Early Recognition of Gestures

Akihiro Mori, Seiichi Uchida, Ryo Kurazume,
Rin-ichiro Taniguchi, Tsutomu Hasegawa, Hiroaki Sakoe

Kyushu University, mori@human.is.kyushu-u.ac.jp

Abstract This paper is concerned with two topics on gesture recognition. The first topic is early recognition for providing the recognition result of a gesture before the gesture is completed. The second topic is motion prediction for guessing the subsequent posture of the person who makes a gesture. Both topics are mutually related and linked to the realization of proactive human-machine interface. For each of those two topics, a simple technique is developed and examined to reveal its limitation. Possible directions to deal with the limitation are also discussed as the future work on those topics.

1 Introduction

This paper is concerned with two methods, i.e., (i) early recognition of gestures and (ii) gesture prediction. Early recognition is the method to determine the recognition result of a gesture at its beginning part. Gesture prediction is the method to estimate a subsequent posture of a performer.

Early recognition is useful for developing efficient gesture-based man-machine interaction systems because a performer can stop his gesture as soon as its recognition result is determined. Early recognition of gestures have not been investigated so far. In fact, most of conventional methods of gesture recognition provide their recognition results when a gesture is inputted completely.

Gesture prediction is useful for a *proactive* man-machine interaction system, which can start its next action before the performer finishes his gesture. In addition to that, the gesture prediction is useful to minimize the delay between performer's action and the system's reaction. Note that the proposed prediction method is developed from the above early recognition method.

In the papers, the technical details and theoretical limitations of those two methods are discussed along with several experimental results.

2 Early recognition of gesture

2.1 Gesture recognition based on conventional continuous dynamic programming (CDP)

The proposed early recognition method is based on continuous dynamic programming (CDP), which has been successfully used for gesture recognition [1, 2, 3]. CDP can compensate for nonlinear time fluctuations, realize spotting recognition, and perform frame synchronous processing. These features of CDP are very suitable to provide a real-time gesture recognition system.

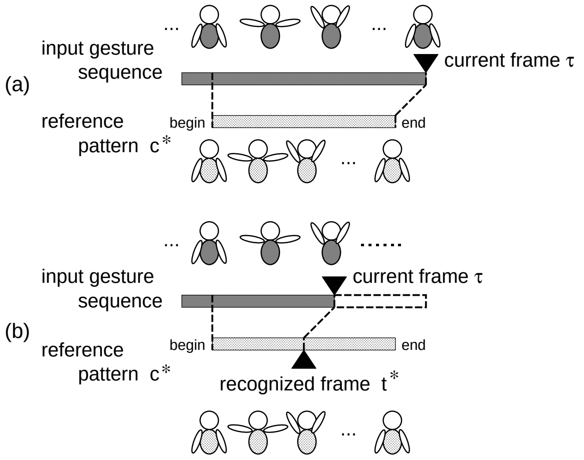

Fig. 1: (a) Conventional gesture recognition method and (b) the proposed early recognition method.

The following is the explanation of a conventional method of gesture recognition based on CDP. Let c denote a gesture category, and feature vector sequence $\boldsymbol{R}_{c,1}, \ldots, \boldsymbol{R}_{c,t}, \ldots, \boldsymbol{R}_{c,T_c}$ represents a registered reference gesture pattern. Each feature vector $\boldsymbol{R}_{c,t}$ represents the posture and the motion of a gesture performer at frame t. In the same way , feature vector sequence $\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_\tau, \ldots$ represents a continuous input pattern to be recognized.

The conventional method computes the following

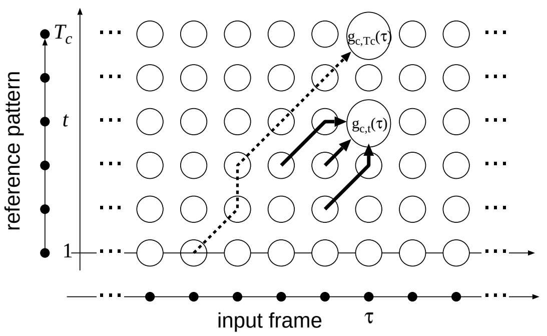

Fig. 2: Gesture recognition based on conventional CDP.

recurrence equation at every input frame τ (Fig. 2);

$$g_{c,t}(\tau) = \min \begin{cases} g_{c,t-1}(\tau-1) + 3d_{c,t}(\tau) \\ g_{c,t-1}(\tau-2) + 2d_{c,t}(\tau-1) + 3d_{c,t}(\tau), \\ g_{c,t-2}(\tau-1) + 3d_{c,t-1}(\tau) + 3d_{c,t}(\tau) \end{cases} \tag{1}$$

where $d_{c,t}(\tau)$ represents local distance between $\boldsymbol{I}_\tau$ and $\boldsymbol{R}_{c,t}$. By computing cumulative distance $g_{c,t}(\tau)$ for all c, t, the recognition result of frame τ is provided as follows;

$$c^* = \operatorname*{argmin}_c g_{c,T_c}(\tau). \tag{2}$$

The above process can realize spotting recognition. That is, recognition results can be provided without segmentation of the input sequence in advance. Since this recurrence equation (1) can be computed in synchronization with input frame τ, the recognition result c^* can be provided at every moment.

Although the conventional CDP works successfully in most cases, it is not suitable for early recognition. This is because the conventional CDP searches an input gesture sequence for a segment similar to the *entire* part of a reference pattern and therefore provides its recognition result after the entire gesture is inputted completely (Fig. 1(a)).

2.2 Early recognition based on CDP

The proposed method can provide the recognition result of a gesture even around the beginning of the gesture. The basic idea of the proposed method is illustrated in Fig. 1(b).

The modification of the proposed method from the conventional CDP for early recognition is rather straightforward. Specifically, the proposed method uses the following discrimination;

$$(c^*, t^*) = \operatorname*{argmin}_{c,t} \left(g_{c,t}(\tau)/t \right). \tag{3}$$

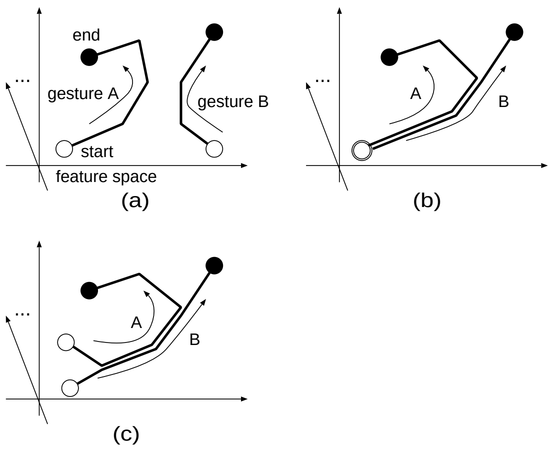

Fig. 3: Relation of two gestures. (a) No common part. (b) Beginning parts are common. (c) Middle parts are common.

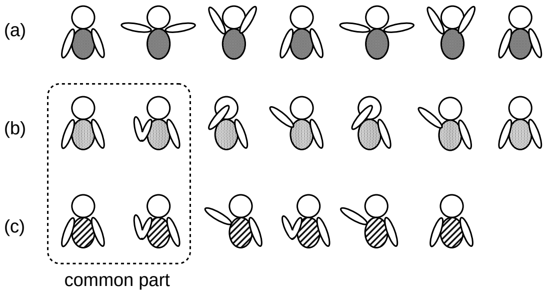

Fig. 4: Three gestures defined in the experiment: (a) “hurrah”, (b) “bye”, and (c) “point”. Note that “bye” and “point” have a common part on their beginnings.

The difference between the discrimination of (2) and (3) is that the beginning part of reference pattern, i.e., $\boldsymbol{R}_{c,1}, \boldsymbol{R}_{c,2}, \ldots, \boldsymbol{R}_{c,t}$ $(t \leq T_c)$. Each beginning part is compared with input pattern after that the beginning part is normalized by its length t. From this discrimination, the system can provide the recognition result that current input τ corresponds with the frame t^* of the reference pattern c^*.

For stabilizing early recognition results, the constraint that $t \geq t_{\min}$ is imposed on the discrimination of (2). This constraint excludes very early recognition results which often become inaccurate. With this constraint, it is impossible to recognize input gesture before the first $t_{\min}/2$ frames.

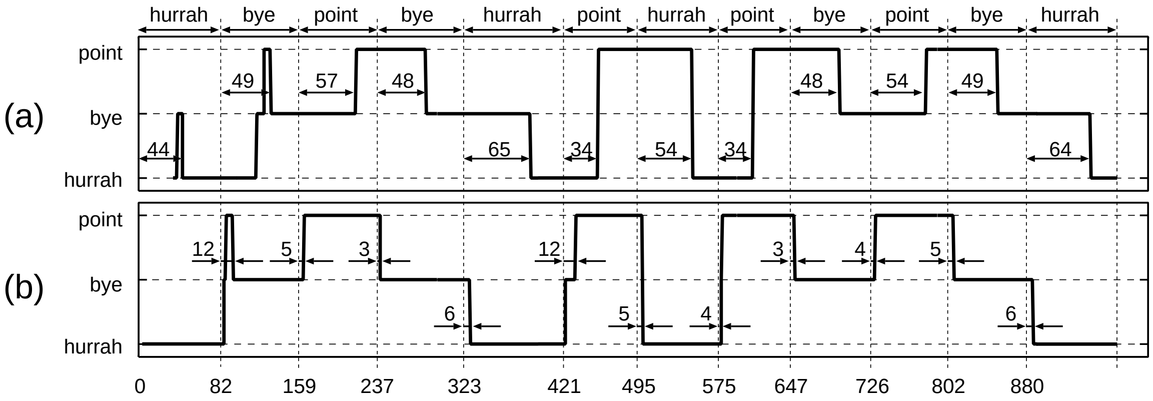

Fig. 5: Recognition results by (a) conventional CDP and (b) proposed early recognition method.

2.3 Theoretical limitation of early recognition

The early recognition has an intrinsic limitation when two or more gestures can not be distinguished from each other due to the ambiguity around their beginning part. Figure 3 illustrates trajectories of the reference patterns of gesture A and gesture B in feature space. The relation of these gestures can be classified to three types as (a)-(c) of this figure. Figure 3(b) is the case which has the limitation. In this case, gesture A and gesture B is not discriminable at their beginning part and therefore the result of early recognition can not be fixed.

2.4 Experimental result

An experiment was conducted to measure how fast the proposed method can provide its recognition result. For this experiment, three gestures are assumed: “hurrah”, “bye”, and “point” (Fig. 4). The gesture “hurrah” is the motion that a performer raises then lowers his both hands, and repeats these motions again. The gesture “bye” is the motion that a performer raises his right hand to his face, shakes it twice from side to side and then lowers it. The gesture “point” is the motion that a performer raises his right hand to his face, moves it back and forth two times, and then lowers it. Thus, beginning parts of “bye” and “point” are the same motion. Note that the performer lowers his both hands at start and end point of these three gestures. For this experiment, 30 patterns (10 patterns for each gesture) of male adult were used. The average frame length of these patterns is about 83.

Each gesture was represented by a sequence of a 12-dimensional feature vector whose elements are the position and the motion of both hands obtained by stereo cameras (Sony, DFW-X700, 15 frames/sec) and skincolor-based tracking. This position of each hand was relative position from position of performer’s head.

Figure 5 shows the recognition results by the conventional method (a) and the proposed early recognition method (b). The input data was obtained by concatenating 12 gestures randomly from above 30 patterns. The remaining 18 patterns were used as the reference pattern. The parameter t_{min} was set 5 in this experiment.

The conventional CDP required about 34 $\sim$ 65 frames for recognition. This delay is caused by the nature of the conventional CDP referred in Section 2.1. In contrast, the proposed method could provide its recognition result within the first 12 frames of gestures. Therefore it can say that the proposed method reduce the delay considerably more than the conventional method. This experimental result implied that the discrimination (3) for early recognition works effectively.

Figure 5 ensures that the ambiguity of gestures degrades the performance of the early recognition as discussed in 2.3. As mentioned previously, the motion that a performer raises his right hand to his face is common to the gesture “bye” and the gesture “point”. The effect of this ambiguity is shown in the recognition result of Fig. 5(b). Specifically, this ambiguity affects the first input of “bye” (12 frame delay) and the second input of “point” (12frame delay); they have larger delay than other input patterns. On the other hand, the input of “hurrah”, which has no common part with other two gestures, can provide the recognition result with a very short delay. Consequently, the fact that the early recognition becomes uncertain by ambiguity like Fig. 3(b) was ensured by this experiment.

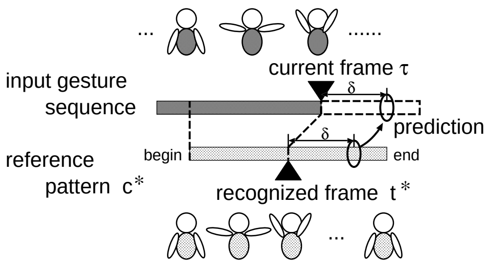

Fig. 6: Gesture prediction based on early recognition.

3 Gesture prediction

3.1 Gesture prediction based on early recognition result

The subsequent posture of the performer can be predicted by simply using the early recognition method of Section 2. Assuming that the current input frame τ is corresponded to the frame t^* of the reference pattern c^* by (3), $\boldsymbol{I}_{\tau+\delta}$ can be predicted as (Fig. 6)

$$\widehat{\boldsymbol{I}}_{\tau+\delta} = \boldsymbol{R}_{c^*,t^*+\delta}. \tag{4}$$

This simple prediction method assumes that the speeds of input and reference gestures are the same. The extension to more general cases that their speeds are different is our future work.

As noted in Section 1, this prediction can be used for compensating the delay of the interaction system; even if the system requires δ frames to react to performer's actions, the delay can be compensated by giving the predicted posture $\widehat{\boldsymbol{I}}_{\tau+\delta}$ instead of $\boldsymbol{I}_\tau$ to the system.

3.2 Theoretical limitation of prediction

In this section, the limitation of the prediction by (4), i.e. the maximum value of δ, is investigated.

In the case of Fig. 3(a),(c), the system can provide the correct prediction result *fully*. At any frame of the gestures of Fig. 3(a),(c), the correct recognition result can be provided, and the subsequent input trajectory can be fixed by the recognition result. Therefore, if (c^*, t^*) is obtained by (3) then the subsequent posture can be predicted as $\boldsymbol{R}_{c^*,t^*}, \boldsymbol{R}_{c^*,t^*+1}, \ldots, \boldsymbol{R}_{c^*,T_c^*}$.

In the case of Fig. 3(b), the system can provide the correct prediction result *partially*. For example, let us consider two gestures "bye" and "point" of Fig. 4. Those gestures have the common part in their beginnings, thus their relation can be illustrated as Fig. 3(b). As discussed before, this common part often induces misrecognition. However, even if the input frame of the beginning part of "bye" are misrecognized as that of "point", it can be predicted that the performer will raise his right hand to his face. That is, the prediction of (4) will be successful if $t^* + \delta$ does not exceed the common part.

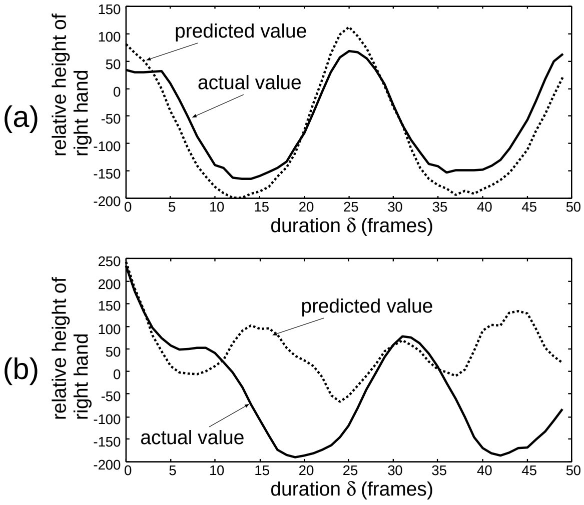

Fig. 7: Prediction results. The relative height of right hand was predicted for various prediction duration δ.

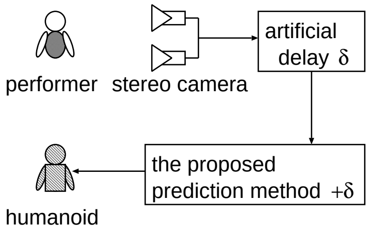

Fig. 8: The experimental scheme to evaluate delay compensation performance.

3.3 Experimental results

Figure 7 shows the prediction result based on (4). Figure 7(a) illustrates predicted data $\widehat{\boldsymbol{I}}_{\tau+\delta}$ at a frame where the correct result of the early recognition (specifically, $\tau = 170$ of Fig. 5(b)) was obtained. In this way, the system predicted subsequent posture correctly for $\delta \sim 50$. On the other hand, Fig. 7(b) illustrates predicted data at a frame ($\tau = 170$ of Fig. 5(b)) where a wrong recognition result ("point" → "bye") was provided because of the ambiguity mentioned in Section 2.4. Note that even in this case, the approximately correct subsequent posture can be provided within the common part of "point" and "bye" ($\delta < 10$).

Another experiment was conducted to ensure the effect of this delay compensation. Figure 8 shows

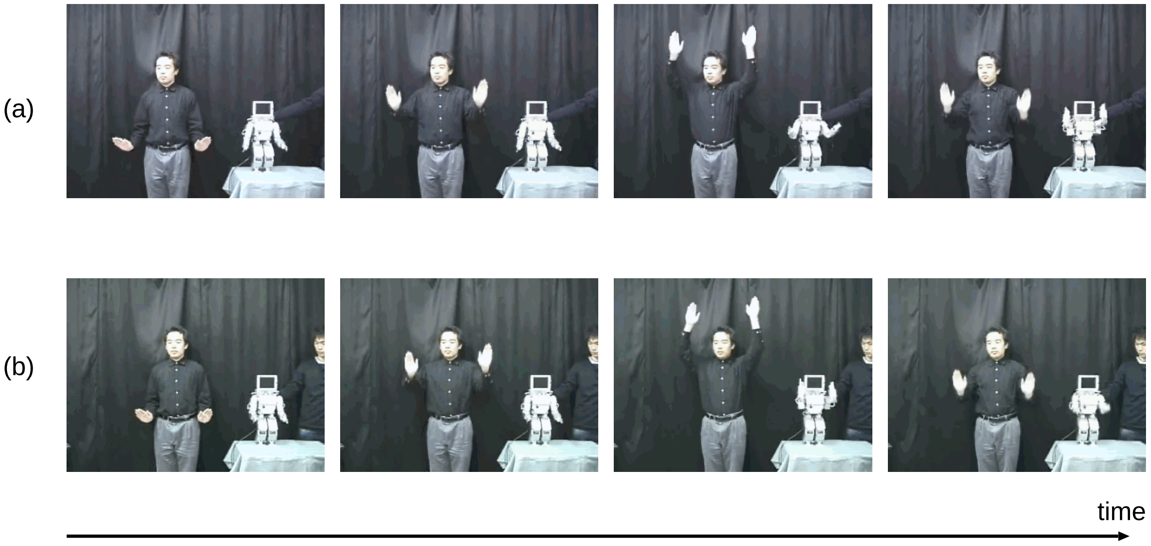

Fig. 9: Effect of delay compensation by the proposed prediction method; the humanoid was driven by (a) delayed input posture and (b) the predicted posture.

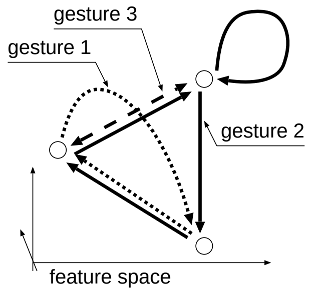

Fig. 10: Network representation of gestures.

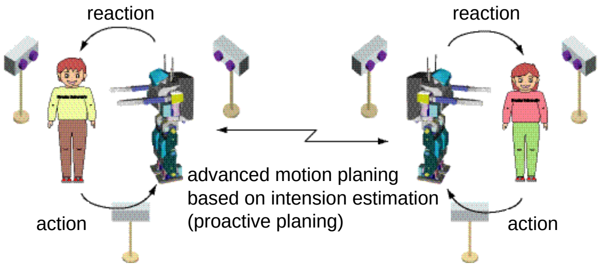

Fig. 11: A proactive man-machine interface for long-distance communication.

the scheme of the experiment. Artificially delayed gestures are firstly inputted to the proposed prediction method. Then the system outputs the predicted posture and a humanoid is driven by this predicted posture. The result shows that the humanoid moved simultaneously to the gesture performer and the effect of proposed prediction method was confirmed.

4 Conclusion

Novel methods for early recognition and prediction of gestures have been proposed in this paper. Experimental results showed that these methods have expected abilities. Theoretical limitation of those methods is also investigated.

Future work will focus on the following points:

- *Construction of gesture network*: As noted in Section 2.3, common parts of gestures should be known to figure out the limitation of the proposed early recognition and prediction methods. Thus, construction of the network represented as Fig. 10 is an important issue.

- *Development of proactive man-machine interface*: Early recognition and delay compensation are useful to realize the humanoid-based man-machine interface for long-distance communication of Fig. 11. That is, the system can continue to move the humanoid by the result of the early recognition even if a performer stops his gesture at its beginning. Furthermore, the system can compensate the delay, which is caused by congestion of communication line or hardware limitation of humanoid. Recall the second experiment of Section 3.3.

Acknowledgment This work was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications in Japan under Strategic Information and Communications R&D Promotion Programme (SCOPE).

References

[1] S. Seki, K. Takahashi and R. Oka: Gesture recognition from motion image by spotting algorithm, Proc. of ACCV1993, vol.2, pp.759-762, 1993.

[2] T. Nishimura, T. Mukai and R. Oka: Non-monotonic Continuous Dynamic Programming for Spotting Recognition of Hesitated Gestures from Time-Varying Images, 3rd ACCV1998, vol.2, pp.734-741, 1997.

[3] T. Nishimura, S. Nozaki, and R. Oka: Spotting Recognition of Gestures by Using a Sequence of Spatially Reduced Range Image, ACCV2000, vol.2, pp.937-942, 2000.

Mori, Akihiro: is an undergraduate student of Dept. of Electrical Engineering and Computer Science, Kyushu University. His research interest includes image processing and gesture recognition.

Uchida, Seiichi: received B.E., M.E., and Dr. Eng. degrees from Kyushu University in 1990, 1992 and 1999, respectively. From 1992 to 1996, he joined SECOM Co., Ltd., Tokyo, Japan where he worked on speech processing. Since 2002, he has been an associate professor at Faculty of Information Science and Electrical Engineering, Kyushu University. His research interests include pattern analysis and speech processing.

Kurazume, Ryo: is an associate professor at Faculty of Information Science and Electrical Engineering, Kyushu University. He received Ph.D. degree from the Department of Mechanical Engineering Science, Tokyo Institute of Technology in 1998. His M.E. and B.E. were from the Department of Mechanical Engineering Science, Tokyo Institute of Technology in 1991 and 1989, respectively. His research interests include multiple mobile robots, computer vision, and walking robots.

Taniguchi, Rin-ichiro: RT received B.E., M.E., PhD in computer science and communication engineering from Kyushu University, Japan in 1978, 1980, 1986 respectively. He became an associate professor of Interdisciplinary Graduate School of Engineering Sciences, Kyushu University in 1989. In 1996, he became a professor of Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests covers image processing, computer vision, human computer interaction and parallel processing.

Hasegawa, Tsutomu: received the B.E. degree in 1973 in electronic engineering and the Ph.D. degree in 1987, both from the Tokyo Institute of Technology, Tokyo, Japan. He was associated with the Electrotechnical Laboratory of the Japanese Government from 1973 to 1992 where he performed research on robotics. From 1981 to 1982, he was a Visiting Researcher at the Laboratoire d'Automatique et d'Analyse des Systemes (LAAS/CNRS), Toulouse, France. He joined Kyushu University, Fukuoka, Japan, in 1992 and is currently a Professor with the Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests are in robotic manipulator control, geometric modeling and reasoning, motion planning, and man-machine interaction.

Sakoe, Hiroaki: received the B.E. degree from Kyushu Institute of Technology in 1966, and the M.E. and D.E. degrees from Kyushu University in 1968 and 1987, respectively. In 1968, he joined NEC Corporation and engaged in speech recognition research. In 1989, he left NEC Corporation to become a Professor of Kyushu University. His research interests include speech recognition and pictorial pattern analysis.