People Tracking by Cross Modal Association of Vision Sensors and Acceleration Sensors

Tetsushi Ikeda, Hiroshi Ishiguro, *Member, IEEE* and Takuichi Nishimura, *Member, IEEE*

***Abstract*— To realize accurate tracking of people in the environment, many studies have been proposed using vision sensors, floor sensors, and wearable devices. The problem of using vision sensors is that they do not provide ID information of each people and there are ambiguities when people come across. To solve the problem, we propose to combine acceleration sensors that are attached to the human body. Since the signals from vision sensors and acceleration sensors synchronize when they observe same person who are acting or walking in the environment, these signals are not independent. The correlation between the signals is evaluated based on the canonical correlation analysis. Experimental results are shown to detect gesture and to track people to confirm the effectiveness of the proposed method.**

I. Introduction

In order to realize intelligent environment that supports human activities, much work on sensor network has been done by integrating many kinds of sensors in the environment and wearable devices. Accurate and reliable tracking of people and knowing their positions in the environment is one of fundamental problems in sensor networks. The sensors used in previous approaches are classified into three main groups.

(1) Vision sensors

Vision sensors are widely used to understand the scene in the environment, and much works have been done to recognize human behavior using vision sensors [1]. Vision sensors provide much information about people in the environment, not only their positions but shape, color, and gestures. The defects are that vision sensors do not provide ID information and there are ambiguities in association when two people are coming across.

(2) Floor sensors

By spreading touch sensor network on the floor, the positions of people are accurately detected [2] [3] [4]. Floor sensors are very reliable, but it does not provide any information to distinguish each person.

(3) Wearable devices

Positions of people are detected by using wearable device on the body. Many systems have been proposed using infrared [5], ultrasonic wave [6], RFID [7], and wireless LAN [8]. Since the ID information to distinguish each person is explicitly sent to the system, personal identification during tracking is perfect. However, we have to install many reader devices in the environment to obtain positions of people accurately. And it is desirable that people carry the device in natural manner.

In this paper, we propose to integrate vision sensors in the environment and acceleration sensors that are attached to the body. A typical problem to track people with only vision sensors is that it is difficult to recognize ID of people in the images. Since an acceleration sensor measures the motion of each person and it contains the ID of the person, the association problem is solved effectively. We suppose that our method is applied to track people who are carrying cellular phones that have acceleration sensors inside. So we are carrying acceleration sensors in daily life and our method works under current information infrastructure.

The problem is how to integrate vision sensors and acceleration sensors in different representations. Many works has been done in the research area of the sensor fusion, and typical approach is to convert each sensory signal to a common representation before integration. For example, to integrate signals from microphones and video cameras, locations of a sound source is estimated in sound and video independently. Then the precise position is computed by integration of estimated locations. However, since acceleration sensors do not provide location information directly and there are large drift in estimating position by integrating acceleration sensor signal, it is difficult to apply previous sensor fusion methods.

In this paper, we propose novel integration method to integrate floor sensors and acceleration sensors based on statistical method. We solve the association problem by evaluating correlation between sensory signals by the canonical correlation analysis (CCA).

In section 2, the sensor fusion framework based on signal correlation is described. In section 3, the algorithm to integrate vision sensors and acceleration sensors are described in detail. In section 4, experimental results are shown. In section 5, we conclude the paper.

Manuscript received April 9, 2007.

T. Ikeda and H. Ishiguro are with the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Osaka, JAPAN (e-mail: ikeda@ed.ams.eng.osaka-u.ac.jp).

T. Nishimura is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, JAPAN.

II. Cross Modal Association of Different Kinds of Sensors based on Computing Correlation

A. Cross modal association based on computing correlation

When different kinds of sensors observe same information source, the observed signals display synchrony and the signals are not independent. Recently, several studies of sensor integration have been done by extracting synchrony between the signals from different kinds of sensors based on statistical methods (Fig. 1).

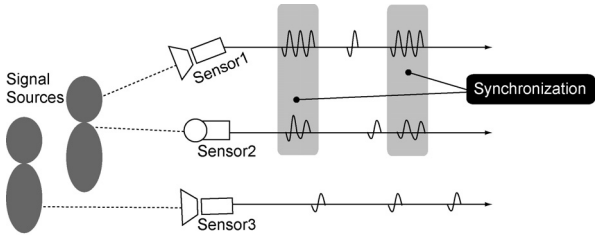


Fig. 1. Integrating different kinds of sensors based on computation of correlation between signals.

Hershey et al. [9] observed people speaking alternately with a camera and a microphone. They extracted synchrony between the audio signal and the brightness of the pixel around the speaker's mouth. They localized the speaker in the image by computing mutual information between the signals. This method has extended and has been applied to especially sound source localization problem [10] [11] [12]. A limitation of the method is the assumption that the target does not move in the images. In coping with a moving target, object detection methods using object models are applied before the integration process [13] [14]. In these methods, the object detection stage and the integration stage is separated

B. Signal source detection and tracking by maximizing correlation between signals

The problem of the two-stage approach is that it is not robust integration method. Since the object detection stage and the integration stage are separated, the integration fails if the object detection fails. In this paper, these stages are integrated into one process. We have proposed to detect and track the sound source simultaneously based on the criteria of mutual information maximization [15]. However, number of the sound source is limited to one in the method.

C. Integration of vision sensors in the environment and wearable acceleration sensors

In this paper, we propose to integrate vision sensors in the environment and wearable acceleration sensors that are attached to the body. By sending the acceleration sensor signal and the ID of each person wirelessly, everyone in the environment are detected and tracked in video images by computing correlation between sensory signals.

The sensor integration approach by detecting synchrony between sensory signals is a general approach and it does not depend on the type of sensors. However, previous studies mainly focused on the sound source localization in video

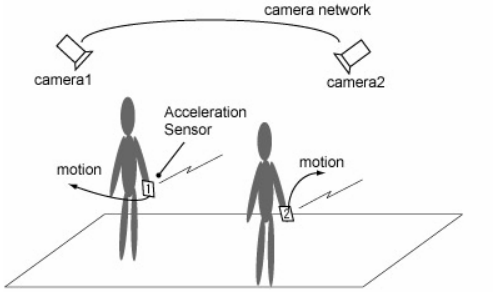


Fig. 2. Integrating different kinds of sensors based on computation of synchrony between signals. The motions of people in the environment are measured by acceleration sensors on the body and transmitted to the sensor network.

images. It is important to apply the approach to many types of sensors and investigate the possibility and the limitation of the approach. This paper is the first trial of integrating vision sensors and acceleration sensors using this approach.

D. Estimating the most correlated direction of the acceleration signal by the canonical correlation analysis

Many kinds of motion are measured by a three dimensional acceleration sensor. In order to detect the correlation between a camera and an acceleration sensor, it is important to estimate the direction of the acceleration with the largest correlation. In this paper, we applied the canonical correlation analysis (CCA) to estimate the direction that maximizes correlation between signals [12][13].

III. Algorithm to Integrate Video Cameras and Acceleration Sensors

A. Preprocessing

Fig. 3 shows the typical acceleration sensor signal when two people walks. Each person has an acceleration sensor on the right hand. The acceleration signal is averaged in each video frame.

Fig. 4 shows the examples of images from two cameras in the environment. Frame difference between subsequent video frames is computed.

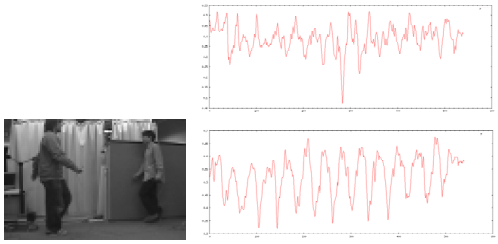


Fig. 3. Signals from acceleration sensors when two people walk in the environment. Each person has an acceleration sensor on the right hand.

Fig. 4. An example images and the frame difference image.

B. Computing canonical correlation between acceleration signals and video signals

We estimate the direction of the motion by computing correlation between signals. In this paper, we apply canonical correlation analysis that finds the linear mapping that maximizes correlation between video signal and acceleration signal.

Given three dimensional acceleration sensor signal at each of the T video frames, the signals can be expressed as:

$$\mathbf{X} = \left[\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z\right]$$

where

$$\mathbf{a}_x = \left[a_x(1) \quad a_x(2) \quad \cdots \quad a_x(N)\right]^T$$

and $(a_x(t), a_y(t), a_z(t))$ is the observed acceleration signal at frame t. A sequence of intensity of a pixel can be expressed as:

$$\mathbf{Y} = \left[I_{x,y}(1) \quad I_{x,y}(2) \quad \cdots \quad I_{x,y}(N)\right]^T$$

where $I_{x,y}(t)$ is an intensity of the pixel (x,y) at frame t.

By subtracting the average of each row, $\widetilde{X}$ and $\widetilde{Y}$ represents deviation matrices:

$$\widetilde{\mathbf{X}} = \left[\mathbf{a}_x - \overline{\mathbf{a}}_x, \mathbf{a}_y - \overline{\mathbf{a}}_y, \mathbf{a}_z - \overline{\mathbf{a}}_z\right]$$

$$\widetilde{\mathbf{Y}} = \mathbf{Y} - \overline{\mathbf{Y}}$$

Canonical correlation analysis finds canonical correlation coefficients a, b that maximizes correlation $r(a,b)$:

$$r(a,b) = \frac{a^T S_{XY} b}{\sqrt{a^T S_{XX} a}\sqrt{b^T S_{YY} b}}$$

that satisfies following conditions.

$$a_i{}^T S_{XX} a_i = 1$$

$$b_i{}^T S_{YY} b_i = 1$$

where S_{XX}, S_{YY}, S_{XY} are variance covariance matrices:

$$S_{XX} = \frac{1}{N}\widetilde{X}^T\widetilde{X}$$

$$S_{YY} = \frac{1}{N}\widetilde{Y}^T\widetilde{Y}$$

$$S_{XY} = \frac{1}{N}\widetilde{X}^T\widetilde{Y}$$

Then the vector a and b are computed by SVD:

$$S_{XY}{}^{-1} S_{XY} S_{YY}{}^{-1} = \mathbf{A}\,\mathbf{\Lambda}\,\mathbf{B}^T$$

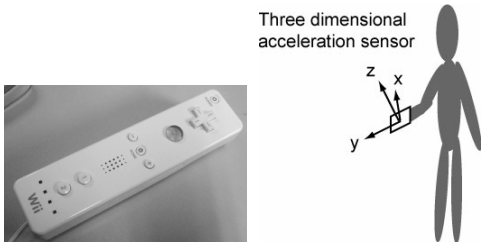
where A and B is orthogonal matrices and Λ is a diagonal matrix. The first column vector of A and B is the canonical correlation vectors a, b, respectively.

IV. EXPERIMENTS

To confirm the effectiveness of the proposed method, we apply the method to detect and track people in a room. The video signal is sampled at 30 frames/second, and the image size is 360x240.

A. Sensors

The acceleration sensor used in the experiments is ADXL330 (Analog Devices, Inc.) (Fig. 5). All people in the environment are assumed to have the sensor with the right hand. The motion of the Wiimote is sensed by a 3-axis linear accelerometer located slightly left of the center of the controller. The signal is sampled at 70 kHz, and the signals



Output signal	8 bit integer
Frame rate	70 Hz

Fig. 5. Acceleration sensor used in the experiments.

are transmitted via Bluetooth. The average of the signal is computed in each video frame. An example signal is shown in Fig. 3.

Standard digital video cameras are used in the experiments. In the second experiment, we placed two video cameras in the environment.

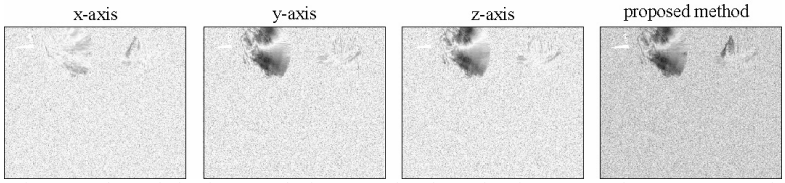
B. Experiment (1)

First we recorded two people that are shaking their arms. Fig. 6 shows the computed correlation function between each pixel in the images and each acceleration sensor.

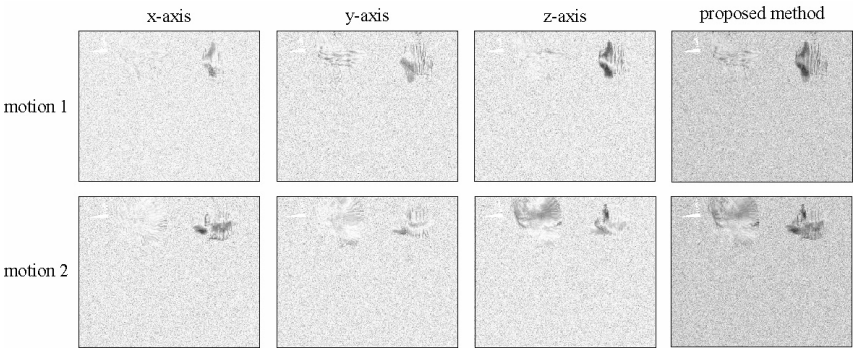
C. Experiment (2)

Next we recorded two people that walk across. Fig. 7 shows the detection and tracking results when two people go across. The region with the highest average correlation is detected and tracked in images for each acceleration sensor signal. The right of the figure shows the trajectory of the region that maximizes correlation between sensory signals.

a) Images taken by a camera.



b) Intensity of computed correlation between the images and each acceleration sensor. The Pearson's correlation function is computed for each pixel in images. Darker pixels show larger value of correlation. In this motion, the acceleration signals of y-axis and z-axis have large correlation.



c) Intensity of correlation of the right person for different types of motions. For the first motion, significant correlation is detected only in the z-axis. However, the axis with the largest correlation is different depending on the type of motion. In the proposed method, the direction with the largest correlation is selected using canonical correlation analysis and significant correlation is detected.

Fig. 7. Experimental result (1).

V. CONCLUSION

In this paper, we proposed a novel method to track multiple people in the environment by integrating vision sensors and acceleration sensors that are attached to the human body. Since many cellular phones have an acceleration sensor, the proposed framework can be applied in the common environment. By using only vision sensors, it is difficult to know the ID of each people and to track each people without correspondence ambiguity. We proposed to evaluate synchrony between vision sensors and acceleration sensors on the body based on canonical correlation analysis. By selecting regions in images that changes of the intensity are correlated to acceleration sensor signal, the correct associations are estimated.

To confirm the effectiveness of the proposed method, two experimental results are shown. In the first experiment, two people who shake their hands are located independently in the images. In the second experiment, two people who walk in the environment are detected and tracked.

In the future, we plan to apply the proposed method to track more than two people who show various behaviors like skipping and dancing.

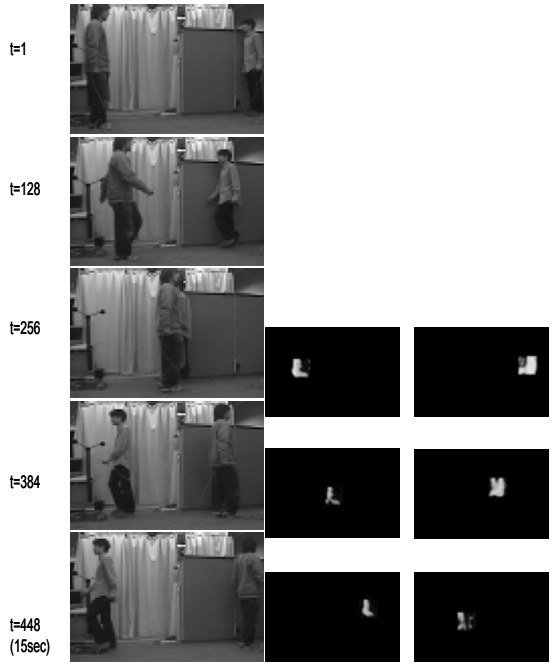


Fig. 6. Experimental result (2). Two people walk across. The left figures show the original images and the right show computed correlation function between intensity of each pixel and the signal from the acceleration sensor on the left person.

REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," IEEE Trans. on Systems, Man and Cybernetics, Part C, Vol. 34, No. 3, pp. 334-352, 2004.

[2] R. J. Orr and G. D. Abowd, "The smart floor: A mechanism for natural user identification and tracking," Proc. Conference on Human Factors in Computing Systems (CHI 2000), pp. 303-306, 2000.

[3] T. Murakita, T. Ikeda, and H. Ishiguro, "Human tracking using floor sensors based on the markov chain monte carlo method," Proc.17th Int. Conf. on Pattern Recognition (ICPR 2004), Vol.4, pp. 917-920, 2004.

[4] T. Mori, Y. Suemasu, H. Noguchi and T. Sato, "Multiple people tracking by integrating distributed floor pressure sensors and RFID system," Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, Vol. 6, pp.5271-5278, 2004.

[5] R. Want, A. Hopper, V. Falc a″ o and J. Gibbons, "The active badge location system," ACM Trans. Inf. Syst., Vol. 10, No. 1, pp. 91-102, 1992.

[6] A. Harter, A. Hopper, P. Steggles, A. Ward and P. Webster, "The anatomy of a context-aware application," Proc. 5th Annual ACM/IEEE Int. Conf. on Mobile Computing and Networking (Mobicom '99), 1999.

[7] T. Amemiya, J. Yamashita, K. Hirota and M. Hirose, "Virtual leading blocks for the deaf-blind: A real-time way-finder by verbal-nonverbal hybrid interface and high-density RFID tag space," In Proc. of IEEE Virtual Reality Conference 2004, pp. 165-172, 2004.

[8] P. Bahl and V. N. Padmanabhan, "Radar: An in-building RF-based user location and tracking system," Proc. of IEEE INFOCOM 2000, Vol. 2, pp. 775-784, 2000.

[9] J. Hershey, H. Ishiguro, and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," Proc. Neural Information Processing Systems (NIPS'99), 1999.

[10] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," Proc. Neural Information Processing Systems (NIPS'00), 2000.

[11] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," Proc. of ACM International Conf. on Multimedia 2002, pp. 303–306, 2002.

[12] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association," Proc. ACM Int. Conf. on Multimedia 2003, pp. 604.

[13] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," Proc. Neural Information Processing Systems (NIPS'00), pp.814-820, 2000.

[14] T. Ikeda, H. Ishiguro, and M. Asada, "Attention to clapping - a direct method for detecting sound source from video and audio -," Proc. IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI'03), 2003, pp. 264–268.

[15] T. Ikeda, H. Ishiguro, and M. Asada, "Sensor fusion as potimization: maximizing mutual information between sensory signals," Proc. 17th Int. Conf. on Pattern Recognition (ICPR'04), 2004, pp. 501–504.