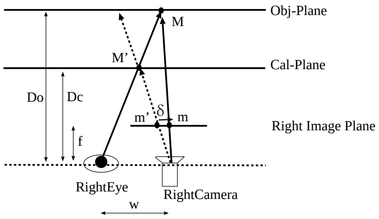


Figure 4: Blink point farther than the calibration plane.

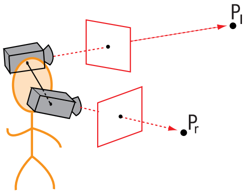


Figure 5: Binocular independent fixation control.

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.

4.2 Sensor calibration

To observe person'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 viewing-line angle coordinates. The algorithm for this calibration is provided with EMR-8. Namely, a person gazes at given nine points on a plane (called a calibration plane) one by one in a given order, and then the nine pairs of viewing-line angles and the images of the points are used to calibrate the two coordinates. This algorithm allows the system to overlay a person'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 person is not large and that the person always keeps his blink points on the calibration plane. These assumptions cause the problem that eye marks do not accurately reflect person's blink points in the image for the case where the distance between the person and his blink points dynamically changes; this case always occurs in our daily life.

For example, we consider the case where person's blink points are farther than a calibration plane (Fig. 4). Let M be the blink point of a person. EMR-8 then (incorrectly) identifies M' on the calibration plane as the blink point of the person and overlays its image m' onto the right-camera image as the person's eye mark at that time. As seen above, 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. 4:

$$\delta = wf\left(\frac{1}{D_{\rm c}} - \frac{1}{D_{\rm o}}\right), \tag{4.1}$$

where f and w respectively denote the focal length of the right camera and the horizontal component of the distance between the rotation center of the person's right eyeball and the projection center of the right camera. $D_{\rm c}$ and $D_{\rm o}$ are the distance of the calibration plane and the blink point from the projection center of the right camera, respectively. This is the formulation of errors of detected eye marks in terms of the depth, $D_{\rm o}$, in blink points. We remark that the vertical $(y-)$ components of the residual can be also derived in the same way. For the case where person's blink points are nearer than the calibration plane, the residual is represented in the same equation as (4.1).

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

5 ESTIMATION OF HUMAN-MOTION TRAJECTORY BY BINOCULAR INDEPENDENT FIXATION CONTROL

In the previous section, to detect his/her blink points two cameras shared the common field of view with a person wearing the cameras: we used the active wearable vision sensor in the context of acquiring subjective information. In this section, by contrast, the sensor is used for acquiring objective information: the field of view of the cameras is independent of the person's.

To estimate the human-motion trajectory with two active wearable cameras, we introduce the *fixation control*, i.e., the camera control in which the camera automatically fixates its optical axis to a selected point (called the *fixation point*) in 3D, and apply the fixation control independently to each active camera. That is, while the person moves, we control the two active cameras independently so that each automatically fixates its optical axis to its own fixation point. We call this camera control the *binocular independent fixation control* (Fig. 5).

5.1 Binocular independent fixation control vs. stereo vision framework

In the robotics literatures, the framework of stereo vision is widely used to estimate the position and motion of a moving robot [4, 9, 10]. When we employ the stereo vision algorithm, however, we have to make two cameras share the common field of view and, moreover, establish feature correspondences between the images captured by the two cameras. This kind of processing has difficulty in its stability. In addition, we have another problem in using the stereo vision framework in the context of wearable cameras. Namely, though the accuracy of the estimation is well known to highly depend on the baseline of the two cameras, keeping the baseline of two cameras wide is hard when we wear cameras. Therefore, the estimation accuracy of the motion trajectory is limited if we employ the stereo vision algorithm.

In the binocular independent fixation control, on the other hand, the two cameras need not share the common field of view because each camera fixates its optical axis to its own fixation point in 3D. We do not face the problem of feature correspondences between the images captured by two cameras. Moreover, the estimation accuracy becomes independent of the baseline of two cameras. This can be understood as follows. If we assume that we set a camera at each fixation point and that the optical axis of each camera is toward a person, then the binocular independent fixation control can be regarded as the situation where we apply the stereo vision framework to estimating the position of the person from the two fixation points. The baseline in this case is identical with the distance of the two fixation points. This means that the estimation accuracy is independent of the baseline of the two cameras that the person wears; selecting fixation points as far as possible from each other allows the estimation accuracy to become high.

5.2 Constraint derivation on human motion

We set the right camera is base and assume that the motion of a person wearing the cameras is identical with the motion of the base-camera coordinates. Moreover, for simplicity, we assume that the orientation of the camera coordinates does not change even though we change pan and tilt of the camera for the fixation control. This means that only the human motion causes changes in orientation and translation of the camera coordinates. We then develop a method for estimating a human motion below.

We assume that the translation vector and the rotation matrix to make the left-camera coordinates identical with the right-camera coordinates are $\boldsymbol{T}_{\rm in}$ in the left-camera coordinates and $R_{\rm in}$ in the right-camera coordinates, respectively. $\boldsymbol{T}_{\rm in}$ and $R_{\rm in}$ are both assumed to be known.

5.2.1 Constraints from fixation correspondence

The fixation control gives us the correspondence of the viewing lines of a camera toward the fixation point over time-series frames. We call this correspondence a *fixation correspondence*.

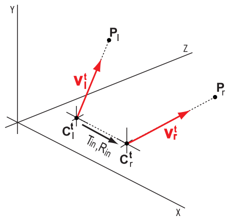


Figure 6: Relationship between the projection centers and the fixation points at time t.

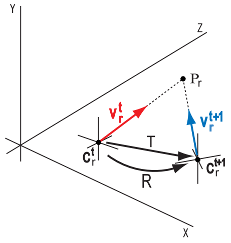


Figure 7: Geometry based on the fixation correspondence of the right camera.

Let the projection centers of the left camera and the right camera be C_ℓ^t and $C_{\rm r}^t$ in 3D at time t. We assume that the both cameras have their own fixation points P_ℓ and $P_{\rm r}$. We denote by $\boldsymbol{v}_{\rm r}^t$ the unit vector from $C_{\rm r}^t$ to $P_{\rm r}$ in the right-camera coordinates at time t. We also denote by $\boldsymbol{v}_\ell^t$ the unit vector from C_ℓ^t to P_ℓ in the left-camera coordinates at time t (Fig. 6).

We first focus on the right camera. We assume that the projection center of the right camera moves from $C_{\rm r}^t$ to $C_{\rm r}^{t+1}$ in 3D due to the human motion from time t to $t+1$ (Fig. 7). We also assume that the rotation and the translation of the right camera incurred by the human motion are expressed as R in the right-camera coordinates at time t and $\boldsymbol{T}$ in the world coordinates. We remark that the orientation of the world coordinates is assumed to be obtained by applying rotation matrix R_0^{-1} to the orientation of the right-camera coordinates at time t.

It follows from the fixation correspondence of the right camera that

$$\lambda R_0 \boldsymbol{v}_{\rm r}^t \quad = \quad \lambda' R_0 R \boldsymbol{v}_{\rm r}^{t+1} + \boldsymbol{T},$$

where λ and λ' are non-zero constants. This equation is rewritten by

$$\det \left[\begin{array}{ccc} R_0 \boldsymbol{v}_{\rm r}^t & R_0 R \boldsymbol{v}_{\rm r}^{t+1} & \boldsymbol{T} \end{array} \right] \quad = \quad 0, \qquad (5.1)$$

which gives the constraint on the human motion derived from the fixation correspondence of the right camera. We

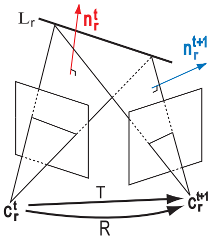


Figure 8: Geometry based on the line correspondence of the right camera.

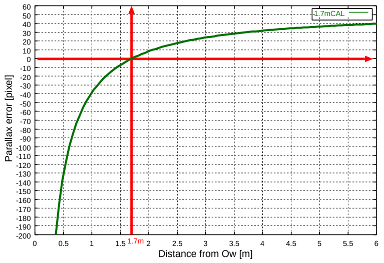


Figure 9: Calibration curve ($w = 114$mm).

see that this constraint is homogeneous quadratic on the unknowns, i.e., R and $\boldsymbol{T}$.

Since we calibrate the two cameras in advance, we can represent $\boldsymbol{v}_\ell^t$ and $\boldsymbol{v}_\ell^{t+1}$ in the right-camera coordinates. We therefore obtain the constraint similar to (5.1) from the the fixation correspondence of the left camera.

5.2.2 *Constraints from line correspondence*

To derive sufficient constraints to estimate the human motion, we employ lines nearby the fixation point. This is because (i) we find many lines in the indoor scene, (ii) we can easily and accurately detect lines with less computation by using the Hough transformation, and (iii) we can easily establish line correspondences over time-series frames due to their spatial extents.

We first focus on the right camera. We assume that we establish the correspondence of images of line L_r in 3D over time t and $t+1$, where line L_r is selected nearby the fixation point of the right camera. Line L_r is called a *focused line* in this paper. We denote by $\boldsymbol{L}_\mathrm{r}$ the unit direction vector of the focused line L_r in the world coordinates[1] . Observing a line in 3D is identical to determining the plane in 3D on which both the projection center at the observation time and the line exist. We thus obtain the unit normal vector of the plane. For the focused line L_r, this unit vector in the right-camera coordinates at time t is denoted by $\boldsymbol{n}_\mathrm{r}^t$ (Fig. 8).

From the relationship of the orientations among the world coordinates, the right-camera coordinates at time t and the right-camera coordinates at time $t+1$, we see that $\boldsymbol{n}_\mathrm{r}^t$ and $\boldsymbol{n}_\mathrm{r}^{t+1}$ are expressed as $R_0\boldsymbol{n}_\mathrm{r}^t$ and $R_0R\boldsymbol{n}_\mathrm{r}^{t+1}$ in the world coordinates. Since $R_0\boldsymbol{n}_\mathrm{r}^t$ and $\boldsymbol{L}_\mathrm{r}$ are orthogonal, and $R_0R\boldsymbol{n}_\mathrm{r}^{t+1}$ and $\boldsymbol{L}_\mathrm{r}$ are also orthogonal, we obtain the following constraint on the human motion from the line correspondence over two frames captured by the right camera:

$$\mu_\mathrm{r}\boldsymbol{L}_\mathrm{r} \quad = \quad (R_0\boldsymbol{n}_\mathrm{r}^t) \times (R_0R\boldsymbol{n}_\mathrm{r}^{t+1}), \qquad (5.2)$$

where μ_r is a non-zero constant and depends on the focused line. We see that this constraint is linear homogeneous with respect to the unknowns, i.e., R and the non-zero constant.

In the similar way, we obtain the constraint on the camera motion derived from the line correspondence of the left camera.

5.2.3 *Estimation of rotation and translation*

The constraints derived from line correspondences depend only on the rotation of the human motion. We can thus divide the human motion estimation into two steps: the rotation estimation and the translation estimation.

The first step is the rotation estimation of the camera motion. We suppose that we have correspondences of n focused lines over two time-series frames. Then, we have $n+3$ unknowns (n are from scale factors and 3 are from rotation) whereas we have $3n$ constraints in this case. Therefore, we can estimate the rotation of the camera motion if we have correspondences of more than two focused lines.

When we finish estimating the rotation of the camera motion, unknowns are only the translation factors. The constraint derived from the fixation correspondence thus becomes homogeneous linear with respect to the unknowns. Hence, we can obtain the translation of the camera motion up to scale from two fixation correspondences with only linear computation[2] .

6 EXPERIMENTS

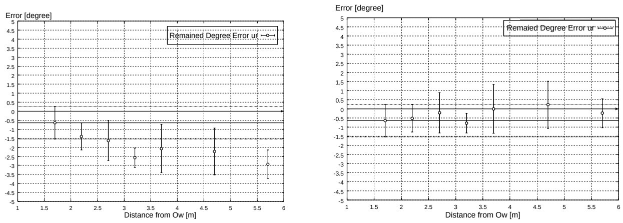
6.1 Precision evaluation of detected blink points

We evaluated the effect of our correction to eye marks detected by EMR-8 under the condition that the depth of person's blink points changes.

We first set a calibration plane whose distance is 1.7m from a person, and then instructed the person to 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 person changes by 0.5m from 1.7m to 5.7m in turn. At each

[1] We assume here that the unit direction vector of a focused line in the world coordinates is known. The vector, however, can be estimated from (5.2) during the motion estimation. Namely, we can compute $\boldsymbol{L}_\mathrm{r}$ (with $\|\boldsymbol{L}_\mathrm{r}\| = 1$) from (5.2) because we know R_0 and R if we have estimated the human motion up to time $t+1$.

[2] Whenever we estimate the translation of the camera motion over two frames, we have one unknown scale factor. The trilinear constraints [5] on corresponding points over three frames enable us to adjust the unknown scales with only linear computation.



(a) not corrected (b) corrected

Figure 10: Errors of detected blink points.

distance, we instructed the person to 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. The nine points here, on the other hand, were also captured by the right camera and formed their images in the right-camera image, whose coordinates were used as the ground truths to evaluate the precision of our correction. Next, we applied our correction described in (4.1) to the eye marks detected by EMR-8 to obtain the corrected coordinates of images of the blink points. We remark here that we carefully measured D_{c} and w to obtain $D_{\mathrm{c}} = 1.7$m and $w = 114$mm. We therefore had the calibration curve shown in Fig. 9.

For the cases with/without our correction, we computed the average and the variance of the residuals from the ground truths over the given set of nine points, and compared the two cases (Fig. 10). Note that error bars in Fig. 10 represent the standard deviation. Fig. 10 shows that the residuals of corrected coordinates almost stably remain small independent of the change in distance of the plane from the person. 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.

6.2 Experiments on estimating a human-motion trajectory

We set up a simulated active wearable vision sensor where two cameras with the baseline of about 27cm were mounted on the stage of a tripod (Fig. 11). Here we employed EVI-G20 and a PC with PentiumIII 750MHz and 1GB memory, both of which are used in our active wearable vision sensor (Fig. 1). We then calibrated the intrinsic and extrinsic parameters of the two cameras using the method proposed by Zhang [16]. The size of images captured by each camera was 640×480 pixels.

We moved the simulated active wearable vision sensor in the scene. The trajectory of the right-camera motion is shown in Fig. 12. The length of the trajectory was about 6m. We marked 35 points on the trajectory and regarded them as samples during the motion. We then applied the binocular independent fixation control only to the samples to estimate the right-camera motion.

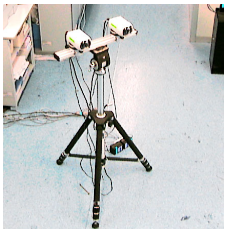
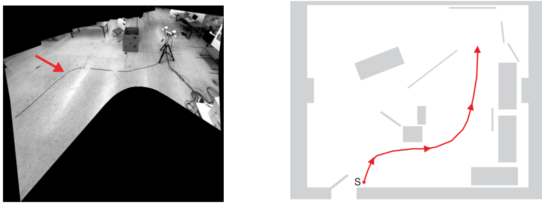


Figure 11: Simulated active wearable vision sensor.

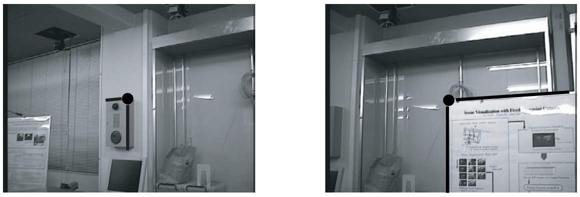


(a) wide-view representation (b) top-view representation

Figure 12: Camera motion trajectory.

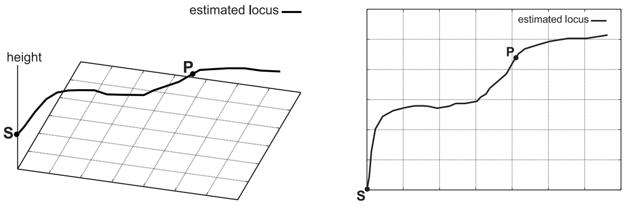
In each image captured by each camera at the starting point of the camera motion, we manually selected a point to set as the fixation point. During the estimation, we updated fixation points 8 times. This updating was also conducted by hand. We used two focused lines for each camera (we thus used four focused lines in total). In detecting lines, we applied the Hough transformation to the edges detected from each image. Fig. 13 shows an example of image pairs captured by the right and left cameras at a marked point. We see that little field of view of the two cameras is common. We remark that the fixation point (the black circle) and two focused lines (the black thick lines) are overlaid onto the images in Fig. 13.

Under the above conditions, we estimated the right-camera motion at each marked point. Fig. 14 shows the trajectory of the right-camera motion that was obtained by concatenating the estimated motions at the marked points.



(a) left-camera image (b) right-camera image

Figure 13: Example of images acquired by the two cameras during the camera motion.



(a) 3D representation (b) top-view representation

Figure 14: Estimated trajectory of the camera motion.

We note that S means the starting point of the motion.

The height from the floor was almost accurately estimated over the trajectory. In fact, the estimated height was almost constant. As for the component parallel to the floor, however, the shape of the former part (from S to P in Fig. 14) of the estimated trajectory fairly coincides with that of the actual trajectory whereas the latter part (after P) of the estimated trajectory has great aberration from the actual trajectory. We have two reasons that may cause this aberration. One is the incorrect estimation of the motion at P and the other is the effect of the estimation error at P upon the subsequent estimations. In other words, since the motion is incrementally estimated, the accumulation of estimation errors and an incorrect estimation at just one marked point cause aberration. The estimation error can be caused by errors in the fixation correspondence or errors in the line detection. Calibration errors of the two cameras also may cause estimation errors.

7 CONCLUSION

We developed an active wearable vision sensor for versatile man-machine interactions based on understanding human intention and activities. The sensor consists of the detector of person's viewing lines and two active cameras.

We first proposed a method for calibrating the sensor so that it detects person's blink points accurately even in a real situation such that the depth of blink points changes. We then proposed a method for incrementally estimating the motion trajectory of a person wearing the sensor. In the former method, we aimed at acquiring information from the person's viewpoint where the vision sensor shares the common field of view with the person. In the latter method, on the other hand, we aimed at autonomously acquiring information for understanding the person's motion trajectory where the field of view of the vision sensor is independent of that of the person. In this way, our active wearable vision sensor enables us to versatilely acquire information for understanding human intention and activities.

Eliminating the accumulation errors in estimating the motion trajectory and improving the accuracy of the estimation are included in the future work. We also plan to develop methods for estimating the position of a person wearing our active wearable vision sensor and for identifying the fixation of his/her viewing lines to make the computer understand his/her interests.

Acknowledgements. The authors are grateful to Akihiro Nakayama and Wataru Nagatomo for the discussion and comments on this work as well as helping to perform the experiments. This work is supported by Grant-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] 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.

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

[4] A. J. Davison and D. W. Murray: Mobile Robot Localisation Using Active Vision, *Proc. of ECCV*, Vol. 2, pp. 809–825, 1998.

[5] R. Hartley and A. Zisserman: *Multiple View Geometry in Computer Vision*, Cambridge Univ. Press, 2000.

[6] 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*, pp.107-114, 2001.

[7] S. W. Lee and K. Mase: Incremental Motion-Based Location Recognition, *Proc. of Int. Symposium on Wearable Computers*, pp. 123–130, 2001.

[8] W. W. Mayol, B. Trdoff and D. W. Murray: Wearable Visual Robots, *Proc. of Int. Symposium on Wearable Computers*, pp. 95–102, 2000.

[9] N. Molton and M. Brady: Practical Structure and Motion from Stereo When Motion is Unconstrained, *Int. J. of Computer Vision*, Vol. 39, No. 1, pp. 5–23 (2000).

[10] D. W. Murray, I. D. Reid and A. J. Davison: Steering and Navigation Behaviors Using Fixation, *Proc. of British Machine Vision Conference*, 1996.

[11] S. Oviatt and P. Cohen: Multimodal Interfaces that Process What Comes Naturally, *Communications of the ACM*, Vol. 43, No. 3, pp. 45–53 (2000).

[12] A. Pentland: Perceptual Intelligence, *Communications of the ACM*, Vol. 43, No. 3, pp. 35–44 (2000).

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

[14] A. Sugimoto, A. Nakayama and T. Matsuyama: Detecting a Gazing Region by Visual Direction and Stereo Cameras, *Proc. of the 16th International Conference on Pattern Recognition*, Vol. III, pp. 278–282, 2002.

[15] M. Turk and G. Robertson: Perceptual User Interfaces, *Communications of the ACM*, Vol. 43, No. 3, pp. 33–34 (2000).

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