6—4
A Multi-Modal Interface for Recognizing Gestures Expressed by Cyclically Repeated Motion of the Hand

Masanori Nishimura Atsushi Nishikawa* Kengo Koara Fumio Miyazaki
Department of Systems and Human Science, Graduate School of Engineering Science
Osaka University

Abstract

Human naturally uses not only voices but also the gestures composed of cyclically repeated motion of the hand in daily life to express the direction to which he/she wants another person or machine to go. In this paper, we first mention that the vocalization timing correlates closely with the human intention expressed by such gestures. Then we propose a multi-modal method to recognize a gesturer's intention in real-time based on the hand motion velocity at the moment of the gesturer's vocalization estimated by the combination of the optical flow from motion images and the voice trigger extracted from the microphone input. Experimental results of the proposed method applied to mobile robot navigation are also presented.

1 Introduction

Human naturally uses not only voices but also the gestures composed of cyclically repeated motion of the hand in daily life to express the direction to which he/she wants another person or machine to go. If machines or computer systems are able to recognize such human intention in real-time and without any contact devices such as keyboard and joystick, a person can operate the machine as easily as he/she instructs another person by voice and/or gestures.

Methods have been so far proposed that involve real-time visual recognition of gestures expressed by cyclically repeated motion of the hand[1][2][3][4]. Most of these studies, however, assume that the repetitive motion follows a pre-defined model pattern. That is, computer forces human to make a special gesture motion— for example, *(1)at first you make both hands static, (2)then move one of the hands cyclically only two times, (3)and make the state of rest again*[2], or *(1)at first you turn the palm of your left hand toward the camera, (2)then keep its position, (3)and move the right hand cyclically*[4]. This means human always has to take care whether his/her action may satisfy the features of the corresponding gesture model. Nobody would say such systems are user-friendly!

*Address: 1-3 Machikaneyama, Toyonaka, Osaka 560-8531 Japan. E-mail: atsushi@me.es.osaka-u.ac.jp

In this paper, we first mention that the vocalization timing correlates closely with the human intention expressed by cyclically repeated gestures. Then we propose a multi-modal method to recognize a gesturer's intention in real-time based on the hand motion velocity at the moment of the gesturer's vocalization estimated by the combination of the optical flow from motion images and the voice trigger extracted from the microphone input. With the proposed method, human does not have to give attention to the repetition number of motion, hand position and velocity, and the trajectory shape. Experimental results of the proposed method applied to mobile robot navigation are also presented.

2 Analysis of Vocalization Timing in Cyclic Gesture Motion

In order to analyze the relationship between the vocalization timing and the intention of gestures expressed by cyclically repeated motion of the hand, we at first had 6 subjects make natural gestures 20 times with vocalization of their favorite instruction word such as "KOCCHI"(the Japanese for "this way") in the following two cases respectively: one is the case when the direction to which the gesturer wants the virtual communication agent to move is "left", the other is the case when the direction to which the gesturer wants the agent to move is "right". That is, the total number of sample data is 240 (2 cases × 20 times × 6 person). In the experiment, a CCD camera with focal length of 8mm was placed at the distance of about 1000mm ahead of the gesturer in advance. Both a special hardware with a high speed correlation processor chip(Color Tracking Vision TRV-CPW5 by Fujitsu Co. Ltd) and a 166MHz Intel MMX Pentium PC(OS: Linux) were then used to determine the hand motion velocity(vector) in real-time(at the rate of 15Hz) as the mean of optical flow vectors from motion images[1], while the vocalization period was estimated by thresholding the sound-level of the microphone

[1]See [5] and [6] for details about the performance of the Color Tracking Vision(CTRV) and the methods for controlling the CTRV based on Linux respectively. Details of the methods in relation to estimation of hand motion velocity from optical flow can be found in [2][3].

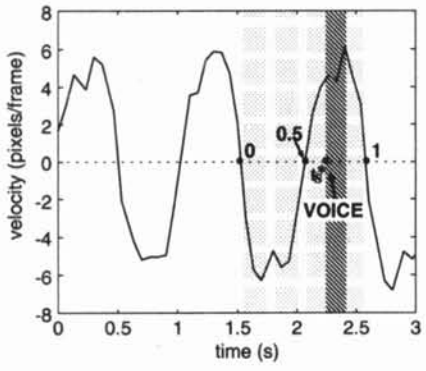


Figure 1: The transition of hand motion velocity and the vocalization timing.

input. In this experiment, we instructed the subject neither how many times, how much velocity, nor how much amplitude he/she should move the hand cyclically. Also we did not care what kind of instruction words he/she would say in his/her gesturing. Furthermore, we did not tell them about the purpose of the experiment in advance. These are all because we wanted to measure "natural" vocalization timing in "natural" gesture motion.

Fig. 1 shows an example of the transition of hand motion velocity(horizontal element only) and the vocalization timing for a gesturer where the horizontal and vertical axes indicate "time[unit:seconds]" and "hand motion velocity[unit: pixels/frame]" respectively. In the figure, the hand motion velocity is negative(positive) while the gesturer is moving his/her hand to the right(left). The hatching part indicates the period of vocalization of instruction words such as "KOCCHI(this way)".

Assuming that the transition of the hand motion velocity follows a sin curve, we can normalize the vocalization timing t_S^* within $[0, 1)$. ($0 < t_S^* < 0.5$ if the gesturer is moving his/her hand to the right, $0.5 < t_S^* < 1$ if the gesturer is moving his/her hand to the left, and $t_S^* = 0$ or 0.5 if the hand motion velocity is zero.) For example, see Fig. 1. In this case, the vocalization timing (marked as "VOICE") is about 0.7.

Fig. 2 shows the two histograms of the normalized vocalization timings. The "white" and "black" bars respectively indicate the histogram in case that the human intention is "left" and "right". As shown in this figure, in almost all the cases, the direction of instruction the gesturer wants to give was consistent with the direction of movement of the hand at the moment of vocalization of the instruction word. This result suggests that the vocalization timing correlates closely with the human intention in "natural" gesture motion and basically we can recognize the gesturer's intention only by estimating the hand motion velocity at the moment of the gesturer's vocalization. However, it can be also seen that the white histogram partially overlaps with the black one and vice versa, which means the system based

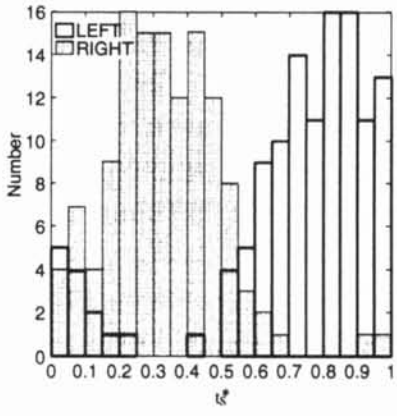


Figure 2: Histogram of vocalization timing.

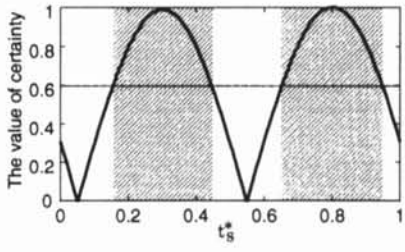


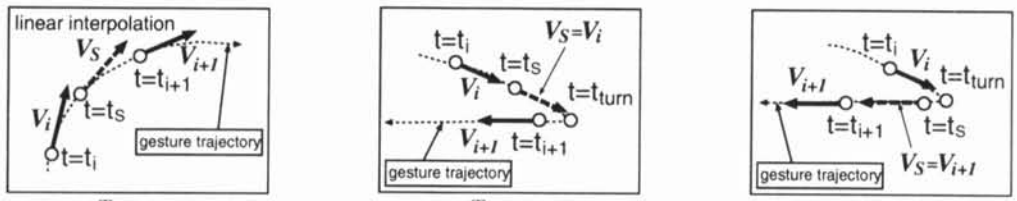
Figure 3: The value of certainty.

on this approach may sometimes mistake the gesturer's instruction. To cope with this "overlapping histogram" problem, we introduce a criterion(the value of certainty) for evaluating the result of recognition.

Definition 1 (value of certainty) *We define the value of certainty C for the normalized vocalization timing t_S^* by*

$$C(t_S^*) = |\sin(2\pi(t_S^* - 0.05))| \quad (1)$$

As shown in Fig. 3, this function is based on the shape of the histograms of vocalization timing(also see Fig. 2). Notice that the phase shift "0.05" in Eq. 1 corresponds to the difference between the peak of the histogram of vocalization timing and the peak of the magnitude of hand motion velocity. Clearly, the domain of C is $[0, 1]$. As can be seen from Fig. 2 and Fig. 3, the larger C is, both the larger the histogram value and the smaller the histogram-overlap probability are. That is, C can be used as a criterion for evaluating the result of recognition: when the gesturer's intention is estimated from the hand motion velocity at the moment of the gesturer's vocalization, the estimation result can be regarded as "reliable" if and only if C is greater than a threshold T. (For example, the hatching parts in Fig. 3 indicate the domain of t_S^* for $T = 0.6$ in which the estimation result is regarded as reliable.) Although the case that the direction of movement of the gesturer's



(a) $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1}) > 0$ (b) $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1}) \leq 0$ and $t_S < t_{turn}$ (c) $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1}) \leq 0$ and $t_S \geq t_{turn}$

Figure 4: Estimation of hand motion velocity at the moment of vocalization.

hand is "left" or "right" was only dealt with so far, it should be noted that the same argument holds true in case of more general cyclic-gestures with arbitrary direction of movement.

As can be seen from Eq. 1, $C(t_S^*) = C(t_S^* + 0.5)$ for arbitrary $t_S^* \in [0, 1)$. Therefore, we can define the *effective* vocalization timing for convenience.

Definition 2 (effective vocalization timing) *We define the effective vocalization timing* $\tilde{t}_S^* \in [0, 0.5)$ *such that* $C(\tilde{t}_S^*) = C(t_S^*)$ *for arbitrary* t_S^* *by*

$$\tilde{t}_S^* = \begin{cases} t_S^* & \text{if } 0 \leq t_S^* < 0.5 \\ t_S^* - 0.5 & \text{if } 0.5 \leq t_S^* < 1 \end{cases}$$

We will use $\tilde{t}_S^*$ instead of t_S^* in the following sections.

3 Gesture Recognition

In this section, we describe how to recognize gestures expressed by cyclically repeated motion of the hand using multi-modal informations.

[Step 1] (Estimation of the direction of the gesturer's instruction) Assume that the hand motion velocity vector can be obtained at the sampling period $\Delta t (> 0)$. Let $\boldsymbol{V}_i$ be the hand motion velocity at time t_i where $i = 0, 1, 2, \ldots$ and $t_{i+1} - t_i = \Delta t$. Also suppose that the vocalization is detected at time $t_S \in [t_i, t_{i+1})$. Now let $\boldsymbol{V}_S$ be the hand motion velocity at the moment of the vocalization. The method for estimating $\boldsymbol{V}_S$ is divided into the following two cases according to whether the inner product of the two vectors $\boldsymbol{V}_i$ and $\boldsymbol{V}_{i+1}$, say $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1})$, is positive or not.

<CASE I> the inner product $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1}) > 0$

In this case, the hand motion velocity $\boldsymbol{V}_S$ is given by simple linear interpolation between $\boldsymbol{V}_i$ and $\boldsymbol{V}_{i+1}$ (see Fig. 4(a)). That is,

$$\boldsymbol{V}_S = \frac{t_{i+1} - t_S}{t_{i+1} - t_i} \boldsymbol{V}_i + \frac{t_S - t_i}{t_{i+1} - t_i} \boldsymbol{V}_{i+1} \tag{2}$$

<CASE II> the inner product $(\boldsymbol{V}_i^T \cdot \boldsymbol{V}_{i+1}) \leq 0$

This means the direction of movement of the hand has rapidly changed during $[t_i, t_{i+1})$. Let t_{turn} be the time when the instantaneous hand motion velocity is zero between t_i and t_{i+1}. The time t_{turn} is estimated using a linear interpolation technique:

$$t_{turn} = \frac{t_{i+1}|\boldsymbol{V}_i| + t_i|\boldsymbol{V}_{i+1}|}{|\boldsymbol{V}_i| + |\boldsymbol{V}_{i+1}|} \tag{3}$$

$\boldsymbol{V}_S$ is then approximated as either $\boldsymbol{V}_i$ or $\boldsymbol{V}_{i+1}$ according to whether or not the vocalization is before the time t_{turn} (see Fig. 4(b),(c)). That is,

$$\boldsymbol{V}_S = \begin{cases} \boldsymbol{V}_i & \text{if } t_S < t_{turn} \\ \boldsymbol{V}_{i+1} & \text{otherwise} \end{cases} \tag{4}$$

Based on the observation described in Section 2, the direction of vector $\boldsymbol{V}_S$ is considered as a *candidate* of the direction of the gesturer's instruction.

[Step 2] (Calculation of the effective vocalization timing) Next, the magnitude of the hand motion velocity $\boldsymbol{V}_S$ is normalized so as to estimate the effective vocalization timing $\tilde{t}_S^*$. Denoting the normalized magnitude of the vector $\boldsymbol{V}_S$ by $V_S^* (0 \leq V_S^* \leq 1)$,

$$V_S^* = \frac{|\boldsymbol{V}_S|}{V_{MAX}} \tag{5}$$

where V_{MAX} indicates the maximum magnitude of the hand motion velocity estimated from the time sequence of the gesturer's hand motion velocity before vocalization, $\{\boldsymbol{V}_i\}; i \in \{i \mid i \text{ is integer and } t_i \leq t_S\}$. As a result, the effective vocalization timing $\tilde{t}_S^* (0 \leq \tilde{t}_S^* < 0.5)$ is given by

$$\tilde{t}_S^* = \frac{\sin^{-1}(V_S^*)}{2\pi} \tag{6}$$

Although two possible solutions may be obtained from Eq. 6, we can always select the appropriate one for $\tilde{t}_S^*$ using the gradient information between the hand motion vectors just before and after vocalization.

[Step 3] (Evaluation of the reliability of the estimation result) Finally, the value of certainty C is obtained by substituting the effective vocalization timing $\tilde{t}_S^*$ into Eq. 1 (by Definition 2, $C(\tilde{t}_S^*) = C(t_S^*)$) and the direction of $\boldsymbol{V}_S$ is "formally accepted" as the direction of the instruction the gesturer wants to give if and only if C is greater than a threshold, say T.

The domain of the threshold T is $[0, 1]$. By definition, the larger T, the smaller the "misrecognized" number but the larger the "unrecognized"(not formally accepted) number. On the other hand, the smaller T, the smaller the unrecognized number but the larger the misrecognized number. Therefore, the threshold T should be carefully selected according to applications. An example for selection of T is shown in the next section.

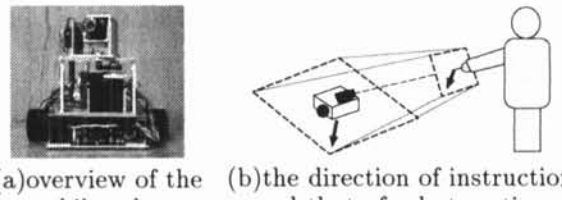
4 A Multi-Modal Interface

The proposed method was applied to mobile robot navigation. As shown in Fig. 5(a), the robot has an active CCD camera that not only tilts up/down but also pans and two wheels that can be driven independently. It is connected with the main computer system(the same as described in Section 2) by wireless. The main computer processes human gesture images from the CCD camera (via transmitter) mounted on the robot as well as the voice input from the microphone worn by the gesturer and decides the corresponding robot motion and the camera pose, and then sends commands to the robot through the microwave link.

Gestures are made on a virtual 2-D plane. The projection of the resulting instruction vector($\boldsymbol{V}_S$) onto the floor plane where the robot moves is associated with the direction of robot motion(see Fig. 5(b)). The robot follows the gesturer's instruction and moves to the instruction direction by a fixed distance. After moving, the robot turns the camera to the gesturer and waits for next instructions. If the instruction direction is ambiguous, that is, the value of certainty is not greater than the threshold T, then the robot does not move and instead requests the gesturer to make a second gesture by shaking its head (the camera platform). If the robot mistakes the gesturer's instruction for another direction, the robot moves to the wrong direction.

We navigated the robot using gestures from a start point to a goal(the same as the start position) via the turn point as shown in Fig. 6. Consequently, our system went well. Only using natural gestures with natural vocalization, the robot was able to be navigated smoothly to the goal while avoiding obstacles. In the experiment, we set the threshold $T = 0.9$. The total number of the gesturer's instructions was 13. The number of unrecognized commands and that of misrecognized commands(during 13 trials) were 3 and 0 respectively.

Next, for comparison, we conducted the same experiment after setting $T = 0.3$. In this case, the total number of the gesturer's trials was 12. The number of unrecognized commands and that of misrecognized commands were 0 and 1 respectively. Notice that the number of gesture commands required to complete the navigation task was less than that in case of $T = 0.9$, while the misrecognized number was not zero. Because T is large it does not necessarily follow that all of the performance is good.



(a)overview of the mobile robot (b)the direction of instruction and that of robot motion

Figure 5: A mobile robot and a gesturer.

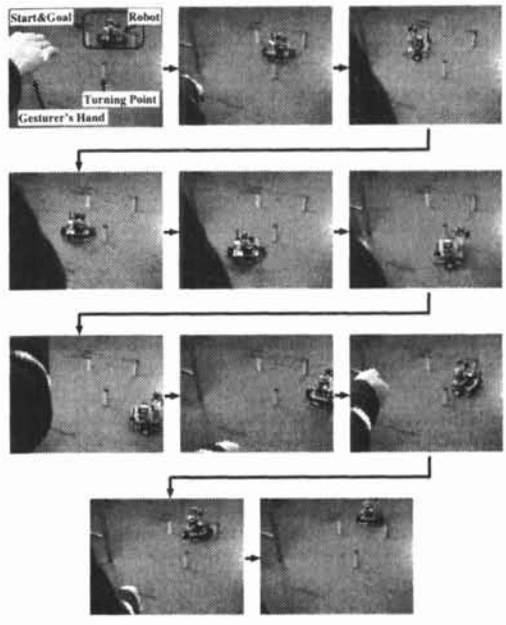


Figure 6: Mobile robot navigation using gestures.

5 Concluding Remarks

We presented a novel method for understanding the human intention expressed by cyclically repeated motion of the hand. Our method is based on the use of a correlation between verbal and non-verbal informations in natural human communication. Now we are introducing voice recognition techniques for the purpose of extending the application range of our system. Our approach opens potential applications to development of truly user-friendly computer/machine systems.

References

[1] R. Cutler and M. Turk: View-based Interpretation of Real-time Optical Flow for Gesture Recognition, *Proc. the 3rd IEEE Int. Conf. on Automatic Face & Gesture Recognition*, pp. 416-421, Nara, Japan, 1998.

[2] A. Nishikawa, A. Ohnishi, and F. Miyazaki: Description and Recognition of Human Gestures Based on the Transition of Curvature from Motion Images, *Proc. the 3rd IEEE Int. Conf. on Automatic Face & Gesture Recognition*, pp. 552-557, Nara, Japan, 1998.

[3] A. Nishikawa, M. Nishimura, A. Hirano, K. Koara, and F. Miyazaki: Systematic Selection of Local Correlation Parameters for Optical Flow-based Gesture Recognition, *Proc. of the 8th IEEE Int. Workshop on Robot & Human Interaction*, pp. 183-188, Pisa, Italy, 1999.

[4] T. Kiriki, Y. Kimuro, and T. Hasegawa: A 4-Legged Mobile Robot Control to Observe a Human Behavior, *Proc. of the 8th IEEE Int. Workshop on Robot & Human Interaction*, pp. 195–200, Pisa, Italy, 1999.

[5] T. Morita, N. Sawasaki, T. Uchiyama, and M. Sato: Color Tracking Vision System, *Proc. the 14th Annual Conf. Robotics Soc. of Japan*, pp. 279-280, 1996 (in Japanese).

[6] Y. Matsumoto, K. Sakai, T. Inamura, M. Inaba, and H. Inoue: PC-based Hypermachine–A Kernel System for Intelligent Robot Application, *Proc. the 15th Annual Conf. Robotics Soc. of Japan*, pp. 979-980, 1997 (in Japanese).