Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

NEUROCOMPUTING

ELSEVIER

Neurocomputing 51 (2003) 213–224

www.elsevier.com/locate/neucom

Object segmentation using maximum neural networks for the gesture recognition system

Noriko Yoshiike*, Yoshiyasu Takefuji

Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa 2520816, Japan

Received 3 July 2001; accepted 8 May 2002

Abstract

In this paper, we present a new clustering method for segmentations of moving target and non-target objects. We assume that the moving target object has the following conditions: (1) object motion data continuity inter-frame, and (2) object motion data continuity intra-frame. In our model, clusters tend to form as filling these two conditions. The experimental results showed the effectiveness of the proposed algorithm and the performance of this model in terms of the quality of the recognition results. Our algorithm is able to clean the input noise by removing non-target objects before the recognition process.

Keywords: Gesture recognition; Object segmentation; Maximum neural networks

1. Introduction

The gesture recognition system is one of the most important applications translating human communication media to symbols that are understandable for the computer. While many kinds of informational household appliances have been researched for developing in recent years, there are many problems in the human and computer interfaces. For handicapped people including aged people, the keyboard and the mouse are not easy to use. Even for the others, they are too complicated to use in daily lives. The gesture recognition system will help us to control these useful appliances more easily. For example, if you want to turn off the light, you have only to point at the ceiling, or if you want to open the curtain in the morning, you have only to wave hands.

* Corresponding author.
E-mail address: yosike@sfc.keio.ac.jp (N. Yoshiike).

0925-2312/03/$ - see front matter
PII: S0925-2312(02)00617-3

To recognize gesture patterns that contain time-series data, hidden Markov model (HMM), artificial neural network model and dynamic programming matching (DP matching) have been used in conventional researches. Takeshi et al. [9] proposed the HMM-based gesture recognition system. While HMM is more appropriate for the independent subjects than other models, it needs a complex learning algorithm for setting parameters. Some artificial neural network models were proposed for association and recognition of time-series patterns by Yamasaki et al. [10], Shigematsu and Matsumoto [6], Tabuse et al. [7] and Nishiyama and Yagi [3]. These neural models for time-series patterns are challenging in terms of the connection to the biological systems, but these researches are at fundamental stages. Sagawa et al. [5], Nishimura et al. [2] and Osaki et al. [4] proposed the gesture recognition system using DP matching. To recognize a small scale of gesture patterns depending on the subject, it is known that DP matching works very well. DP matching is a method for adjusting distorted input data to template data by making a map between them and recognizing the input by evaluating the matching cost after mapping. While DP matching is well suited to recognize an object in a noiseless scene, there are many cases that other moving objects appear in a same scene of video captured images. It needs to remove noise motions for using DP matching in the real scenes.

A preprocessing method is proposed where it detects and removes noise motions for the DP matching method. In our applications, moving objects, for instance the heads or the other hands, are regarded as noise motions. In the first phase, we use maximum neural networks for segmenting objects from input motion data, maximum neural networks were originally proposed by Takefuji et al. [8] in order to force the state of the system to converge to the solution in neural dynamics. Amartur et al. [1] showed the applications for the segmentation of magnetic resonance images of the maximum neural networks. In our system, maximum neural networks are used for segmentation problems in time-series images by adding the following new conditions. We assume that the moving target objects have two conditions: (1) object motion data continuity inter-frame, and (2) object motion data continuity intra-frame. In our model, clusters tend to form as filling these two conditions. In the second phase, we use DP matching to recognize the target motion. The experimental results show the performance of this model by comparing the recognition results using DP matching after the convergence of maximum neural networks to that of DP matching without any preprocessing.

2. Object segmentation

The purpose of the first phase is to detect the target object from input motion data including non-target objects. By grouping the motion data as time-space segments, the target object can be detected and the non-target objects can be removed. We assume that the moving target objects have the following two conditions: (1) object motion data continuity inter-frame, and (2) object motion data continuity intra-frame. The former condition is based on the assumption that the target object should appear in almost the same position and size of the object in post and previous frames. The latter condition

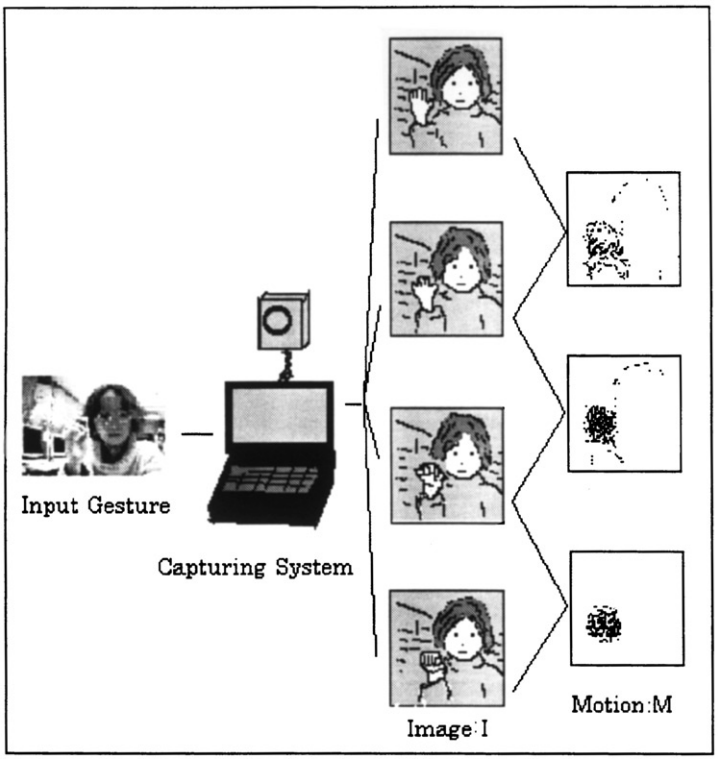

Fig. 1. Gesture inputs.

is based on the assumption that the target object should move as a group in the same time. These two conditions are considered when the clusters are formed in the iterative calculation.

2.1. Input

To translate video captured images, I, to sequential motion data, M, differences of each pixel of the succession frames are calculated (see Fig. 1). If the difference is larger than a threshold value at time t in the position x, motion data M_{ti} is set to x, where i is the current number of the motion data. The motion data is extracted from the frame in which the first motion appeared to the frame in which the motion vanished for the next few frames.

2.2. Clustering algorithm

Initially, the number of the clusters is set as 1, and the representative point of each frame, R_{1t}, is set as the gravity point of the M_t (see Fig. 2):

$$R_{1t} = \frac{1}{n} \sum_{i=0}^{m} M_{ti}. \tag{1}$$

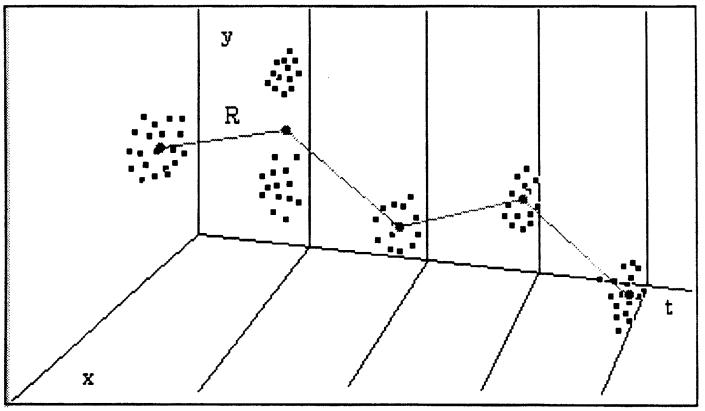

Fig. 2. Initial state of the representative point.

The cost for point i to categorize class j, C_{ij}, is calculated as

$$C_{ij} = \begin{cases} \alpha(|M_{ti} - R_{j,t-1}| + |M_{ti} - R_{j,t+1}|) + \beta|M_{ti} - R_{jt}| & \text{if } V_{ij} = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where V_{ij} represents if the point i belongs to the class j, and α and β are constant values. The left term of Eq. (2) coincides with the summation of the distances to the representative point of the previous frame and that of post frame. It agrees with condition (1) that the object should appear in almost the same position of that in previous and post frames. The right term of Eq. (2) coincides with the distance to the representative point of the same frame. It agrees with condition (2) that the object should move as a group in the same time.

The maximum cost is selected from class 1 in all the frames to compare to the average cost. If the ratio of the maximum cost and the average cost is less than a threshold value, the calculation is at an end. Else the new representative point, $R_{j't'}$ where j' is the current id of the class in the frame and t' is the current id of the frame that has the maximum cost, is set as the position of the motion data that has the maximum cost (see Fig. 3):

$$R_{j't'} = M_{t'i'}, \tag{3}$$

where i' is the id of the motion data that has the maximum cost.

To make the new cluster, $R_{jt'}$ is updated continuously by Eq. (4) and the new clusters are formed by Eq. (5) as maximum neuron rules for the constant number of

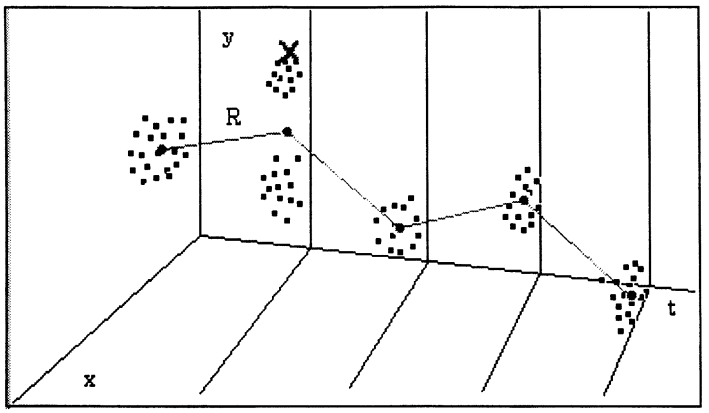

Fig. 3. A new representative point.

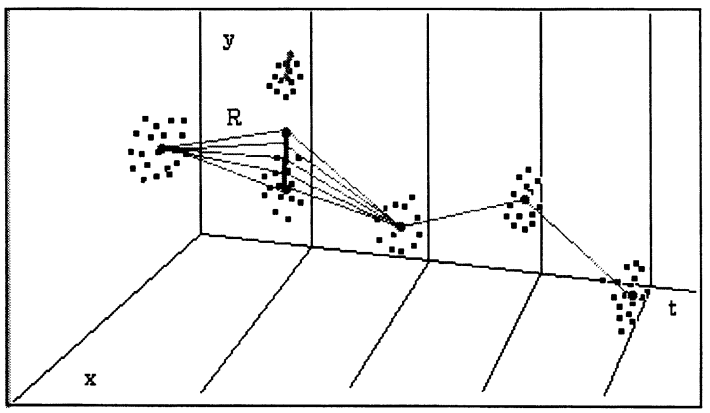

Fig. 4. Modification process of the representative points.

iterative steps (see Fig. 4):

$$\frac{\mathrm{d}R_{jt'}}{\mathrm{d}t} = \gamma\left(\frac{1}{n}\sum_{i=1}^{m} M_{t'i}\cdot V_{ij} - R_{jt'}\right), \tag{4}$$

$$V_{ij*} = \begin{cases} 1 & \text{if } |M_{t'i} - R_{j*t'}| = \min(|M_{t'i} - R_{jt'}\|\forall j), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

After the iterative steps, go to Eq. (2).

2.3. Output

Finally, the set of representative points of class 1, R_1, is regarded as the target object movement and that of other classes is regarded as the noise movements (see Fig. 5).

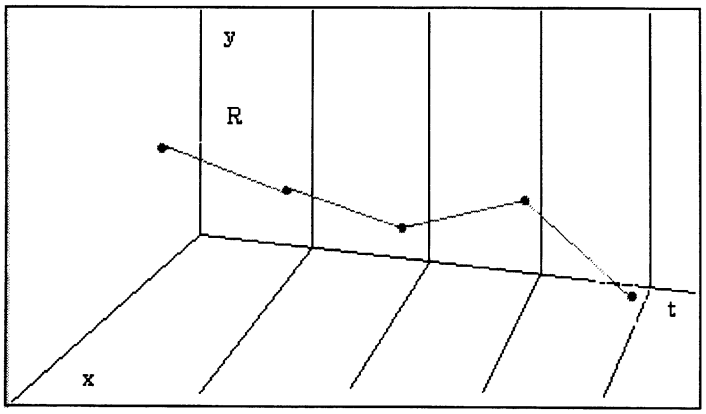

Fig. 5. Output as the movement of the target object.

3. Recognition

In the second phase, DP matching is used to recognize the target object. Input pattern is translated from the output sequence of phase 1 to the sequence of the differential data. By using difference of each two representative points of the succession frames, the input sequence becomes invariant for the object translation. The input sequence is represented as J and the distance from input pattern to the template pattern is represented as d. The accumulation of the distance from input pattern at time t to template pattern at time u, represented by $S(t,u)$, is calculated by the DP matching rule:

$$S(t,u) = \min \begin{pmatrix} S(t-2,u-1) + 2d(t-1,u) + d(t,u) & \text{(a)} \\ S(t-1,u-1) + 2d(t,u) & \text{(b)} \\ S(t-1,u-2) + 2d(t,u-1) + d(t,u) & \text{(c)} \end{pmatrix}. \tag{6}$$

By calculating Eq. (6), the minimum distance for making the map between input pattern and template pattern is obtained when the slopes are limited three ways. The length of the path, L, is calculated by

$$L(t,u) = \begin{cases} L(t-2,u-1) + 3 & \text{(a)}, \\ L(t-1,u-1) + 2 & \text{(b)}, \\ L(t-1,u-2) + 3 & \text{(c)}. \end{cases} \tag{7}$$

The total cost, C, is the normalized total distance by the path length:

$$C(t) = \frac{S(t,u)}{L(t,u)}. \tag{8}$$

The input can be recognized as the template that has the minimum cost C among the whole templates.

4. Simulation

In our simulation, the resolution of a captured image is 320×240 and the average frame rate of the motion data is 3.39 frame/s, which is measured after differential calculations. The following two subsections show a result for the maximum neural networks and results for the recognition using DP matching after convergence of the maximum neural networks respectively.

4.1. A clutering result

The input gesture pattern for the maximum neural networks is shown in Fig. 6 as motion data. The left hand is waving horizontally and the right hand and the head are sometimes moving. This gesture continued for 9.16 s and was captured into 31 frames. In Fig. 6, the head and shoulder are moving at the first three frames, and the right hand is moving in middle frames. These are considered as non-target objects. The differential sequence of the gravity points for each frame before segmenting objects is shown in Fig. 7. Fig. 8 shows an object segmentation process of the input pattern by applying the maximum neural networks. The selected frames are 27, 11, 12, 20, 8, 4, 2, 28, 29, 18 and 19 in order. In Fig. 8, the gray point 1 presents the trace of the target representative point and the arrow presents the direction of the point for each frame. The gray point 2 presents the non-target representative point and the noise area is encircled. Fig. 9 shows the differential sequence of the gravity points after the object segmentation process. This gravity sequence and the gravity sequence without the object segmentation process (Fig. 7) are matched with a template pattern (see Fig. 10) using DP matching. The template gesture pattern of the same hand motion is captured in a perfect condition without non-target objects. The result is shown in Fig. 11. The gray lines present mapping points decided by using DP matching. The matching costs of raw input pattern and template pattern is 38.58, and the cost of noise removed input pattern and template pattern is 36.85.

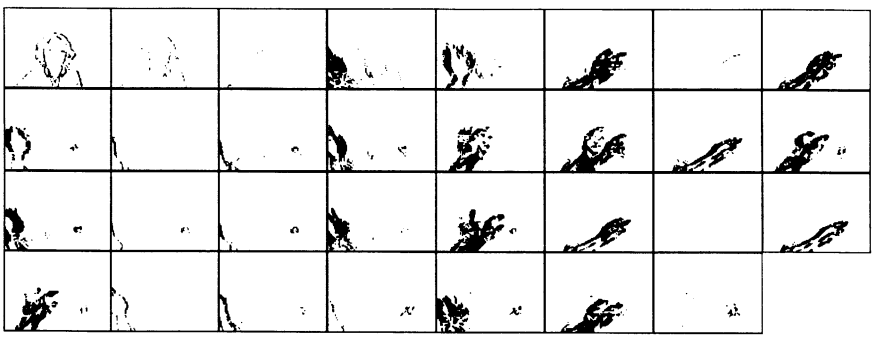

Fig. 6. A gesture input as motion data (frames 1–31).

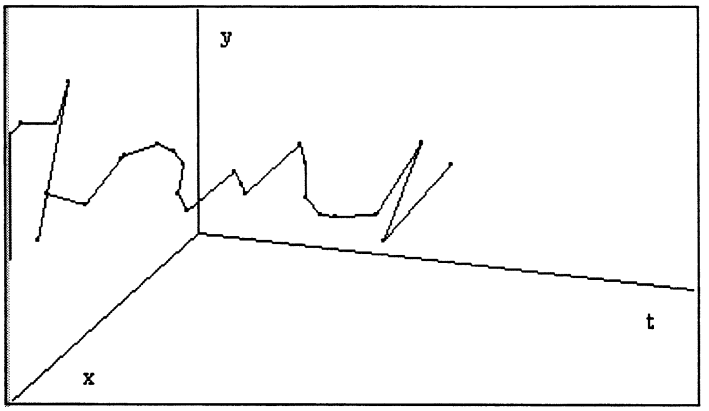

Fig. 7. The differential sequence of the gravity points without segmentation process.

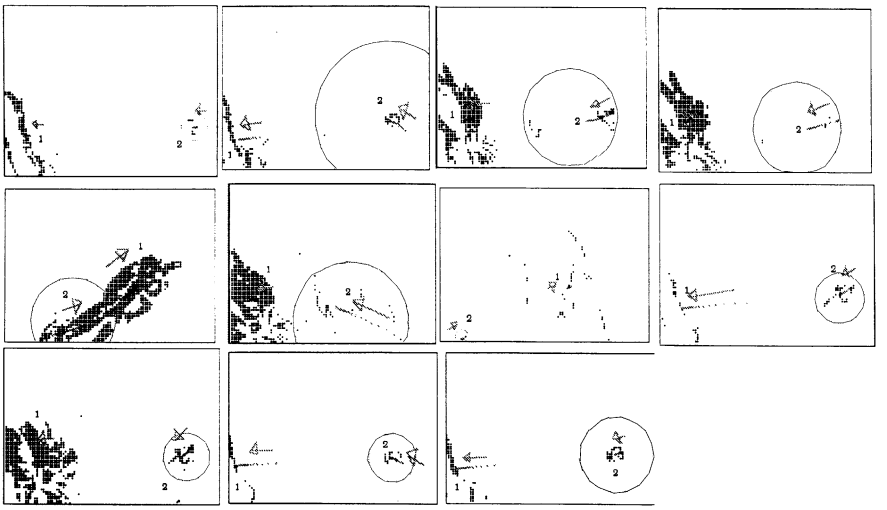

Fig. 8. A noise removal process of the input pattern by applying the maximum neural networks. The selected frames are 27, 11, 12, 20, 8, 4, 2, 28, 29, 18 and 19 in order.

4.2. Recognition results

Five kinds of gesture patterns are presented for testing the performance of our algorithm: a waving hand, a rounding hand, a hand tracing the figure of eight, a hand waving horizontally and a hand waving vertically. Video images are captured in noisy condition which include non-target objects and noise less condition which do not include non-target objects. To make the template data, each pattern is presented only once in each condition. Fifty different input patterns are presented for

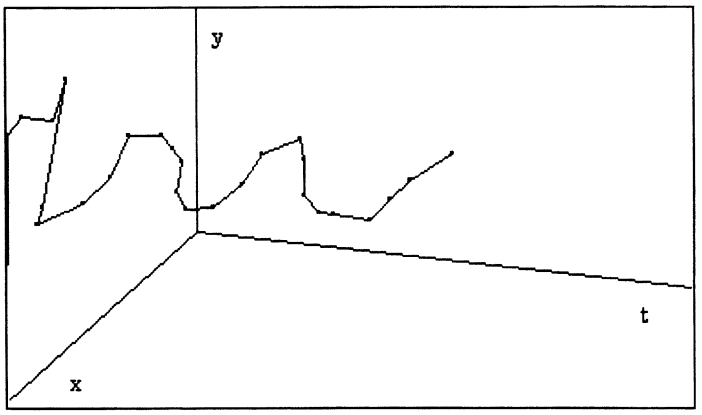

Fig. 9. The differential sequence of the gravity points after segmentation process.

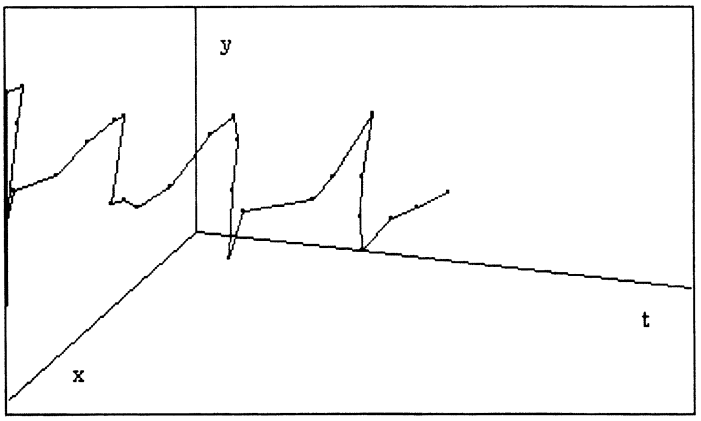

Fig. 10. A template pattern.

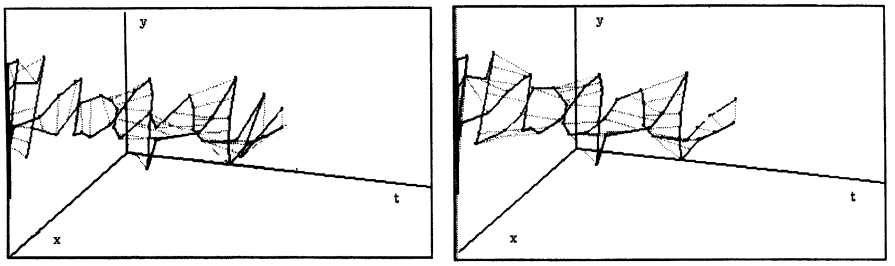

Fig. 11. DP matching result: the matching result of raw input pattern and the template pattern, and the matching result of cleaned input pattern and the template pattern, respectively.

Table 1
Recognition results

Input/template	DP matching	Clustering and DP matching	False → correct	Correct → false
With non-targets/ with non-targets	32/50	32/50	3	3
With non-targets/ only the target	31/50	34/50	3	0
Only the target/ with non-targets	48/50	49/50	2	1
Only the target/ only the target	50/50	50/50	0	0

each condition and tested using both template patterns. Table 1 shows the number of correct answers without any preprocessing and the number of correct answers after convergence of the clustering algorithm, and the number of changes from false answer to correct answer and the number of changes from correct answer to false answer.

5. Discussion

In the clustering result shown in the previous section, there are some remarkable points. First of all, the motion of the right hand, which does not present a gesture, is detected as a non-target object by using the maximum neural networks (see Fig. 8). In frame 27, 11, 12, 20, 4, 28, 29, 18 and 19 have a common feature that the target gravity point moves toward the real center point of the left hand because of the object segmentation. In these cases, non-target objects appeared as groups and that agrees with the assumption that the objects should move as a group in the same time. Also, the result that the attributions of the clusters are not mistaken means the assumption that the target object should appear in almost the same position and size of the object in post and previous frames is plausible. In frame 8, the arm of the hand is removed as a noise. In this case, it is not clear if the arm is a target object or not, but the target representative point moves closer to the center of the hand by removing the arm. Second, comparing the sequence of gravity points before object segmentation and that after the calculation of object segmentation, the latter presents the feature of the wave clearly (compare Fig. 9 to Fig. 7). The latter pattern has more salient peaks than the former, and greater self-correlation. Third, the matching result is better after the object segmentation. In Fig. 11, the latter waves are mapped successfully as peaks to peaks but the former are not. Consequently, the latter mapping cost is smaller than the former. However, in most frames the clustering is succeeded and the mapping result is finer than the former, but in a few frames the clustering is not succeeded. In frames 1 and 2, head and shoulder are considered as the target objects. These frames in which the target object disappears are difficult to segment by using our model. The recognition results in the previous section show the performance of the maximum neural networks.

In three cases (with non-targets/with non-targets, only the target/with non-targets and only the target/only the target), the results are not changed meaningfully. But in the remaining case (with non-targets/only the target), the recognition result increases 6% of all input patterns. In a real situation, input patterns rarely can be captured in the same good conditions in which template patterns can be captured previously. Our segmentation model is an effective algorithm for revision of input data in those cases.

We presented the simulation results in which the target object mainly appeared in the whole motions. If a non-target object has larger movement than the target object, noise detection would fail. This is because this system extracts the largest movement in the image sequence by removing other small motions.

6. Conclusion

In this paper, we presented the maximum neural networks for the object segmentation of a moving target and non-target objects. We assumed the following conditions for the moving target segment: (1) object motion data continuity inter-frame, and (2) object motion data continuity intra-frame. The experimental results showed the effectiveness of the proposed algorithm and the performance of this model in terms of the quality of the recognition results. Our algorithm is able to clean the input noise by segmenting objects before the recognition process.

Acknowledgements

The authors thank the reviewers for their valuable comments and suggestions, which have helped to improve the quality of this paper.

References

[1] S.C. Amartur, D. Piraino, Y. Takefuji, Optimization neural networks for the segmentation of magnetic resonance images, IEEE Trans. Med. Imag. 11 (2) (1992).
[2] T. Nishimura, T. Mukai, R. Oka, Spotting recognition system of gestures using shape features from gray-scale image sequence, Technical Report of IEICE, 1998-02, PRMU97-237, 1998.
[3] K. Nishiyama, S. Yagi, A consideration on memorialization and recall ability of TDNN for time series patterns, Technical Report of IEICE, 1997-03, NC96-167, 1997.
[4] R. Osaki, M. Shimada, K. Uehara, Extraction of primitive motions by using clustering and segmentation of motion-captured data, JSAI 15-5 (2000) 878–886.
[5] H. Sagawa, H. Sakou, E. Oohira, T. Sakiyama, M. Abe, Sign-language recognition method using compressed continuous DP matching, IEICE, 1994-04, D-II Vol. J77-D-II No. 4, pp. 753–763, 1994.
[6] Y. Shigematsu, G. Matsumoto, Selforganizing process for temporal associative memory—a physiologically plausible associative learning rule, Technical Report of IEICE, 1995-03, NC94-131, 1995, pp. 131–138.
[7] M. Tabuse, M. Kinouchi, M. Hagiwara, Recurrent neural network using mixture of experts for time series processing, Technical Report of IEICE, 1997-03, NC-96-168, 1997, pp. 99–106.
[8] Y. Takefuji, K.C. Lee, H. Aiso, An artificial maximum neural network: a winner-take-all neuron model forcing the state of the system in a solution domain, Biol. Cybernet. 67 (1992) 243–251.

[9] N. Takeshi, S. Haruyama, T. Kobayashi, HMM-based human gesture recognition, Technical Report of IEICE, 1996-05, PRMU96-8, 1996, pp. 53–59.
[10] T.Yamasaki, Y. Kataoka, K. Kameyama, K. Nakano, Neural networks memorizing sequential patterns, Technical Report of IEICE 1998-03, NC97-115, 1998, pp. 109–116.

Noriko Yoshiike received her M.Sc. degree in media and government in 2000 from the Department of Media and Governance, Keio University. Since 2000 she has been working towards her Ph.D. in the same university. Her current research interests focus on neural network models and their applications.

Yoshiyasu Takefuji is a tenured professor on faculty of environmental information at Keio University since April 1992 and was on tenured faculty of Electrical Engineering at Case Western Reserve University since 1988. Before joining Case, he taught at the University of South Florida for 2 years and the University of South Carolina for 3 years. He received his BS (1978), MS (1980), and Ph.D. (1983) in Electrical Engineering from Keio University under the supervision of Professor Hideo Aiso. His research interests focus on neural computing and hyperspectral computing. He received the National Science Foundation/Research Initiation Award in 1989 and received the distinct service award from IEEE Trans. on Neural Networks in 1992.