Early Recognition Based on Co-occurrence of Gesture Patterns

Atsushi Shimada, Manabu Kawashima, and Rin-ichiro Taniguchi

Department of Advanced Information Technology, Kyushu University
744 Motooka, Nishi-ku, Fukuoka, Japan
{atsushi,kawashima,rin}@limu.ait.kyushu-u.ac.jp
http://limu.ait.kyushu-u.ac.jp/

Abstract. We propose an approach to achieve early recognition of gesture patterns. We assume that there are two people who interact with a machine, a robot or something. In such a situation, a gesture of a person often has a relationship with a gesture of another person. We exploit such a relationship to realize early recognition of gesture patterns. Early recognition is a method to recognize sequential patterns at their beginning parts. Therefore, in the case of gesture recognition, we can get a recognition result of human gestures before the gestures have finished. Recent years, some approaches have been proposed. In this paper, we expand the application range of early recognition to multiple people based on the co-occurrence of gesture patterns. In our approach, we use Self-Organizing Map to represent gesture patterns of each person, and associative memory based approach learns the relationship between co-occurring gestures. In the experiments, we have found that our proposed method achieved the early recognition more accurately and earlier than the traditional approach.

Keywords: Gesture Recognition, Early Recognition, Co-occurring Gesture, Self-Organizing Map.

1 Introduction

A man-machine seamless interaction is an important tool for various interactive systems such as virtual reality systems, video game consoles, human-robot communication, and so on[5,6]. To realize such a interaction, the system has to estimate human gestures in real-time. Generally, a gesture recognition result is acquired after the gesture has finished. Therefore, if a long gesture is observed, we have to wait for the response until the recognition result is determined. This is a problem to realize a "real-time" man-machine interaction.

Recent years, a new approach called "early recognition" has been proposed for gesture recognition[4,8,1]. The early recognition means that a system outputs a recognition result before a gesture has finished. It is a very useful technique to realize a real-time interaction. The most difficult problem of early recognition is that when the system determines the recognition result. In other words, the system has to ensure the recognition result before the observing gesture has finished. Most traditional approaches suffer from this problem since the gestures

K.W. Wong et al. (Eds.): ICONIP 2010, Part II, LNCS 6444, pp. 431–438, 2010.

comprehend ambiguity. Especially at the beginning part of them, it is very difficult to determinate the recognition result since enough input data has not been observed yet. To solve this problem, we propose a new approach. The biggest difference between traditional approaches and our approach is that we target not only an individual person but also two or more people in the environment. It means that there are two or more people who interact with a machine, a robot, or so on, simultaneously. In such a situation, a gesture of a person is often related to a gesture of another person. We call such a relationship "co-occurring gesture", and we use the information of co-occurrence for realizing early recognition.

Our approach uses Self-Organizing Map (SOM) and its sparse codes to represent gesture patterns. This approach is based on the approach proposed by Shimada *et al.*[7]. In this research, we have modified their approach to adapt for co-occurring gesture recognition. In addition, we introduce an associative memory to describe a relationship between co-occurring gestures.

2 Definition of Early Recognition of Gesture Patterns

In this section, we give conceptual explanation about early recognition of individual gesture and co-occurring gesture.

2.1 Typical Gesture Recognition

Let $\boldsymbol{C}^i = \{c_1^i, \ldots, c_n^i\}$ be a training gesture pattern which belongs to gesture class $i \in \boldsymbol{L}$. The $\boldsymbol{L}$ is a set of class labels. A gesture can be represented in a sequential n-long posture patterns. Therefore, c_n^i means the n-th posture of the gesture. When an unknown gesture $\boldsymbol{X} = \{x_1 \ldots, x_l\}$ is observed, the typical gesture recognition problem is to find the most similar gesture from training patterns by

$$p = \operatorname*{argmin}_i \{f(\boldsymbol{X}, \boldsymbol{C}^i)\} \tag{1}$$

where p is the class label and $f()$ is a distance function which evaluate the similarity between the gesture pattern $\boldsymbol{X}$ and $\boldsymbol{C}^i$.

2.2 Early Recognition of Individual Gesture Patterns

The key issue of early recognition is to output a recognition result before acquiring complete input pattern. In the case of gesture recognition, especially individual gesture patterns, it corresponds to the following problem. When a part of gesture pattern (unfinished gesture) $\boldsymbol{X}^{'} = \{x_1 \ldots, x_k\}, (k < l)$ is observed, the recognition result is determined by

$$p = \operatorname*{argmin}_i \{f(\boldsymbol{X}^{'}, \boldsymbol{C}^i) < TH_I\} \tag{2}$$

where TH_I is a threshold of distance which adjusts the timing of recognition result. If the threshold is not introduced, a recognition result will be output without concrete proof. Therefore, we set a threshold to ensure reliability for the recognition result.

2.3 Early Recognition of Co-occurring Gesture Patterns

Unlike the other recognition strategies mentioned above, the system has to observe gestures of two people simultaneously. Let $\boldsymbol{Y}^{'}$ be a gesture pattern of another person (the gesture has not finished yet). The output of the early recognition of co-occurring gesture patterns can be defined as follows.

$$(p,q) = \underset{(i,j)\in \boldsymbol{M}}{\operatorname{argmin}} \{f(\boldsymbol{X}^{'}, \boldsymbol{C}_A^i) + f(\boldsymbol{Y}^{'}, \boldsymbol{C}_B^j) < TH_C\} \tag{3}$$

where $\boldsymbol{M}$ is a subset of $\boldsymbol{L} \times \boldsymbol{L}$. The subscripts of $\boldsymbol{C}$ denote the person labels, i.e, person A and person B. Note that the $\boldsymbol{L} \times \boldsymbol{L}$ is a set of all combination of co-occurring gestures. Actually, the combination is restricted by the application, environment or so on, and the co-occurrence is not always the all combination of gestures. This is why we introduce the subset $\boldsymbol{M}$. Fig. 1 shows an example of the relationship between $\boldsymbol{M}$ and $\boldsymbol{L}\times\boldsymbol{L}$. The rows denote the label of gesture of person A, and the columns denote the one of person B. The "circle mark" indicates that the corresponding gestures will be observed simultaneously. In the case of Fig. 1(a), all gestures of person A would be observed at the same time with the all gestures of person B. On the other hand, in the case of Fig. 1(b), some cells are "blank", which means that such co-occurrence will not be observed. Therefore, the possible co-occurring gestures between person A and B are a subset of all combination $\boldsymbol{L} \times \boldsymbol{L}$, i.e., $\boldsymbol{M} = \{(g1, g1), (g1, g2), (g2, g1), (g2, g3), (g3, g1)\}$.

		Person B		
		g1	g2	g3
Person A	g1	●	●	●
	g2	●	●	●
	g3	●	●	●

(a) $\boldsymbol{M} = \boldsymbol{L} \times \boldsymbol{L}$

		Person B		
		g1	g2	g3
Person A	g1	●	●	
	g2	●		●
	g3	●		

(b) $\boldsymbol{M} \subset \boldsymbol{L} \times \boldsymbol{L}$

Fig. 1. Description of a set of co-occurring gesture $\boldsymbol{M}$. (a):all possible combination of co-occurring gestures, (b):a subset of all combinations.

The TH_C is a threshold which controls the timing of early gesture recognition. The difference between Eq. 2 and Eq. 3 is that the latter determines the output timing based on two distance functions, i.e, $f(\boldsymbol{X}^{'}, \boldsymbol{C}_A^i)$ and $f(\boldsymbol{Y}^{'}, \boldsymbol{C}_B^j)$. Therefore, even if the system does not have high confidence in one person's gesture recognition, it can output the result when another person's gesture is recognized with higher confidence (i.e, very smaller distance of $f(\boldsymbol{Y}^{'}, \boldsymbol{C}_B^j)$).

3 Early Recognition Strategy

3.1 System Overview

First of all, we show the system overview in Fig. 2. The process can be divided into two phases; training phase and test phase. In the training phase, Self-Organizing Map (SOM) is used to learn postures, which are elements of all gestures. The advantages of using SOM are 1) to reduce dimensionality of

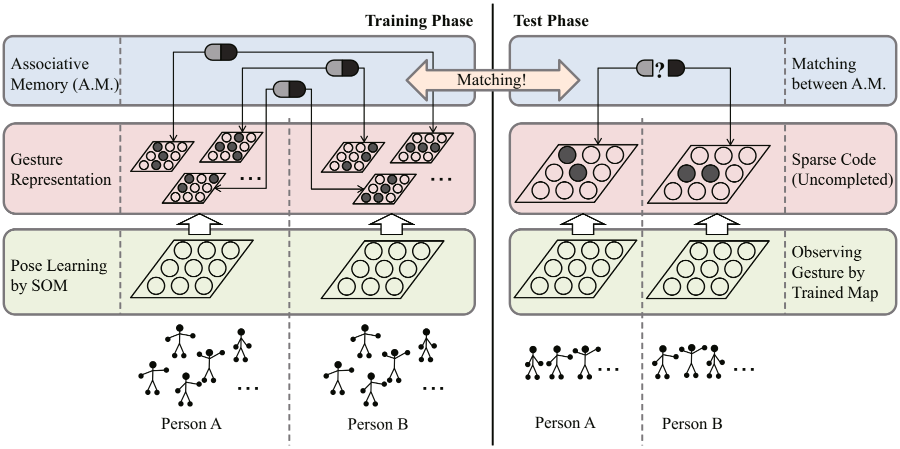


Fig. 2. Processing Flow of Training/Recognition of Co-occurring Gesture Patterns

gesture patterns, 2) to reduce some redundant postures, 3) to represent a gesture pattern by combination of smaller number of neurons and so on. Due to space limitation, we skip the detailed explanation about SOM and how to learn the postures (refer to the literature [7] for detail). After the training of all postures, element postures of each gesture are input to the map again. And then, we can get a “Sparse Code” which represent a gesture pattern on the SOM (see section 3.2). Finally, in the training phase, we associate one person’s gesture pattern (sparse code) with another person’s gesture pattern based on teacher signals given by the relationship as shown in Fig. 1. In this way, all possible co-occurrence gestures are associated by “Associative Memory”(see section 3.3).

In the test phase, the system observes two people’s gesture simultaneously. Then, each person’s parse code is generated/updated immediately whenever a new observation is acquired. Finally, two sparse codes (person A and person B) are examined whether or not they are co-occurring gesture by referring to associative memory acquired in the training phase. Actually, the examination is achieved by measuring the distance between sparse codes(see section 3.4).

3.2 Sparse Coding

When a posture x_k is input to the SOM, one neuron will be selected as winner. When a set of postures which consist of a gesture is sequentially input to the SOM, some neurons will be activated. We regard such an activation pattern as “Sparse Code”, which represents an input gesture. Here, we define the notation of a sparse code. Let $\boldsymbol{S}$ be a sparse code which means a set of activated neuron s. In the training phase, all training gestures are represented by using sparse code $\boldsymbol{C}^i$. Meanwhile, in the test phase, a sparse code of observing gesture is represented by $\boldsymbol{X}^{'}$ or $\boldsymbol{Y}^{'}$, which corresponds to Eq. 3.

Note that the sparse code described here has not an ability to distinguish the gesture patterns whose elements are the same but the sequences are different.

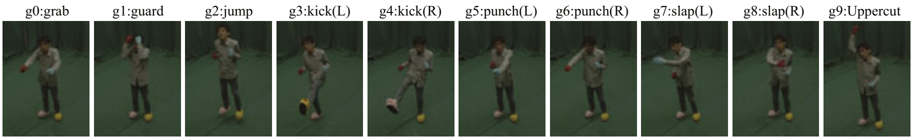


Fig. 3. Gestures used in our experiments

However, we can easily improve introduce temporal information into the system by our previous study[7].

3.3 Associative Memory

To realize early recognition of co-occurring gesture patterns, we introduce an associative memory. After the training of all gestures(actually, the training of sequential postures by SOM), we get sparse code $\boldsymbol{C}_A^i$ and $\boldsymbol{C}_B^j$ which co-occur with each other. The combination of i and j is restricted by $(i, j) \in \boldsymbol{M}$ which is defined by application(see section 4 for our configuration). Our system memorize these relationships between co-occurring gestures as "associative memory". In other words, the system has several combinations between $\boldsymbol{C}_A^i$ and $\boldsymbol{C}_B^j$, which will be observed as co-occurring gestures. Actually, in our implementation, we stored each pair of sparse codes, which indicates the list of winner neurons' indices for the corresponding gesture, in the memory storage of the computer and used them as associative memory.

3.4 Similarity Measure

The number of elements in sparse code $\boldsymbol{S}$ is different from each other since the number of activated neurons depend on the gesture length and the gesture pattern. Therefore, we introduce the Hausdorff distance to measure the similarity between two sets of sparse code. Let X and Y be two non-empty subsets of a metric space. The Hausdorff distance $f(X, Y)$, which corresponds to the distance function in Eq. 1, 2 and 3, is defined by

$$f(X, Y) = \max_i \{ \sup_{x \in \boldsymbol{X}} \inf_{y \in \boldsymbol{Y}} d(x, y), \sup_{y \in \boldsymbol{Y}} \inf_{x \in \boldsymbol{X}} d(x, y) \} \qquad (4)$$

where $d(x, y)$ is the distance function. In our research, we use L2-distance between the coordinates of activated neurons.

4 Experimental Results

4.1 Condition

We demonstrate proposed early recognition of gesture patterns using motion data prepared by ourselves. Each gesture consists of a sequence of postures, and each posture is represented by 5 measured markers. Each marker is composed

		Person B									
		g0	g1	g2	g3	g4	g5	g6	g7	g8	g9
Person A	g0						●				
	g1							●			
	g2								●		
	g3									●	
	g4										●
	g5	●									
	g6		●								
	g7			●							
	g8				●						
	g9					●					

(a) case 1: one-to-one

		Person B									
		g0	g1	g2	g3	g4	g5	g6	g7	g8	g9
Person A	g0			●							
	g1				●	●	●	●			●
	g2	●									
	g3		●		●	●					
	g4		●		●	●					
	g5		●				●	●			
	g6		●				●	●			
	g7									●	
	g8								●		
	g9		●								

(b) case 2: one-to-many

Fig. 4. Configuration of co-occurring gestures

of data of (x, y, z)-axis. We prepared 10 kinds of gesture patterns ($|\boldsymbol{L}| = 10$, see Fig. 3) from 7 examinees. Each person did each gesture 40 times. We used 20 patterns for training and other 20 patterns for test. We conducted the experiment through cross-validation among examinees.

The co-occurring gestures used in the experiments are shown in Fig. 4. The Fig. 4(a) shows the simple configuration that each gesture of person A corresponds to unique gesture of person B. Meanwhile, in the case of Fig. 4(b), the problem becomes more difficult since there are some gesture candidates(one-to-five correspondence at maximum) which occurs at the same time. We can investigate how the co-occurrence information is effective and helpful to determine the recognition result.

4.2 Early Recognition Result of Individual Gesture Patterns

Fig. 5 shows the result of early recognition for an individual person. The horizontial axis denotes the complete ratio of observing gesture pattern, and the vertical axis denotes the recognition accuracy. The bold curve indicates the average ratio of accuracy. For example, the recognition ratio exceeded 90% when more than 50% long gestures had been observed on average. We regards this result as baseline in the following experiments.

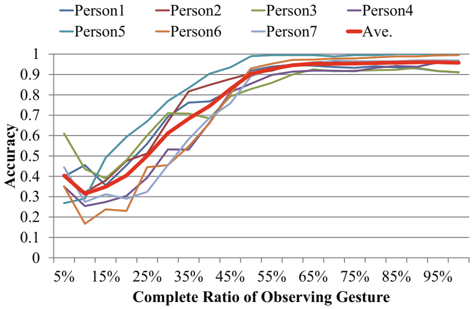
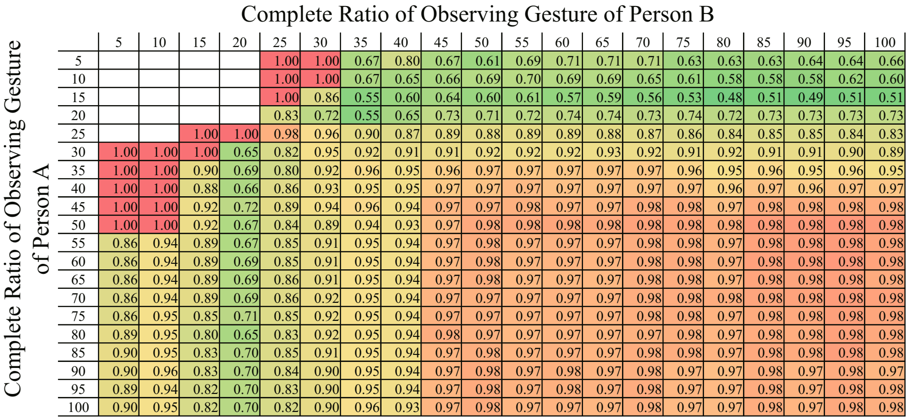


Fig. 5. Early Recognition Result of Individual Gesture Patterns

4.3 Early Recognition Result of Co-occurring Gesture Patterns

As mentioned above, we investigated the recognition accuracy of co-occurring gestures under two conditions(see Fig. 4). First, we examined the case 1 in



Complete Ratio of Observing Gesture of Person B

Complete Ratio of Observing Gesture of Person A

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5					1.00	1.00	0.67	0.80	0.67	0.61	0.69	0.71	0.71	0.71	0.63	0.63	0.63	0.64	0.64	0.66
10					1.00	1.00	0.67	0.65	0.66	0.69	0.70	0.69	0.69	0.65	0.61	0.58	0.58	0.58	0.62	0.60
15					1.00	0.86	0.55	0.60	0.64	0.60	0.61	0.57	0.59	0.56	0.53	0.48	0.51	0.49	0.51	0.51
20					0.83	0.72	0.55	0.65	0.73	0.71	0.72	0.74	0.74	0.73	0.74	0.72	0.73	0.73	0.73	0.73
25			1.00	1.00	0.98	0.96	0.90	0.87	0.89	0.88	0.89	0.89	0.88	0.87	0.86	0.84	0.85	0.85	0.84	0.83
30	1.00	1.00	1.00	0.65	0.82	0.95	0.92	0.91	0.91	0.92	0.92	0.92	0.93	0.92	0.91	0.92	0.91	0.91	0.90	0.89
35	1.00	1.00	0.90	0.69	0.80	0.92	0.96	0.95	0.96	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96	0.95	0.96	0.95
40	1.00	1.00	0.88	0.66	0.86	0.93	0.95	0.95	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.96	0.97	0.97
45	1.00	1.00	0.92	0.72	0.89	0.94	0.96	0.94	0.97	0.97	0.98	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
50	1.00	1.00	0.92	0.67	0.84	0.89	0.94	0.93	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
55	0.86	0.94	0.89	0.67	0.85	0.91	0.95	0.94	0.97	0.98	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
60	0.86	0.94	0.89	0.69	0.85	0.91	0.95	0.94	0.97	0.98	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
65	0.86	0.94	0.89	0.69	0.86	0.91	0.95	0.94	0.97	0.98	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
70	0.86	0.94	0.89	0.69	0.86	0.92	0.95	0.94	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
75	0.86	0.95	0.85	0.71	0.85	0.92	0.95	0.94	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
80	0.89	0.95	0.80	0.65	0.83	0.92	0.95	0.94	0.98	0.97	0.97	0.97	0.97	0.97	0.98	0.97	0.98	0.98	0.98	0.98
85	0.90	0.95	0.83	0.70	0.85	0.91	0.95	0.94	0.97	0.98	0.97	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.98	0.98
90	0.90	0.96	0.83	0.70	0.84	0.90	0.95	0.94	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.97	0.98	0.97	0.98	0.97
95	0.89	0.94	0.82	0.70	0.83	0.90	0.95	0.94	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.97	0.98	0.97	0.98	0.98
100	0.90	0.95	0.82	0.70	0.82	0.90	0.96	0.93	0.97	0.98	0.97	0.97	0.97	0.98	0.97	0.97	0.98	0.97	0.98	0.98

Fig. 6. Early Recognition Result of Co-occurring Gesture Patterns

Fig. 4(a). We got about 100% accuracy when two people's gesture had been observed at least 15%. Due to limitations of space, we left out the detailed result here, but we consider that we could get such good results because of the simple co-occurrence rule between two people's gesture patterns. In other words, the system could narrow the recognition result easily with the help of co-occurrence information.

Second, we examined the case 2 in Fig. 4(b). The early recognition results are shown in Fig. 6. The vertical line shows the complete ratio of observing gesture patterns of person A and the horizontal line shows the one of person B. Each cell in the figure shows each recognition result. The blank cell denotes that the system didn't output the recognition result because the condition of early recognition was not satisfied in Eq. 3. We gave each cell a color based on the accuracy ratio. The red or yellow cell means a good result which exceeds 95% accuracy. Most cells have a red-like or yellow-like color, which indicates higher accuracy. For example, it was enough for the system to determine the recognition result when the one person's gesture patterns had been observed at least 25%.

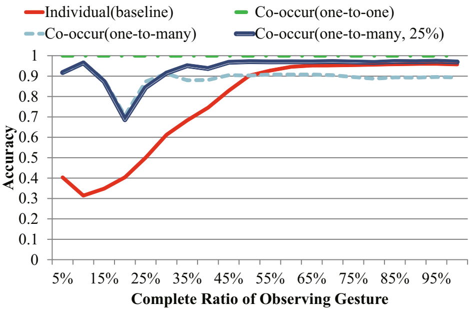


Fig. 7. Comparison between Individual and Co-occurrence Recognition

To compare the early recognition accuracy between individual and co-occurrence, we draw the accuracy curve as shown in Fig. 7. The bold red curve is referred from the average accuracy in Fig. 5. In fact, though there are two complete ratio axes in the case of co-occurring gesture patterns(i.e., person A and person B), we marginalized with respect to one person's complete ratio to represent in the same format with Fig. 5. We can see that the system achieved gesture recognition much earlier than the baseline method. In particular, if one person's gesture had been observed at least 25%, the performance was very high compared with the baseline method(see the bold blue line in Fig. 7).

5 Conclusion

We have proposed a new approach for early recognition of gesture patterns which targets two people. When there is co-occurrence of gestures between two people, the system can recognize the recognition result using its co-occurring information. We have developed prototype of early recognition system using SOM and the associative memory. Through experiments, we confirmed that our proposed method performs well.

In a future work, we are going to use our proposed early recognition system for actual man-machine interaction and investigate its effectiveness. Before that, we will conduct further experiments; increasing the number of gesture classes, the number of people and so on.

References

1. Kawashima, M., Shimada, A., Taniguchi, R.: Early recognition of gesture patterns using sparse code of self-organizing map. In: 7th International Workshop On Self-Organizing Maps, pp. 116–123 (June 2009)
2. Kohonen, T.: Self-Organization and Associative Memory. Springer, Heidelberg (1989)
3. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Science (1995)
4. Mori, A., Uchida, S., Kurazume, R., Taniguchi, R., Hasegawa, T., Sakoe, H.: Early recognition and prediction of gestures. In: Proc. of International Conference on Pattern Recognition, vol. 3, pp. 560–563 (2006)
5. Park, H.S., Jung, D.J., Kim, H.J.: Vision-based game interface using human gesture. In: Chang, L.-W., Lie, W.-N. (eds.) PSIVT 2006. LNCS, vol. 4319, pp. 662–671. Springer, Heidelberg (2006)
6. Park, J., Yi, J.: Gesture recognition based interactive boxing game. International Journal of Information Technology 12, 36–44 (2006)
7. Shimada, A., Taniguchi, R.: Gesture recognition using sparse code of hierarchical som. In: Proc. of International Conference on Pattern Recognition (2008)
8. Uchida, S., Amamoto, K.: Early recognition of sequential patterns by classifier combination. In: Proc. of International Conference on Pattern Recognition (2008)