A Real-time Hand Gesture Recognition Technique and Its Application to Music Display System

Jun-Yong Lee, Joong-Eun Jung, and Ho-Joon Kim
Dept. of Computer Science and Electrical Engineering, Handing University, Pohang, South Korea
Email: {leejunyong, jejung}@hgu.edu, hjkim@handong.edu

***Abstract*—In the paper, we introduce a real-time hand gesture recognition method using a neural network. The underlying system is an automatic music display system which consists of three modules; feature extraction module, pattern classification module, and display control module. To reduce the computation time of the feature extraction process and the pattern classification process, a three-dimensional data representation called motion history volume has been adopted. In addition, we propose a feature selection technique based on a modified fuzzy min-max neural network. We have defined a relevance factor which can measure the relevance of a feature to classify the specific pattern classes. The feature selection method can remove ineffective features and erroneous features in the learning data set by using the relevance factor data.**

***Index Terms*—hand gesture recognition, motion history volume, feature selection**

I. Introduction

It is uncomfortable for musicians to turn pages by themselves while playing music and it can even disturb a musical performance. Thus, in this paper, we consider a method to control pages easily by recognizing simple hand gestures. For the purpose, we introduce an effective data representation technique in order to present the motion information in video.

For the data representation, we have adopted a three-dimensional structure named Motion History Volume (MHV) which is generated by stacking the motion information along the time dimension [1], [2].

Fuzzy Min-Max (FMM) neural network is a hyperbox-based pattern classification model [3], [4]. In our previous research, we proposed a learning method which reflects the frequency factor of a feature value in the learning data [5]. It prevents the performance degradation caused by some abnormal data and overcomes the ambiguity of classification without the contraction process for overlapping hyperboxes during the learning process. Furthermore, it enables us to generate rule-based knowledge by quantification of relevance factors between features and pattern classes from the trained neural network. However, in some problems such as image recognition, the learning data may include erroneous data or ineffective data because there may exist unexpected variations in the image data. They degrade the recognition rate and effectiveness of the learning process.

In this research, we propose a methodology to select an effective feature set by extending the FMM neural network model. We have defined a relevance factor which measures the relevance of a feature with its pattern classes. The feature selection method removes ineffective and abnormal features from the learning data set by using the relevance factor data. The applicability of the music display system and its performance of recognition have been tested with the residual features.

II. Music Display System

As shown in Fig. 1, the music display system consists of two processes: Learning procedure and pattern classification process. The motion history volumes generated from the input video data are converted into three-dimensional feature maps. The feature selection process computes the relevance factors for each unit of the feature map for the pattern classes. Erroneous features are removed through the process and selective features are used for building the pattern classifier.

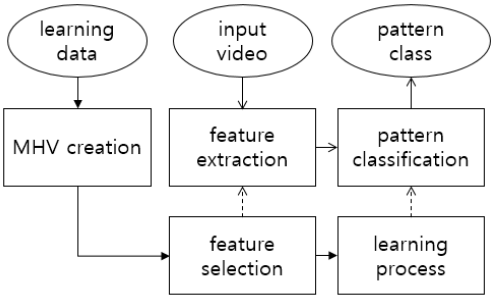

Figure 1. The structure of the music display system

Fig. 2 shows the image of page turning action in the developed music display system. We have defined six types of hand gesture patterns: *next, previous, speed up, speed down, pause and restart.*

Manuscript received January 21, 2015; revised May 11, 2015.

doi: 10.12720/joace.4.2.177-180

Figure 2. A demonstration of the system.

III. Feature Extraction and Recognition Module

Convolutional neural networks (CNN) incorporate constraints and achieve some degree of shift and deformation invariance using spatial subsampling and local receptive fields [6]. When an image pattern is input, spatially-localized subset of units (receptive fields) are passed through the two-dimensional processing element in the subsequent layers. The convolution layers have orientation-selective filter banks where elementary visual features are extracted from the spatial template. The filtered image is then subsampled by the subsampling layer. Spatial resolution is reduced in this process and certain amount of translation is ignored. Therefore, each sub-layer generates a feature map which reflects successively larger ranges of the preceding unit. In our previous study [7], we introduced an extended version of the CNN for temporal feature extraction. The input data for the feature extractor are represented as a spatiotemporal volume which is described in the previous section. The spatial structure of the receptive field in the model is extended along the time axis. The center of the three-dimensional processing element shifts through the spatial and temporal domain of the cube by two positions. Thus, the proposed model is not only robust to spatial variance but also to temporal variance.

Motion history information is used for the feature extraction module in our model. We have CNN model to extract feature maps form the motion history volume (MHV) [8]. Fig. 3 shows an example of the data representations of the real-time gesture signals and the feature map generated from the data. In the figure, the direction of time sequence is from the left column to the right column. For each frame in the image sequence, the object region is cut out by a background subtraction and contour detection method. We refer to motion as the occurrence of object region pixels between contiguous images, i.e. if the object region did not exist at an image point (x, y) at time t and appeared at the same location at time $t + 1$, it indicates that the point is a region of motion. By stacking the motion information along the time dimension, we obtain a spatiotemporal volume data. Since motion is to occur near the boundary of the object region, the template provides a certain degree of shape information as well as the direction of the object movement.

As a preprocessing module, the video data of the dynamic hand gestures are converted into MHVs. As illustrated in Fig. 3, the MHV is generated by extracting the motion information and stacking it for each frame along the time dimension. In order to reduce the number of features, we generate three-dimensional feature map of size (5×5×5). We use a Multi-Layer Perceptron (MLP), a conventional neural network model, for the pattern classification module. Fig. 4 shows the examples of the MHVs generated from the 6 patterns in each video.

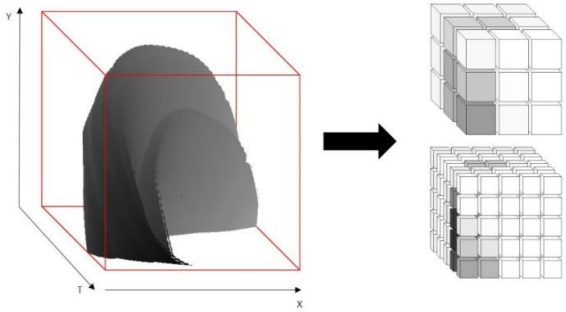

Figure 3. An example of the MHV and its feature maps.

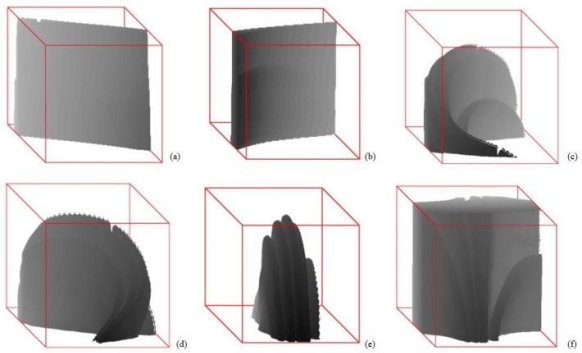

Figure 4. Examples of MHVs of 6 patterns: (a) *next*, (b) *previous*, (c) *speed up*, (d) *speed down*, (e) *pause*, (f) *restart*

IV. Feature Selection with FMM Model

In our previous work, we had proposed a method of a rule extraction for pattern classification using a fuzzy min-max (FMM) neural network [5].We defined new adjustment schemes from the new definition of hyperbox for the four cases. The frequency value and the mean point value as well as the min and max points are updated for the four cases. Consequently the frequency factor is increased in proportion to the relative size of the feature range, and the mean point value is adjusted by considering the expanded feature range.

From the trained FMM neural network, the relevance factor between features and pattern classes can be calculated.

We classified the feature ranges of each dimension as 5 fuzzy partitions. Each partition is represented as a trapezoidal fuzzy membership function. If the relevance factor has positive value, it means that the feature is an excitatory signal for the pattern class. On the other hand,

a negative value of relevance factor means the inhibitory relationship between the feature and the pattern class. From the analysis of the relevance factors we can generate a set of rules for the pattern classification.

In this study, we suggest an extension of the previous model for the feature selection.

We define a relevance factorw_{ik}as follows:

$$w_{ik} = \frac{m_k}{|I_{ik}|} \cdot \sum_{j,j\neq k}^{N} (1.0 - \frac{|I_{ik} \cap I_{ij}|}{MIN(|I_{ik}|,|I_{ij}|)}) \qquad (1)$$

In the equation, w_{ik} is the relevance factor of the i^{th} feature related with the k^{th} pattern class and m_k is the number of patterns used for pattern class k for the learning process. I_{ik} means the feature range of the hyperbox in i^{th} dimension. $|I_{ik} \cap I_{ij}|$is the length of the overlapping area of the two ranges for k and j. The equation represents that it is proportional to the frequency of the pattern classes and inversely proportional to the length of feature range. Also, the relevance factor increases when the size of overlapping area between the different classes is small.

For the entire pattern class set, the final relevance factor is computed by averaging the relevance factors in the same dimension of all the pattern classes.

$$R(i) = \frac{\sum_k^N w_{ik}}{N} \qquad (2)$$

VI. Experimental Result

This section shows the performance of recognition using pattern selection method proposed in this paper. The 6 Patterns have been predefined with total 30 training data: 5 data for each class. Also, total 120 data for 20 data per class have been tested. The feature selection process is applied to both learning and test process. The results show that how the feature selection affects the performance of the recognition.

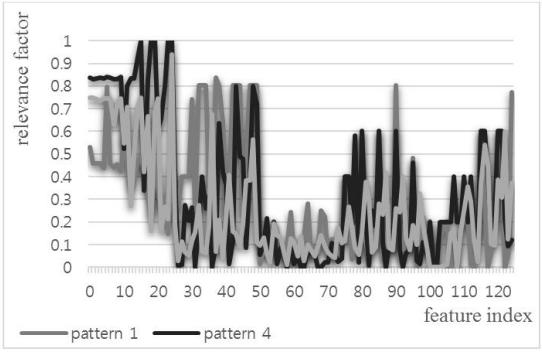

Figure 5. Examples of the relevance factors for two arbitrary patterns

Fig. 5 is the diagram that shows the relevance value of each feature. And three out of six pattern classes are displayed. The relevance factor indicates that how much impact the feature has on the decision for a specific pattern. As the graph shows, each relevance factor has unique value while bundle of them are roughly the same. To further explain this, Fig. 6 shows the average relevance factor of each feature for all 6 patterns that is obtained by (2). Features are selected and removed depend on the value of the average relevance factor.

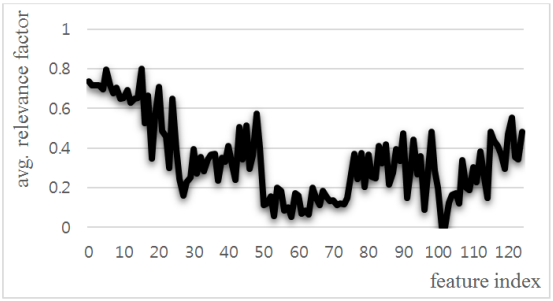

Figure 6. The average value of relevance factors for the 6 patterns.

Fig. 7 and Table I indicate the recognition rates according to the applied number of features. The number of features affected the computation time of both training and recognition processes. This output will provide a guideline of design specification of a system compromising both running time and recognition rate.

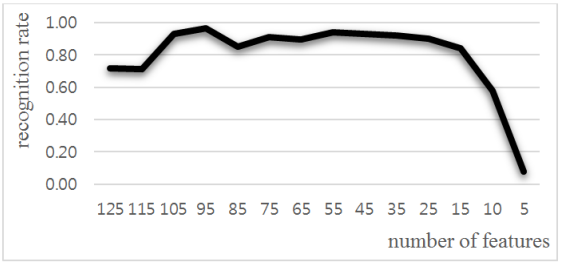

Figure 7. The recognition rate with respect to the number of selected features.

TABLE I. The Recognition Rate with Respect to the Number of Features

Number of Features	Recognition Rate	Number of Features	Recognition Rate
125	0.72	55	0.94
115	0.71	45	0.93
105	0.93	35	0.92
95	0.97	25	0.90
85	0.85	15	0.84
75	0.91	10	0.58
65	0.90	5	0.07

VI. Conclusion

One of the major problems in image recognition is the computation time of the recognition process. This is due to the large number of the features extracted from motion information which eventually complicates the recognition procedure, resulting in increasing of the computation time. Moreover, it is difficult to explicitly define the relationship between motion features and pattern classes with the conventional neural networks such as MLP because it includes the hidden layers. However, the FMM neural network model facilitates the analysis of an association between the features and pattern classes. The feature selection method proposed in this paper combines the advantages of the two neural network models: MLP

and FMM neural network.The method provides relevance factors which are the numeric representation of the relationship between the features and pattern classes. The classification module uses the relevance factors to remove the ineffective features. In other word, the classification module removes features that either do not affect (or have small effect on) the decision of classifying pattern or even tarnish the categorizing process. Hence, this method not only increases the recognition rate but also curtails the computation time of the system. For the future research, we will focus on the performance improvement by grouping of the pattern classes according to the relevance factors.

ACKNOWLEDGMENT

This research was financially supported by the Ministry of Education (MOE) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation (No. 2012H1B8A2025800).

REFERENCES

[1] J. Lin, and Y. C. Ding, "A temporal hand gesture recognition system based on hog and motion trajectory," *Optik - International Journal for Light and Electron Optics*, vol. 124, pp. 6795-6798, Dec. 2013.

[2] X. J. Peng, Y. Oiao, and O. Peng, "Motion boundary based sampling and 3d co-occurrence descriptors for action recognition," *Image and Vision Computing*, vol. 32, pp. 616-628, July 2013.

[3] P. K. Simpson, "Fuzzy min-max neural network-part 1: Classification," *IEEE Transaction on Neural Network*, vol. 3, no. 5, pp. 776-786, Sep. 1992.

[4] A. Quteishat, C. P. Lim, and K. S. Tan, "A modified fuzzy min-max neural network with a genetic algorithm-based rule extractor," *IEEE Transaction on System, Man, and Cybernetics-Part A: System and Humans*, vol. 40, no. 3, pp. 641-650, May 2010.

[5] H. J. Kim and S. J. Park, "Rule extraction for dynamic hand gesture recognition using a modified fmm neural network," *International Journal of Software Engineering and Its Applications*, vol. 7, no. 6, pp. 367-374, Dec. 2013.

[6] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1408-1423, Nov. 2004.

[7] H. J. Kim, S. J. Park, and S. K. Lee, "Sign language recognition using motion history volume and hybrid neural networks," *International Journal of Machine Learning and Computing*, vol. 2, no. 6, pp. 750-753, Dec. 2012.

[8] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249-257, Oct. 2006.

Jun-Yong Leeis in a bachelor's degree of in computer science and electrical engineering from Handong Global University, Pohang, Kyeongbuk, South Korea. His research interests include pattern recognition, neural network, artificial intelligence, image processing, medical vision and contrast enhanced ultrasonography image analysis.

Joong-Eun Jung received a bachelor's degree of in computer engineering from Handong Global University, Pohang, Kyeongbuk, Korea in 2013. He is currently studying for a master's course in Handong Global University. He is knowledgeable in computer vision, artificial intelligence and mobile programming. He received presentation award in Engineering Education Festa 2012.

Ho-Joon Kim received the B.S. degree in Computer Engineering from Kyeongbuk National University, Korea, in 1987 and the Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1995. He worked as a researcher at the Korea Atomic Energy Research Institute from 1987 to 1991. Currently he is a professor at the School of Computer Science and Electric Engineering, Handong Global University, Korea. His research interests include machine vision, pattern recognition, neural network architectures, and medical image processing. Prof. Kim received the Best Paper Award from the technical program committee of IEEE International Conference on Neural Networks and Signal Processing in 2008.