2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Identifying a Moving Object with an Accelerometer in a Camera View

Osamu Shigeta, Shingo Kagami and Koichi Hashimoto

***Abstract*— This paper proposes a method for identifying an object which contains an accelerometer out of many moving objects in the view of a stationary camera using motion data obtained by the camera and the accelerometer. The camera and the accelerometer are assumed to be connected with a network but not synchronized. In order to evaluate the similarity of the motion data despite the unknown time lag between the accelerometer and the camera, NCC (Normalized Cross-Correlation) of the signals is computed and its peak is tracked. Since the coordinate system of the accelerometer is unknown, NCC is computed for the norms of the acceleration vectors, which were compensated for the gravitational acceleration component, obtained by the camera and the accelerometer. The experimental results show that the proposed method successfully identified the person wearing the accelerometer out of three walking people. It is also shown that the hand holding the accelerometer was successfully identified out of three moving hands even though the directions of the accelerometer coordinate axes varied temporally due to the free motion of the hand.**

I. Introduction

It is important to estimate the position of a portable information device such as a PDA (Personal Digital Assistant) or a cell-phone in realizing ubiquitous computing [1]. Therefore many methods have been proposed for position estimation as briefly summarized in Section II.

Cameras have been widely used to localize and track objects [2], but identifying an object out of the ones with similar appearance is difficult and some other clues must be combined. A popular measure is to attach to the object a visually detectable identification tag such as 1D or 2D barcodes or temporally coded light sources [3]. These methods are practical, but they cannot be applied in situations where the target object is hidden, for example, in a pocket or a hand because the camera has to capture the image of the tag itself.

In the meantime, accelerometers have been improved with the advancement of the MEMS technology and they have been rapidly becoming smaller, more lightweight and more inexpensive. Following this trend, many devices such as cell-phones, controllers of game consoles and PDAs equipped with accelerometers have been brought to the market, and applications utilizing them have been proposed [4].

Considering the recent advancement of the sensor network technology, it is expected that information from these accelerometers in portable devices will be communicated through a network and combined with visual information from the cameras, which are installed, for example, in public spaces.

This work was supported in part by the Grant-in-Aid for Young Scientists (A) 18680016 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

O. Shigeta, S. Kagami and K. Hashimoto are with Graduate School of Information Sciences, Tohoku University, Aramaki Aza Aoba 6-6-01, Sendai, Japan {shigeta,swk,koichi}(at)ic.is.tohoku.ac.jp

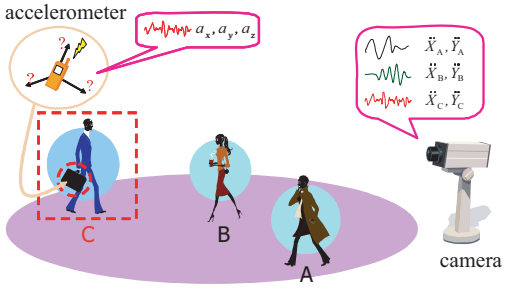


Fig. 1. Application example: Finding a person who has an accelerometer based on motion information gathered by a camera and the accelerometer.

Taking account of these backgrounds, this paper proposes a method to identify an object containing an accelerometer out of many moving objects in a camera view by computing time correlation of motion data obtained by the accelerometer and the camera. Fig. 1 shows an application example. It is assumed that there are many walking people in the camera view, and one of them has a device with an accelerometer in his/her bag. Even if the camera cannot capture the image of the device itself, it is possible to identify the person by combining the signals from the accelerometer and the camera. The identification results will be used, for example, to offer location-aware services for the person and the device.

II. Related work

Other than using cameras, there are many methods for position estimation that could be applied to localizing portable devices.

The most popular and widely used ones are the methods based on GPS (Global Positioning System) [5]. However, it is difficult to estimate location accurately in indoor environments.

Positioning techniques based on the radio frequency technology, using such as RSSI (Received Signal Strength Indication) [6] or TDoA (Time Difference of Arrival) [7], are also actively developed. For indoor use, an interesting method utilizing the power line infrastructure has also been proposed [8]. Although it is reported that their detection accuracy is on the order of sub-meter, it is not accurate enough, for example, to detect gesture motions.

Some positioning systems introducing more specialized equipments can achieve higher degrees of accuracy. Sub-

millimeter accuracy is achieved using a system based on the magnetic field [9]. Some systems based on ultrasonic sensors [10][11] achieve sub-centimeter accuracy. However, it is still unclear whether these specialized equipments will be employed in many portable devices. In addition, magnetic-based methods are susceptible to magnetic interference from the presence of metals or other conductive materials [11].

A related work similar to ours has been reported [12] in which a person is identified out of many people in a camera view by evaluating the correlation of signals from the camera on the ceiling and motion sensors, namely an accelerometer, a gyro and a magnetic sensor, worn by the person. Although it is not clearly described in the paper how the motion sensor data are processed, the method presumably requires some knowledge about the relationship between the coordinate systems of the camera and the motion sensors, which is the biggest difference from the method proposed in this paper.

III. Proposed method

Fig. 2 shows the procedure of the proposed method. It assumes that there is a moving object containing a 3-axis accelerometer among many moving objects in the view of a stationary camera. Moving areas are detected and segmented from the image, although we do not focus on how. The centroids of the moving areas in the image coordinate are denoted by (X_i, Y_i), and the acceleration vectors of them are denoted by $(\ddot{X}_i, \ddot{Y}_i)$, where i denotes the object index, which is omitted unless it causes ambiguity. The acceleration vector of the 3-axis accelerometer is denoted by (a_x, a_y, a_z) where the coordinate axes are fixed to the accelerometer. Here, we apply a low-pass filter to (X_i, Y_i), $(\ddot{X}_i, \ddot{Y}_i)$ and (a_x, a_y, a_z), respectively to get rid of undesired high frequency components.

It is assumed that the accelerometer and the camera have their own internal clocks, so that timestamps can be recorded, which are not synchronized. In order to evaluate the similarity of the motion data, NCC (Normalized Cross-Correlation) of the signals from the accelerometer and the camera is computed.

The problems we have to consider here are two-fold. The first is that the coordinate system of the accelerometer is unknown, and is varying temporally. To solve this problem, we compute NCC for the norms of acceleration vectors, which do not depend on the coordinate system. The second is that the accelerometer reports an acceleration vector including the component of gravitational acceleration, but the gravity direction is unknown. To solve this problem, a roughly estimated gravitational acceleration component is added to the acceleration vector obtained by the camera before calculating its norm.

Fig. 3 shows the time charts of the involved signals. The unknown time lag of the camera signals from the accelerometer signals is denoted by τ, where $-N_2 < \tau < N_1$. N_1 and N_2 specify the range of the time shift for which NCC is computed. Let N_{ws} and N_{wl} $(= N_1 + N_2 + N_{\mathrm{ws}})$ denote the window sizes for NCC. The time t_1, t_2 and t_3 are defined

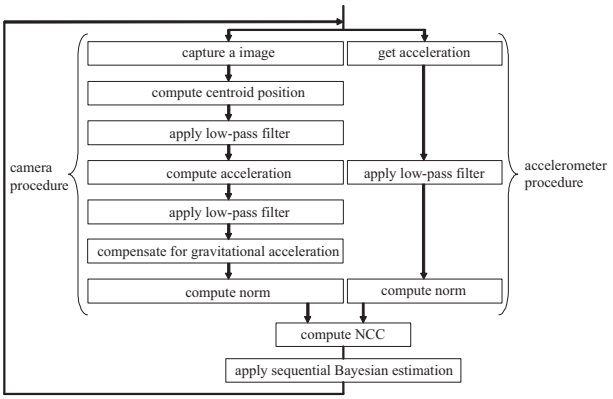


Fig. 2. Procedure of the proposed method.

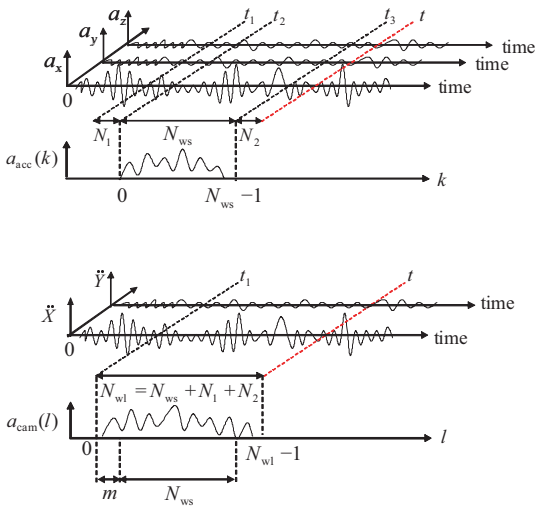


Fig. 3. Time charts of the signals.

as $t_1 = t - N_{\mathrm{wl}}$, $t_2 = t - N_2 - N_{\mathrm{ws}}$, and $t_3 = t - N_2$, respectively, where t is the present time. For $k \in \{0, 1, \cdots, N_{\mathrm{ws}} - 1\}$ and $l \in \{0, 1, \cdots, N_{\mathrm{wl}} - 1\}$, the acceleration norms of the accelerometer and the camera are denoted by $a_{\mathrm{acc}}(k)$ and $a_{\mathrm{cam}}(l)$, which are computed as:

$$a_{\mathrm{acc}}(k) \stackrel{\mathrm{def}}{=} \sqrt{a_x{}^2(t_2 + k) + a_y{}^2(t_2 + k) + a_z{}^2(t_2 + k)}, \quad (1)$$

$$a_{\mathrm{cam}}(l) \stackrel{\mathrm{def}}{=} \sqrt{(\ddot{X}_i(t_1 + l))^2 + (\ddot{Y}_i(t_1 + l) + g_{\mathrm{cam}})^2}, \quad (2)$$

where the gravitational acceleration in the image coordinate is denoted by g_{cam} [pixel/sample^2], and it is assumed that the positive direction of the Y-axis in the image coordinate approximately corresponds to the direction of gravity. Since it is difficult to estimate g_{cam} precisely, we use an approximate value considering that the moving object is around the optical axis:

$$g_{\mathrm{cam}} \simeq -\frac{gf}{dl_{\mathrm{pix}}H^2}, \quad (3)$$

where f [m] is the focal length of the camera lens, l_{pix} [m/pixel] is the pixel size of the image sensor, H [samples/s]

is the sampling frequency of the camera, g [m/s^2] is the gravitational acceleration, and d [m] is an approximate distance between the camera and the object.

NCC between $a_{\mathrm{acc}}(k)$ and $a_{\mathrm{cam}}(l)$ is defined as:

$$r_t(m) \stackrel{\mathrm{def}}{=} \frac{\sum_{n=0}^{N_{\mathrm{ws}}-1} f_t(n)h_t(n+m)}{\sqrt{\sum_{n=0}^{N_{\mathrm{ws}}-1} f_t^2(n)}\sqrt{\sum_{n=0}^{N_{\mathrm{ws}}-1} h_t^2(n+m)}}, \tag{4}$$

where f_t and h_t are defined as:

$$f_t(k) \stackrel{\mathrm{def}}{=} a_{\mathrm{acc}}(k) - \bar{a}_{\mathrm{acc}}, \tag{5}$$

$$h_t(l) \stackrel{\mathrm{def}}{=} a_{\mathrm{cam}}(l) - \bar{a}_{\mathrm{cam}}, \tag{6}$$

where $\bar{a}_{\mathrm{acc}}$ is the average value of $a_{\mathrm{acc}}(k)$ over $k = 0, \cdots, N_{\mathrm{ws}} - 1$, and $\bar{a}_{\mathrm{cam}}$ is the average value of $a_{\mathrm{cam}}(l)$ over $l = 0, \cdots, N_{\mathrm{wl}} - 1$. The symbol $m \in \{0, 1, \cdots, N_1 + N_2\}$ denotes the time shift.

At each present time t, we find the time shift m where the maximum NCC value is obtained, which is denoted by $\hat{m}$. This $\hat{m}$ is expected to correspond to the time lag between the signals from the accelerometer and the camera. The maximum value of NCC is expected to express the similarity between the signals.

Practically, due to various disturbances or coincidental motions, the NCC peak will not always appear at the ground-truth point, or NCC of the signals corresponding to a false object might happen to exhibit a peak instantly. In order to track a consistently appearing peak, sequential Bayesian estimation is applied.

Let the time lag at the time t be denoted by m_t, all the measurement data at the time t, including the ones from the accelerometer and the camera, be denoted by z_t, and let Z_t denote $\{z_1, z_2, \cdots, z_t\}$. The conditional probability $p(m_t|Z_t)$ is computed from $p(m_{t-1}|Z_{t-1})$ recursively as:

$$p(m_t|Z_t) \propto p(z_t|m_t)p(m_t|Z_{t-1}), \tag{7}$$

$$p(m_t|Z_{t-1}) \stackrel{\mathrm{def}}{\propto} p(m_{t-1}|Z_{t-1}) + c, \tag{8}$$

$$p(z_t|m_t) \stackrel{\mathrm{def}}{\propto} r_t(m_t) + 1. \tag{9}$$

Equation (8) represents the assumed dynamics of the time lag m_t, where m_t is assumed to be unchanged, while a small offset c is added in order to prevent a posterior probability from becoming zero. The likelihood function defined in (9) where $r_t(m_t)$ is biased so that it has a positive value. This equation means that the larger the correlation at some time lag is, the larger $p(z_t|m_t)$ is. This likelihood function is not physically grounded, but this kind of ad-hoc definition of a likelihood function based on NCC is sometimes employed [13]. We use a uniform distribution as the initial density $p(m_0)$ because we have no information about the time lag at first.

IV. Experimental Setup

We carried out experiments under the environment shown in Fig. 4. A Point Grey Research Dragonfly Express camera captured images of a moving object. The camera was connected to a PC through IEEE1394b. To get acceleration data of the moving object, we used a Freescale Semiconductor MMA7260Q accelerometer, which was connected to PC through USB. Table I summarizes the specification of the accelerometer and the camera.

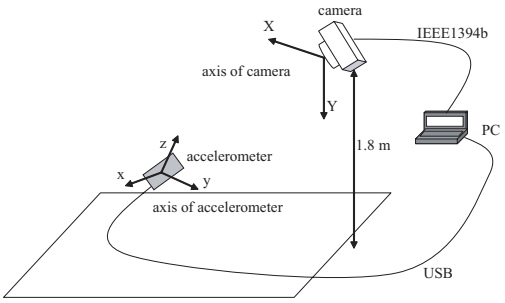


Fig. 4. Experimental setup.

Fig. 5. Snapshots of tracked targets. The left image shows a walking person and the right one shows a moving hand. The red ellipses indicate the tracked target regions, and image processing was done within the green rectangles around the targets.

TABLE I

Specifications of the accelerometer and the camera.

Focal length of the lens f,	4.1 [mm].
Camera resolution,	640 × 480 [pixels].
Pixel size l_{pix},	7.4 [μm].
Format,	color 8 [bit].
Accelerometer resolution,	8 [bit].
Accelerometer measurement range,	$-2.0 \sim 2.0$ [g].

To use the proposed method, we first have to extract moving regions from the images. In this experiment, we used the CAMSHIFT algorithm [14] to track a hand and a jacket motion. Fig. 5 shows some snapshots of the tracked targets in the camera view. The red ellipses indicate the tracked target regions. To reduce the computation time, the tracking image processing was done only within the green rectangles around the targets. As the approximate distance between the camera and the object, which is required in computing a_{cam}, we used the value $d = 3$ [m].

Table II shows the parameters we set in the experiment. These parameters were determined empirically. The average ground-truth time lag between the time stamps on the signals from the camera and the accelerometer, which were recorded immediately after the PC obtained the data, was 15 [ms] and the standard deviation was 0.12 [ms].

TABLE II

PARAMETERS USED IN THE EXPERIMENT.

Sampling frequency H,	63 [Hz].
Number of sampled data,	1000 [samples].
Window size N_{ws},	300 [samples].
Time shift range N_1,	50 [samples].
Time shift range N_2,	50 [samples].
Offset in (8) c,	0.00001.

We carried out the experiments for the following two type of moving objects:

- a walking person with the accelerometer in his/her trouser pocket,
- a moving hand holding the accelerometer.

For the first type, the jacket of the person was tracked in the camera view as shown in the left image in Fig. 5. The person had the accelerometer in his/her trouser pocket and walked freely in the camera view.

For the second type, the hand with the accelerometer, which was moved freely, was tracked as shown in the right image in Fig. 5. We prevented the hand from moving near flesh-color objects such as other hands or faces because the CAMSHIFT tracking fails.

In order to evaluate the effectiveness of the proposed method, we must compare the results for the true object (with the accelerometer) and false objects (without the accelerometer). Instead of moving several objects simultaneously within the camera view, we moved only one object in the view and used the data $(\ddot{X}_1, \ddot{Y}_1)$ and (a_x, a_y, a_z) obtained in a trial of the experiment as true object data, and $(\ddot{X}_2, \ddot{Y}_2)$ and $(\ddot{X}_3, \ddot{Y}_3)$ obtained in two other trials as false object data. These acceleration vectors from the camera are referred to as data 1, 2 and 3, respectively. The situation of identifying a moving object out of three objects was simulated by evaluating the correlation between (a_x, a_y, a_z) and these data 1, 2 and 3 off-line.

V. EXPERIMENTAL RESULTS

A. Walking Person

We evaluated the proposed method for various ways of walking. A set of results corresponding to one of them is described here in detail, where the person was walking at the distance of from 1 to 4 [m] from the camera.

Fig. 6 shows NCC for data 1, 2 and 3 calculated as (4). Fig. 7 shows the maximum values of these NCC, and the time shift values where the maximum NCC were detected. It can be observed that NCC for data 1 was periodical because the period of walking was constant. The results of the sequential Bayesian estimation are shown in Fig. 8. Here, at time t, the time shift value where the maximum $p(m_t|Z_t)$ was detected is referred to by the estimated time shift. Table III shows the average values and standard deviations of the estimated time shift at time $t = 400 \sim 1000$.

Considering that the standard deviation for the data 1 is smaller than the other two, the data 1 can be believed to exhibit a consistent peak at the time shift $m = 60$ and it can be concluded that this object was successfully identified as

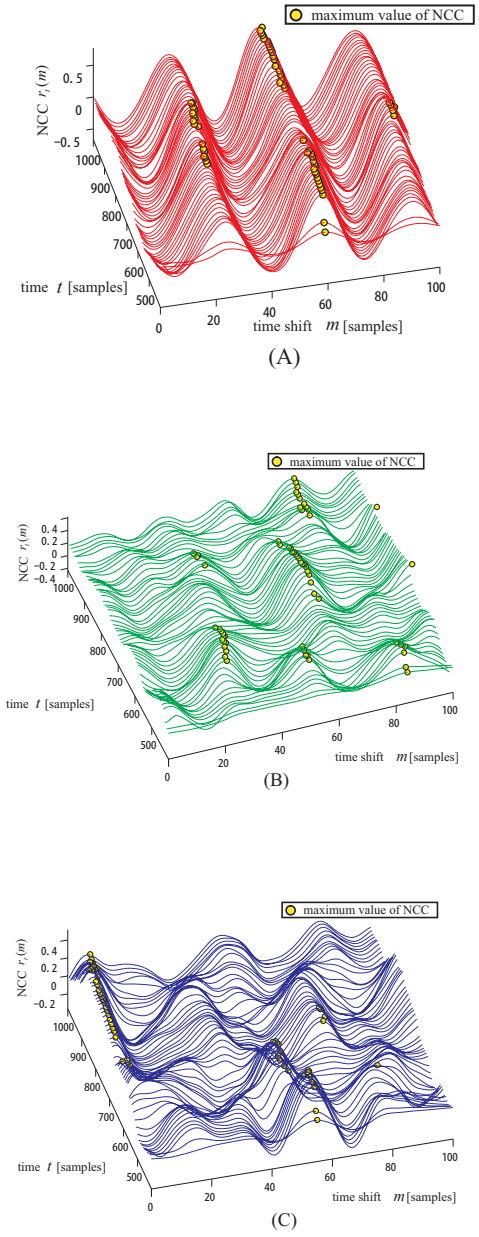


Fig. 6. NCC for the case of a walking person. (A) shows NCC for data 1, which corresponds to the true object which contains an accelerometer. (B) and (C) show NCC for data 2 and 3, respectively.

the one containing the accelerometer. For the other two data, consistent peaks could not be observed.

The other results that were not presented here, including the case of walking sideways or backward, showed that the objects were successfully identified.

B. Moving Hand

We also evaluated the proposed method for various ways of hand moving. A set of results corresponding to one of

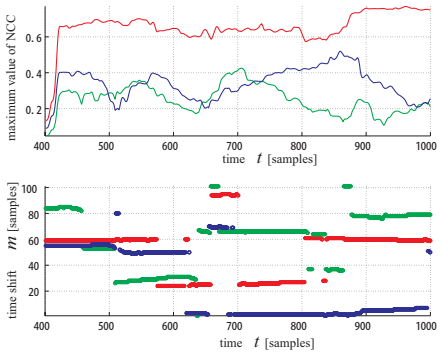


Fig. 7. NCC peaks for the case of a walking person. The upper shows the maximum values of NCC. The lower shows the time shift values where the maximum NCC were detected.

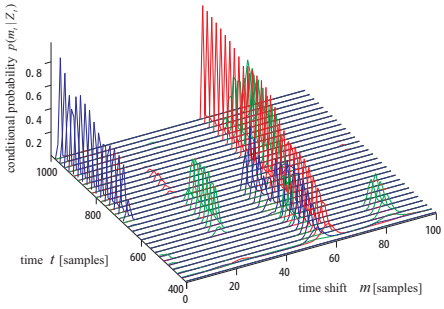


Fig. 8. The results of sequential Bayesian estimation for the case of a walking person.

TABLE III

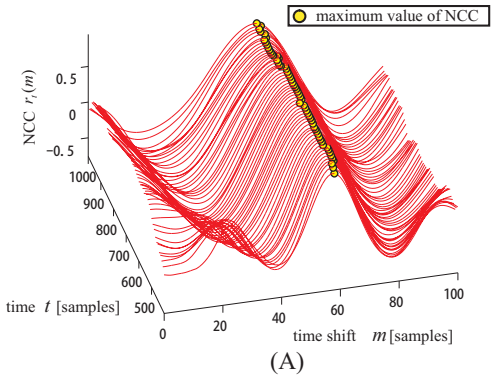
ESTIMATED TIME SHIFT OBTAINED BY SEQUENTIAL BAYSIAN ESTIMATION FOR THE CASE OF A WALKING PERSON.

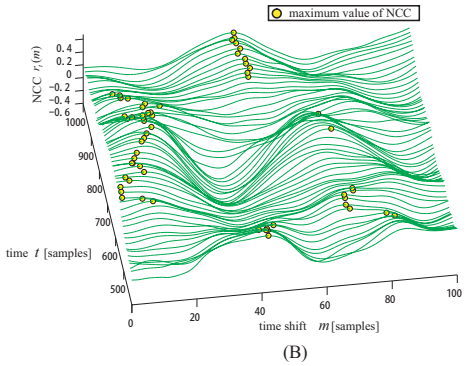
data	average value [samples]	standard deviation [samples]
1	60	0.47
2	60	20
3	26	25

them is described here in detail, where the directions of the coordinate axes of the accelerometer were varied rapidly due to the rapid motion of the hand.

Fig. 9 shows NCC for data 1, 2 and 3. Fig. 10 shows the maximum values of these NCC, and the time shift values of the maximum NCC. The results of the sequential Bayesian estimation are shown in Fig. 11. Table III shows the average values and standard deviations of the estimated time shift at time $t = 400 \sim 1000$.

Considering that the standard deviation for the data 1 is the smallest, the data 1 can be believed to exhibit a consistent peak at the time shift $m = 58$ and it can be concluded that this hand was successfully identified as the one holding the accelerometer.





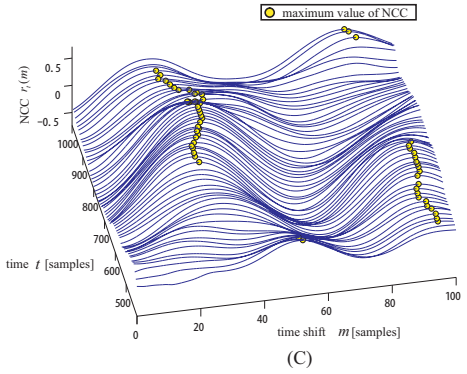


Fig. 9. NCC for the case of a moving hand. (A) shows NCC for data 1, which corresponds to the true object which contains an accelerometer. (B) and (C) show NCC for data 2 and 3, respectively.

TABLE IV

ESTIMATED TIME SHIFT OBTAINED BY SEQUENTIAL BAYSIAN ESTIMATION FOR THE CASE OF A MOVING HAND.

data	mean value [samples]	standard deviation [samples]
1	58	0.33
2	33	28
3	56	31

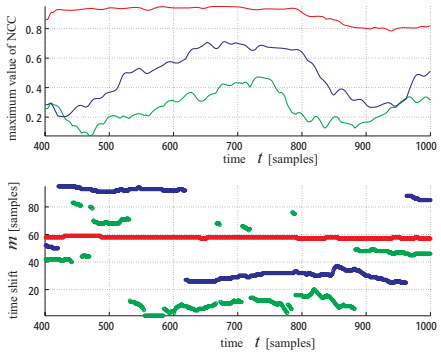


Fig. 10. NCC peaks for the case of a moving hand. The upper shows the maximum values of NCC. The lower shows the time shift values where the maximum NCC were detected.

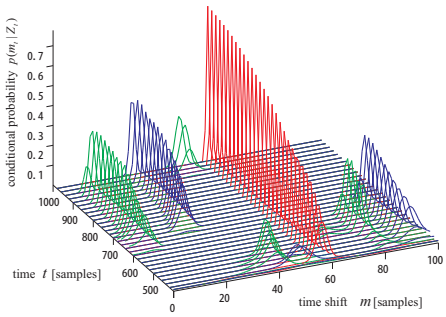


Fig. 11. The results of sequential Bayesian estimation for the case of a moving hand.

VI. Discussion

In the experiment of the walking persons, it should be noted that the person actually walked at the distance of from 1 to 4 [m] while the parameter used to compute the gravitational acceleration was fixed at $d = 3$. This shows that the accurate distance from the camera to the object is not required for identifying an object.

In the experiment of the moving hands, it should also be noticed that the NCC peak exhibited values higher than 0.8 in Fig. 10 even though the directions of the accelerometer coordinate axes rapidly varied. This shows that our method is insusceptible to the change of the coordinate system of the accelerometer.

Although our experimental evaluation was successful on the whole, it is clear that there are many particular cases in which our method will not work. For example,

- the object moves only along the optical axis,
- no acceleration is obtained by the camera, for example, due to a constant velocity motion,
- closely similar motions are detected by the camera.

These particular motions, however, will not last permanently in normal situations and it is expected that a long time observation will deliver sufficient information for identification.

VII. Conclusion

This paper has presented a method for identifying an object containing an accelerometer out of many moving objects in a camera view by computing NCC of the norms of acceleration vectors obtained by the accelerometer and the camera. The experimental results show this method could identify the person with the accelerometer out of three walking people. It is also shown that the moving hand holding the accelerometer was identified out of three hands even though the directions of the accelerometer coordinate axes rapidly varied.

Future work will include establishing a clear criterion for identification. Our method regards the one which has the smallest standard deviation of the estimated time shift as the object with the accelerometer. However, this method fails when the object with the sensor is out of the camera view. In such a case, another object which is not related to the true one will be regarded as the one with the sensor. We must establish a reliable criterion, for example, based on statistical methods.

VIII. Acknowledgements

The authors wish to express their gratitude to Mr. Kei Watanabe and Mr. Shogo Arai at Tohoku University for fruitful discussions.

References

[1] M. Weiser, "The computer for the 21st century," *Scientific American*, Vol. 265, pp. 94–104, 1991.
[2] Y. Ma, S. Soatto, J. Kosecka and S. S. Sastry: *An Invitation to 3-D Vision*. Springer, 2003.
[3] N. Matsushita, D. Hihara, T.Ushiro, S. Yoshimura, J. Rekimoto and Y. Yamamoto, "ID CAM: A Smart Camera for Scene Capturing and ID Recognition," *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 227–234, 2003.
[4] T. Iso and K. Yamazaki, "Gait analyzer based on a cell phone with a single three-axis accelerometer," *in Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pp. 141–144, 2006.
[5] R. Bajaj, S. L. Ranaweera and D. P. Agrawal, "GPS Location Tracking Technology," *IEEE Computer*, Vol. 35, No. 4, pp. 92–94, 2002.
[6] P. Bahl and V. Padmanabhan, "RADAR: An In-Building RF-Based User Location and Tracking System," *in Proceedings of IEEE Infocom 2000, IEEE CS Press, Los Alamitos, California*, pp. 775–784, 2000.
[7] http://www.hitachi-cable.co.jp/en/infosystem/news/20031119la.html
[8] S. N. Patel, K. N. Truong, and G. D. Abowd. "PowerLine Positioning: A Practical Sub-Room-Level Indoor Location System for Domestic Use," *in Proceedings of Ubicomp 2006*, LNCS 4206, 2006.
[9] http://www.polhemus.com/
[10] A. Nishitani, Y. Nishida, and H. Mizoguchi, "Omnidirectional Ultrasonic Location Sensor," *in Proceedings of The 4th IEEE International Conference on Sensors* , pp. 684–687, 2005.
[11] N.B. Priyantha, A. Chakraborty, and H. balakrishnan, "The Cricket Location-Support system," *in Proceedings of the 6th International Conference on Mobile Computing and Networking (ACM MobiCom2000)*, pp. 32–43, 2000.
[12] J. Kawai, K. Shintani, H. Haga and S. Kaneda, "Identification and positioning based on motion sensors and a video camera," *The 4th IASTED International Conference on WEB-BASED EDUCATION*, No.461-809, 2005.
[13] S. Thrun, W. Burgard and D. Fox: *Probabilistic Robotics*. The MIT Press, pp. 174–176, 2005.
[14] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal, 2nd Quarter*, 1998.