Early Recognition and Prediction of Gestures

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

Kyushu University, Fukuoka, 812-8581, Japan
mori@human.is.kyushu-u.ac.jp

Abstract

This paper is concerned with an early recognition and prediction algorithm of gestures. Early recognition is the algorithm to provide recognition results before input gestures are completed. Motion prediction is the algorithm to predict the subsequent posture of the performer by using early recognition. In addition to them, this paper considers a gesture network for improving the performance of these algorithms. The performance of the proposed algorithm was evaluated by experiments of real-time control of a humanoid by gestures.

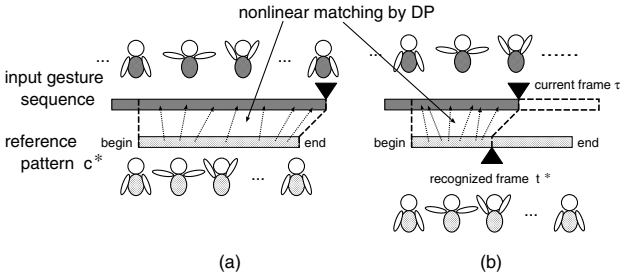

Figure 1. (a) Conventional gesture recognition. (b) The proposed early recognition.

1. Introduction

This paper is concerned with two algorithms, i.e., (i) early recognition of gestures and (ii) motion prediction. The early recognition algorithm determines the recognition result of a gesture at its beginning part. The motion prediction algorithm predicts the subsequent posture of a performer. These algorithms are important to develop practical and intelligent gesture-based man-machine interfaces. For example, the algorithms can be used for compensating delays in the interfaces by driving the interfaces by the predicted posture. In addition, the algorithms provide the basis of "proactive" interfaces where the interfaces react in advance to the end of user's action.

For accurate recognition and prediction, we also define a novel motion primitive as the unit of gesture recognition. This motion primitive is specially designed for accurate and stable recognition/prediction results.

Experiments were performed to evaluate the performance of those algorithms. Technical details and theoretical limitations are also discussed along with experimental results.

2 Conventional Gesture Recognition Algorithm Based on Dynamic Programming

The proposed early recognition algorithm is based on dynamic programming (DP). DP and its stochastic extension called hidden Markov model (HMM) have been successfully used for gesture recognition because they can compensate nonlinear time fluctuations of gestures optimally. Darrell and Pentland [1] have applied DP for recognizing gesture sequences. Oka and his colleagues have proposed a DP-based gesture spotting algorithm called continuous DP algorithm and applied it successfully to gesture recognition [2–4]. Wu and Huang [5] have provided a good review paper which includes HMM-based gesture recognition algorithm.

Let a vector sequence $\boldsymbol{R}_c = \boldsymbol{R}_{c,1}, \ldots, \boldsymbol{R}_{c,t}, \ldots, \boldsymbol{R}_{c,T_c}$ be a registered reference gesture pattern of category c. Each feature vector $\boldsymbol{R}_{c,t}$ represents the posture of a user at frame t. Similarly, a vector sequence $\boldsymbol{I} = \boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_\tau, \ldots$ represents an input gesture sequence comprised of several gestures. A conventional DP algorithm for recognizing the input gesture sequence $\boldsymbol{I}$ is considered as an optimal nonlinear matching algorithm between $\boldsymbol{I}$ and $\boldsymbol{R}_c$. The algorithm is descrived by the following pseudo-code, where $g_{c,t}(\tau)$ is the minimum value of the accumulated distance, (i.e. matching cost) up to frame τ:

Step 1: For $\tau = 1, 2, ...$, repeat Step 2-5.

Step 2: For $c = 1, \ldots, C$, repeat Step 3 and 4.

Step 3: For $t = 1$, calculate the following DP-recurrence equation:

$$g_{c,1}(\tau) = 3d_{c,1}(\tau), \tag{1}$$

where $d_{c,t}(\tau)$ represents the distance between $\boldsymbol{I}_\tau$ and $\boldsymbol{R}_{c,t}$, i.e.,

$$d_{c,t}(\tau) = \|\boldsymbol{I}_\tau - \boldsymbol{R}_{c,t}\|. \tag{2}$$

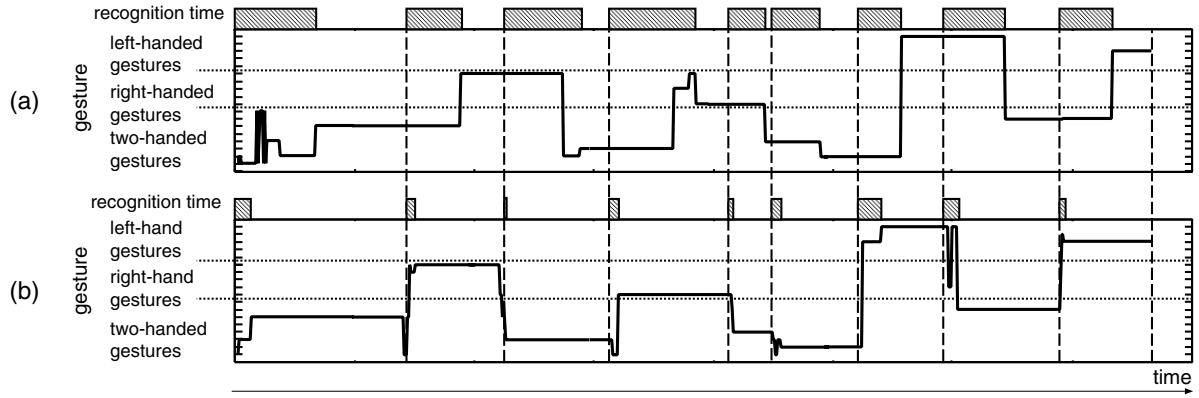

Figure 2. Recognition result of a part of input sequence. (a) Conventional algorithm. (b) Proposed early recognition algorithm. Vertical dashed lines are ground-truthed boundaries of gestures. The width of the hatched rectangular shows the recognition time.

Step 4: For $t = 2, \ldots, T_c$, calculate the following DP-recurrence equation:

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

Step 5: The recognition result at the frame τ is provided as follows:

$$c^* = \underset{c}{\operatorname{argmin}}\, g_{c,T_c}(\tau). \quad (4)$$

Although the above conventional DP algorithm works successfully in most cases, it is not suitable for early recognition. This is because the conventional algorithm 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)). This paper is the first attempt at the early recognition of gestures.

3 Early Recognition Based on DP

3.1 The Algorithm

The proposed early recognition algorithm provides its recognition result of a gesture even at the beginning of the gesture. The proposed algorithm can be derived by the slight modification of the above conventional DP algorithm. Specifically, the proposed early recognition uses the following discrimination rule instead of (4):

$$(c^*, t^*) = \underset{c,t}{\operatorname{argmin}} \left(g_{c,t}(\tau)/t \right). \quad (5)$$

The difference between the discrimination rules of (4) and (5) is that the latter rule allows partial matchings; that is the beginning part of reference pattern $\boldsymbol{R}_{c,1}, \boldsymbol{R}_{c,2}, \ldots, \boldsymbol{R}_{c,t}$ ($t < T_c$) can be a matching candidate in (5) (**Fig. 1**(b)). By the rule (5), it is established that the current input frame τ corresponds to the t^*th frame of the reference pattern c^*.

3.2 Experimental Result

An experiment was conducted to measure how early the proposed algorithm can provide its correct recognition result. The subjected input sequence comprises 54 gestures of 18 categories. Among the 18 categories, 8, 5, and 5 categories are both-handed gestures, right-handed gestures, left-handed gestures, respectively. Each frame is represented as a 6-dimensional feature vector acquired by a stereo motion-capturing system. The average length of these patterns was about 87 frames. For the conventional algorithm, the average and the maximum recognition time were 43.4 and 83 frames, respectively. In contrast, the proposed early recognition could shorten the average and the maximum recognition time successfully to 7.8 and 26 frames, respectively.

The usefulness of the proposed algorithm is also observed in **Fig. 2**. The conventional algorithm required about half length of each gesture at least (i.e., $T_c/2$) to provide correct recognition results. In contrast, the proposed algorithm could provide correct recognition results with far shorter recognition times than the conventional algorithm.

4 Gesture Prediction

4.1 Basic Strategy

The subsequent posture of a performer can be predicted by simply using the result of the early recognition. That is, the subsequent posture after δ frames, i.e., $\boldsymbol{I}_{\tau+\delta}$, can be predicted simply as (**Fig. 3**)

$$\widehat{\boldsymbol{I}}_{\tau+\delta} = \boldsymbol{R}_{c^*,t^*+\delta}, \quad (6)$$

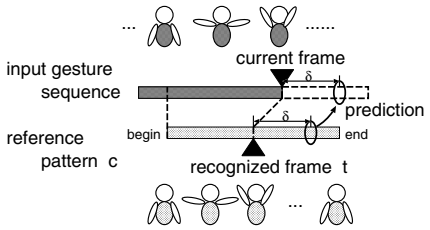

Figure 3. Gesture prediction based on early recognition.

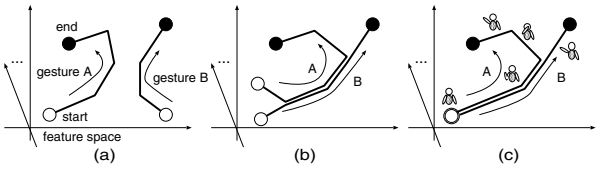

Figure 4. Relation of two gestures. (a) No common part. (b) Middle parts are common. (c) Beginning parts are common.

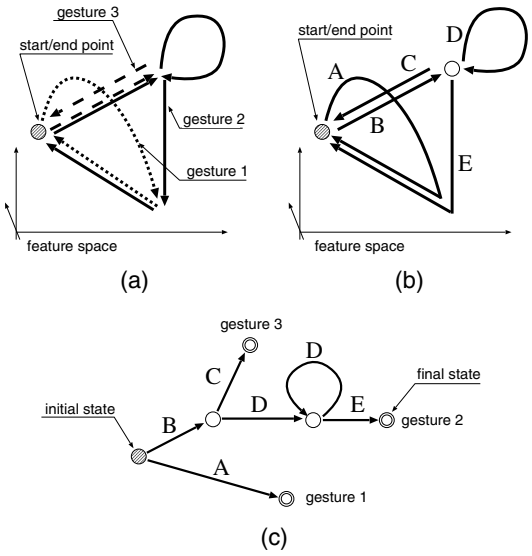

Figure 5. (a) Trajectories of multiple gestures. (b) Unification of common beginning parts. (c) The gesture network. The null-transitions from the final states to the initial state are omitted.

where c^* and t^* are given by (5). Note that this simple prediction algorithm assumes that the speeds of input and reference gestures are the same after the frame τ. The extension for dealing with the speed change after τ can be done by estimating current gesture speed using the optimal path up to τ, although its details are omitted here.

4.2 Theoretical Limitation of Prediction

Early recognition has an intrinsic limitation on its recognition ability. If two or more gestures have a common beginning part, that is, if several gestures are ambiguous in their beginning, we cannot distinguish them at the beginning. Since the subsequent posture prediction is based on the result of early recognition, this ambiguity also degrades the accuracy of the prediction results.

Fig. 4 shows three possible relations between two gestures A and B (depicted as trajectories) in feature space. In the case of **Fig. 4** (a) and (b), there is no common beginning part and thus we can always expect correct early recognition results and correct prediction results. In contrast, in the case of **Fig. 4** (c), we cannot expect correct prediction results if a gesture is misrecognized as another gesture at their common beginning part.

It is important to note that even in the case of **Fig. 4** (c), we can always expect correct prediction results *within* the common beginning part. That is, unless the prediction frame $t^* + \delta$ does not exceed the end of the common part, we can expect a correct prediction. The fact that the common part defines a predictable range provides the basis of the following discussion of gesture network.

4.3 Gesture Network and Motion Primitives

The gesture network proposed in this section provides a useful representation of the predictable range from each frame. **Fig.5** shows the procedure to create a gesture network. Among three gestures shown in **Fig.5** (a), gesture 2 and 3 have a common beginning part. This common beginning part is unified as **Fig. 5** (b), whereas common middle/ending parts are not. Finally, the gesture network of **Fig. 5** (c) is directly obtained from the unified trajectories of **Fig. 5** (b).

Each edge of the gesture network determines the predictable range. For example, the edge B in **Fig. 5** (c) is the common beginning part of gestures 2 and 3 and we cannot predict which is the subsequent gesture, C or D. Thus the end of the edge B is the limit of the predictable range for the frames on B. Gesture 1 has no partner having a common beginning part and thus forms a single edge A by itself.

Now we define *a motion primitive* as an edge of the gesture network and use it as a reference pattern $\boldsymbol{R}_c$ in the proposed early recognition and prediction algorithms. Our motion primitive is somewhat peculiar because it is specialized for prediction, while there are many definitions of motion primitives (e.g., [6, 7].) The use of the motion primitive as a reference pattern can avoid the misrecognition due to the ambiguity and thus avoid serious prediction error. Furthermore, since we can grasp the limitation of predictable range explicitly, we can have the reliability of the current predic-

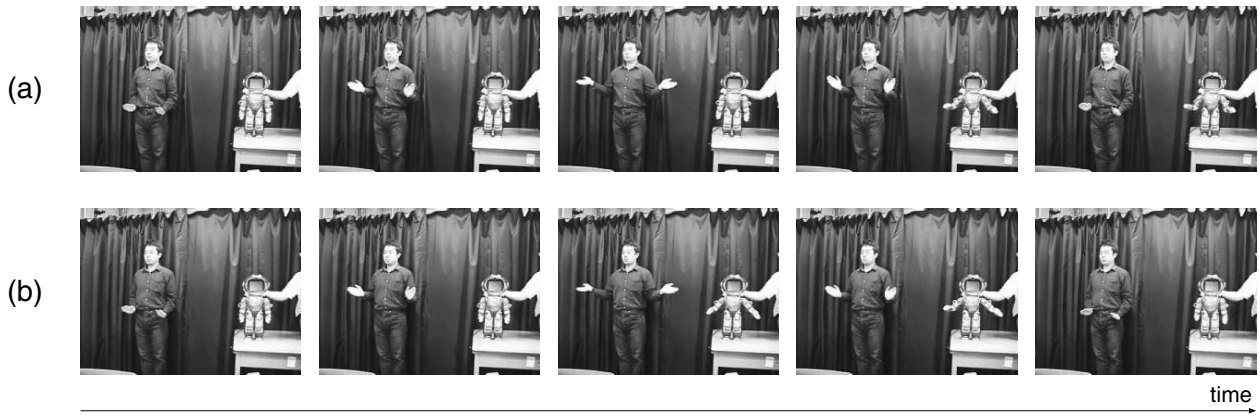

Figure 6. Result of delay compensation by the proposed prediction algorithm; the humanoid was driven by (a) delayed input posture and (b) the predicted posture.

tion result. Note that even in this case, the early recognition algorithm of Section 3 can be used with a slight modification. Also note if a posture beyond the end of a motion primitive should be predicted (i.e. if $t^* + \delta$ exceeds the frame length of the motion primitive), the motion primitives which can follow the present motion primitive are averaged as a tentative trajectory and the predicted posture is given as a point on the trajectory.

4.4 Experimental Results

The experiment was conducted to evaluate the proposed prediction algorithm on compensating delays in a man-machine interface. Specifically, a gesture sequence with an artificial delay of 1 sec was firstly inputted to the proposed early recognition / prediction algorithm. Then the predicted posture was provided by the prediction algorithm (6) with $\delta = 1\text{sec} = 15\text{frames}$. Finally, a humanoid was actuated with this predicted posture. If an accurate prediction is provided, it is expected that the user and the humanoid get synchronously.

The 18 gesture categories of Section 3.2 were assumed. Then, a gesture network comprised of 31 motion primitives were created manually by considering the fact that the 15 categories have common parts in their beginnings.

Fig. 6 shows the snap-shots of delay compensation results around a gesture input of "shrug". In **Fig. 6** (a), no delay compensation was performed and thus a gap between user's action and humanoid's reaction (i.e., imitation) is observed. In contrast, in **Fig. 6** (b) where the delay of 1 sec was compensated by the proposed algorithm, the gap was clearly minimized and thus the action of the user and the reaction of humanoid were almost synchronized.

5 Conclusion

The main contributions of this paper are twofold: (i) the proposal of an early recognition algorithm and (ii) the proposal of a gesture prediction algorithm. These two algorithms are closely related because the result of the former algorithm is used in the latter algorithm. In addition, a novel motion primitive was introduced to deal with the theoretical limitation of the gesture prediction.

The performance of the proposed algorithms were evaluated via experiments on actual gesture sequences. Furthermore, a delay compensation experiment with a humanoid was performed and its result showed that the proposed algorithms could compensate the delay of 1 sec.

Acknowledgement 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] T. Darrell and A. Pentland, "Space-time gestures," Proc. CVPR, pp. 335-340, 1993.

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

[3] T. Nishimura, T. Mukai and R. Oka, "Non-monotonic continuous dynamic programming for spotting recognition of hesitated gestures from time-varying images," Proc. ACCV, vol. 2, pp. 734-741, 1998.

[4] T. Nishimura, S. Nozaki, and R. Oka, "Spotting recognition of gestures by using a sequence of spatially reduced range image," Proc. ACCV, vol. 2, pp. 937-942, 2000.

[5] Y. Wu and T. S. Huang, "Vision-based gesture recognition: a review," Lecture Notes in Computer Science, vol. 1739, pp. 103–114, 1999.

[6] T.D. Sanger, "Optimal movement primitives," Advances in Neural Information Processing Systems, vol. 7, pp. 1023-1030, 1995.

[7] A.Fod, M.J. Mataric, and O.C. Jenkins, "Automated deviation of primitives for movement classification," Autonomous Robots, vol. 12, no. 1, pp. 39-54, 2002.