社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－MPS－52　(2)
2004／12／20

Second-order neural network と 自己組織化マップを使ったジェスチャー認識のための動作特徴抽出

青葉雅人*,　武藤佳恭**

*慶應義塾大学大学院 政策・メディア研究科, **慶應義塾大学 環境情報学部

概要：本論文では，ビデオベースのジェスチャー認識に対する前処理のニューラル手法を提案する．Second-order neural network (SONN) と自己組織化マップ (SOM) を動作手領域抽出と動作特徴の正規化に用いる．SONN はフレーム差分と比較してノイズに対して頑健性があり，SOM の位相保持特性は DP マッチングのデータ正規化に対して極めて有効である．実験結果は，これらのニューラル手法がジェスチャーパターン認識に有効であることを示している．
キーワード：ジェスチャー認識，SONN，SOM，位相保持マップ

Motion feature extraction using second-order neural network and self-organizing map for gesture recognition

Masato Aoba* and Yoshiyasu Takefuji**

* Graduate School of Media and Governance, Keio University

** Faculty of Environmental Information, Keio University

Abstract: We propose a neural preprocess approach for video-based gesture recognition. Second-order neural network (SONN) and self-organizing map (SOM) are employed for extracting moving hand regions and for normalizing motion features respectively. The SONN is more robust to noise than frame difference technique, and the topological property of the SOM is quite suited to data normalization for the DP matching technique. Experimental results show that those neural networks effectively work on the gesture pattern recognition.
keywords: hand gesture recognition, SONN, SOM, topological map.

1. Introduction

Using hand gestures is a common way for communications between human and human, therefore the gesture recognition system has a potential to be a useful human-computer interaction (HCI) tool.

In case video based gesture recognition system, motion feature extraction has much effect on its recognition performance. Some neural models have been proposed for motion extraction as prototypes [1][2], however few real time approaches were not presented for applications [3].

In this paper, we propose a neural preprocess approach for video-based gesture recognition system using two neural network models; second-order neural network (SONN) for extracting moving hand regions, and self-organizing map (SOM) for normalizing motion features. Time sequential motion feature pattern is classified by DP matching. Chashikawa *et al.* reported that SONN has robustness to noise in extracting moving objects [4]. The SOM is introduced by Kohonen [5] and it translates feature vectors into another feature space with keeping its topology and data distribution. This is quite suited to the DP matching technique. We applied those ideas for recognizing twelve hand gestures.

2. System Overview

We design a system to recognize hand gestures. RGB video images are translated into *L*a*b** images. Moving hand regions are extracted by SONN. Then velocity vector is calculated and translated into motion feature by motion feature map trained by SOM. The system feeds the motion features in time order as motion feature array throughout a gesture. The motion feature array is classified by DP matching and the system outputs the recognition results.

3. Motion Feature Extraction

3.1. Moving Hand Extraction

The RGB colors in the video images are translated into *L*a*b** color space in order to extract a moving hand-region. We modified second-order neural network (SONN) for moving hand region extraction. The binary output $O_{ij}(t)$ is calculated as follows,

$$O_{ij}(t)=\begin{cases}1 & if\ U_{ij}(t)\ge\Theta_{ij}(t)\\ 0 & otherwise\end{cases}$$

$$\Theta_{ij}(t)=\theta_o\left(1+\xi\sum_{i,j}U_{ij}(t)/(l_h\times l_w)\right)$$

$$U_{ij}(t)=F_{ij}(t)\left(1+\beta L_{ij}(t)\right)$$

$$F_{ij}(t)=\exp(-\tau_F)F_{ij}(t-1)+\gamma_F\sum_{k,l}W^F_{ijkl}O_{kl}(t-1)+\sum_{k,l}W^R_{ijkl}R_{kl}(t)$$

$$L_{ij}(t)=\exp(-\tau_L)L_{ij}(t-1)+\gamma_L\sum_{k,l}W^L_{ijkl}\left(O_{kl}(t-1)-1\right)$$

$$R_{ij}(t)=\gamma_R\left(S_{ij}(t)+\exp(-\tau_R)R_{ij}(t-1)\right)$$

$$S_{ij}(t)=\frac{D^{L^*}_{ij}(t)+D^{a^*}_{ij}(t)+D^{b^*}_{ij}(t)}{3}$$

$$D^{L^*}_{ij}(t)=C_{L^*}\left|I^{L^*}_{ij}(t)-I^{L^*}_{ij}(t-1)\right|$$

$$D^{a^*}_{ij}(t)=\exp^2\left(-\frac{\left(I^{a^*}_{ij}(t)-m_{a^*}\right)^2}{\sigma^2_{a^*}}\right)\left|I^{a^*}_{ij}(t)-I^{a^*}_{ij}(t-1)\right|$$

$$D^{b^*}_{ij}(t)=\exp^2\left(-\frac{\left(I^{b^*}_{ij}(t)-m_{b^*}\right)^2}{\sigma^2_{b^*}}\right)\left|I^{b^*}_{ij}(t)-I^{b^*}_{ij}(t-1)\right|$$

where U_{ij} is the internal activity and S_{ij} is the input stimuli. $I^{L^*}_{ij}$, $I^{a^*}_{ij}$ and $I^{b^*}_{ij}$ are the input value at pixel (i, j) for L^*, a^* and b^* respectively. W^F_{ijkl}, W^L_{ijkl}, W^R_{ijkl} are Gaussian kernels. Θ_{ij} is the dynamic threshold. An example of hand gesture is shown in Figure 1.

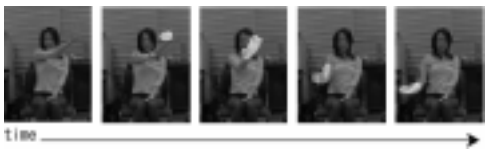

Figure 1 Moving hand extraction

3.2. Motion Feature Map

The velocity of the gravitation center **G** is defined as

$$\mathbf{v}(t)=\mathbf{G}(t)-\mathbf{G}(t-\Delta t)$$

Then we define velocity array vector $\mathbf{V}(t)$ as an array of $\mathbf{v}(t)$.

$$\mathbf{V}(t)=\left[\mathbf{v}(t),\mathbf{v}(t-1),\cdots,\mathbf{v}(t-n_V-1)\right]$$

The output signal y^f_{ij} of the ith jth output neuron is calculated as follows,

$$y^f_{ij}=\begin{cases}1 & if\ i=i_{win}\cap j=j_{win}\\ 0 & otherwise\end{cases}$$

$$\left\|\mathbf{m}_{i_{win}j_{win}}-\mathbf{V}\right\|=\min_{i,j}\left\|\mathbf{m}_{ij}-\mathbf{V}\right\|$$

where i_{win} and j_{win} are the indices of the winner neuron, $\mathbf{m}_{ij}$ is the codebook vector. We define motion feature as following.

$$\mathbf{x}(t)=\left[i_{win},j_{win}\right]$$

The codebook vectors $\mathbf{m}ij$ are adjusted by SOM learning rule.

$$\mathbf{m}_{ij}(s_f+1)=\mathbf{m}_{ij}(s_f)+\eta(s_f)\exp\left(-\frac{\left\|[i,j]-[i_w,j_w]\right\|^2}{\sigma^2_n(s_f)}\right)\left\{\mathbf{V}_p-\mathbf{m}_{ij}(s_f)\right\}$$

$$\left\|\mathbf{m}_{[i_w,j_w]}(s_f)-\mathbf{V}_p\right\|=\min_{i,j}\left\|\mathbf{m}_{[i,j]}(s_f)-\mathbf{V}_p\right\|$$

4. Recognition

In the recognition part, dynamic programming (DP) matching is implemented. The motion pattern **X** is defined as a sequence of the input motion feature $\mathbf{x}(t)$,

$$\mathbf{X}=\left\{\mathbf{x}(1),\mathbf{x}(2),\cdots,\mathbf{x}(t),\cdots\mathbf{x}(t_{\max})\right\}$$

The template $\mathbf{R}_q$ of the category q is also defined as a sequence of the motion feature $\mathbf{r}_q(u)$.

$$\mathbf{R}_q=\left\{\mathbf{r}_q(1),\mathbf{r}_q(2)\cdots,\mathbf{r}_q(u),\cdots\mathbf{r}_q(u_{\max})\right\}$$

An accumulated cost $C_q(\mathbf{X}, t, u)$ and a length of the path $L_q(\mathbf{X}, t, u)$ is calculated by the DP matching rule. Normalized accumulation cost $y^{DP}{}_q(\mathbf{X})$ is acquired by following.

$$y^{DP}_q(\mathbf{X})=\frac{C_q(\mathbf{X},t_{\max},u_{\max})}{L_q(\mathbf{X},t_{\max},u_{\max})}$$

Recognition result is obtained by finding the category with minimum $y^{DP}{}_q(\mathbf{X})$. The template is figured out as averaged vectors of time normalized input patterns.

5. Experiments

5.1. Training Conditions

The system is trained to recognize twelve hand-gesture patterns. Training data were obtained from three examinees at different backgrounds. We label them as scene A, B and C. Three examinees performed all gesture patterns 6 times.

The obtained motion feature map calculated by SOM is shown in Figure 2, and an example of feature trajectory for a test movie is shown in Figure 3.

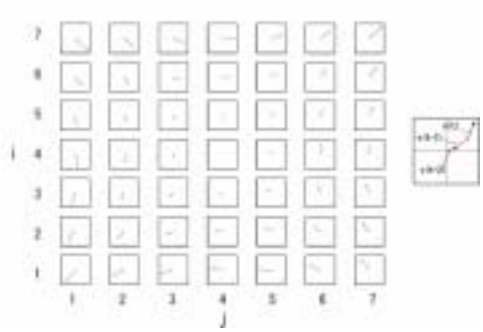

Figure 2 Motion Feature Map

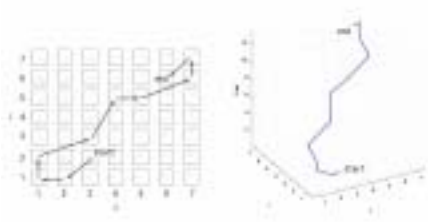

Figure 3 Example of feature trajectory

5.2. Experimental Results

5.2.1. Recognition Rates

At first, we have tested other 360 untrained data to recognize the gestures in "known" situations. They performed all gesture patterns 10 times. Then we have tested other 600 untrained data to recognize the gestures at "unknown" situations. We label them as scene 1 to 6. The six examinees performed all gestures 10 times. The results are shown in Figure 4.

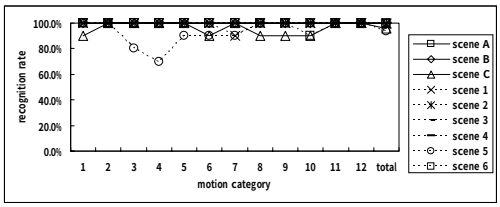

Figure 4 Recognition rates

5.2.2. Comparative Experiments

At first, we replace the SONN in our system with frame difference technique. This system also uses skin-color regions using $L^*a^*b^*$ color space. The recognition results for this modification are shown in Figure 5.

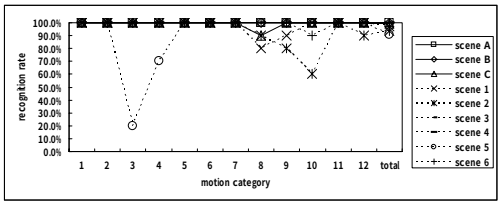

Figure 5 Recognition rates of the system using frame difference technique

In order to verify the noise reduction ability of the SONN, we prepared additional test data as scene N. The scene N contains an ornament waving by wind at the background. The recognition rates of the system using frame difference are compared with those of the system using SONN for the scene N in Figure 6.

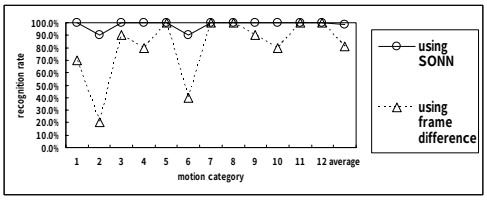

Figure 6 Comparison of the recognition rates for scene N

The second comparative system does not employ the motion feature map. The recognition results are shown in Figure 7.

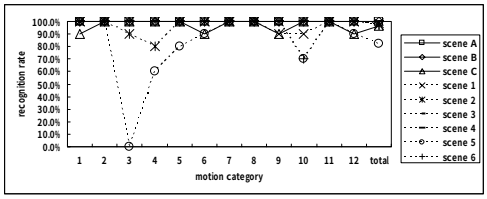

Figure 7 Recognition rates of the system without motion feature map

In addition, we translated the video images of the scene 3 into various sized images. Comparisons of the recognition rates for the image distortions are shown in Figure 8 and Figure 9.

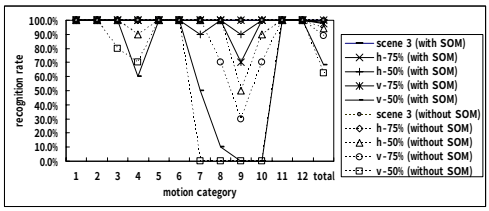

Figure 8 Comparison for diminution

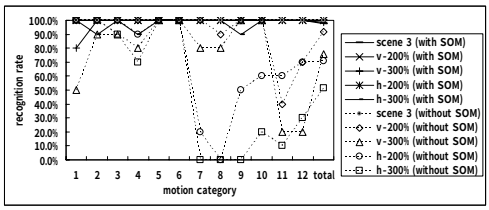

Figure 9 Comparison for expansion

6. Discussion

The recognition results of our system are shown in Figure 4. The results show that the system has a high performance for recognizing gestures by various persons at various backgrounds.

As illustrated in Figure 1, SONN well extracts moving hand regions. Figure 6 shows the recognition rates of the both systems for noisy background. The SONN acts on scenes at noisy backgrounds more appropriately than the frame difference technique.

The results of the comparative experiments in Figure 8 and Figure 9 show the robustness. The results in Figure 8 show that the motion feature map alleviates the effects of scaling down distortions. Figure 9 significantly indicates the robustness of the motion feature map to scaling up distortions. This is caused by the fact that the SOM optimizes upper and lower thresholds for input vectors are defined automatically.

Topological distances between the competitive neurons in the map approximate statistical distances in the feature space since the SOM quantizes and approximates data distribution with keeping their topology. This trait is suited to data normalization for the DP matching.

7. Conclusion

We propose a neural preprocess approach for video-based gesture recognition. Our experimental results show that the system has a good performance to classify twelve hand gesture patterns. For situations with noisy backgrounds, the SONN acts on more appropriately than frame difference technique. The SOM provides the robustness to spatial scaling distortion of input video images, and topological property of SOM is quite suitable to normalizing feature vectors for DP matching technique.

References

[1] Kubota, T. : Massively parallel networks for edge localization and contour integration – adaptable relaxation approach, Neural Networks, Vol.17, pp.411-425, (2004).

[2] Katayama, K., Ando, M. and Horiguchi, T. : Models of MT and MST areas using wake-sleep algorithm, Neural Networks, Vol.17, pp.339-351, (2004).

[3] Yoshiike, N. and Takefuji, Y. : Object segmentation using maximum neural networks for the gesture recognition system, Neurocomputing 51 (2003) 213-224

[4] Chashikawa, T. and Takefuji, Y. : Extracting Moving Object Areas Based on Second-order Neural Network, IPSJ Vol.44, No.SIG 14(TOM 9), pp. 31-47, 2003.

[5] Kohonen, T. : Self-Organizing Maps, Springer-Verlag, Berlin (1995).