Sign Language Recognition Using Motion History Volume and Hybrid Neural Networks

Ho-Joon Kim, So-Jeong Park, and Seung-Kang Lee

***Abstract*—In this paper, we present a sign language recognition model which does not use any wearable devices for object tracking. The system design issues and implementation issues such as data representation, feature extraction and pattern classification methods are discussed. The proposed data representation method for sign language patterns is robust for spatio-temporal variances of feature points. We present a feature extraction technique which can improve the computation speed by reducing the amount of feature data. A neural network model which is capable of incremental learning is introduced. We have defined a measure which reflects the relevance between the feature types and the pattern classes. The measure makes it possible to select more effective features without any degradation of performance. Through the experiments using six types of sign language patterns, the proposed model is evaluated empirically.**

***Index Terms*—Sign language recognition, neural network, feature extraction, pattern classification.**

I. Introduction

Many researchers have been working on the recognition of various sign languages and gestures, but this research poses major difficulties due to the complexity on hand and body movements in sign language expression[1]-[3]. Recently several approaches to represent the motion information for the human action recognition in video have been reported. Weinland et al. proposed a human action recognition model using 3D volume structures called Motion History Volume (MHV) as a free-viewpoint representation for human actions in the case of multiple calibrated video cameras [4]. Yilmaz et al. proposed a novel action representation method named action sketch which is generated from a view-invariant action volume [5]. Convolutional neural networks (CNN) have been successfully applied to object recognition in 2D images [6]. The CNN model is a biologically inspired hierarchical multilayered structure, where each sub-layers incorporate feature extraction and feature reduction. Thereby, the feature extractor can achieve partial invariance of shift, rotation, and scale. Simpson introduced a fuzzy min-max (FMM) neural network based on fuzzy hyperbox sets representing the data clusters [7]. Gabrys et al. developed a general fuzzy min-max (GFMM) neural network by integrating the classification and clustering features of the original two FMM models and generalizing some features[8]. In our previous works, we proposed a weighted fuzzy min-max (WFMM) neural network[9] which has a modified activation function.

In this study, we present a hybrid neural network model for sign language recognition. The model consists of two types of neural networks, a modified CNN model and the WFMM model. For the feature extraction stage, we use the motion history volume which is generated by stacking the motion information along the time dimension. The CNN model generates a set of feature maps from the three-dimensional input data. The modified CNN model is not only robust to spatial variance but also to temporal variance. The WFMM model classifies activity patterns using the generated feature maps. In this paper, we first describe the system structure and object representation method. Then we present a modified CNN model which has three-dimensional receptive field structure for feature extraction. Finally we introduce the pattern classification and feature analysis technique using the WFMM model and discuss its validity from the experimental results.

II. Underlying System

As shown in Fig.1, our underlying system model consists of three modules: preprocessing module, feature extraction module and pattern classification module. In the preprocessing module, the feature regions are detected through the skin color analysis process. We have used the motion energy data and the motion history data for the feature extraction. An extended version of CNN model is used for the feature map generation. The model generates a three dimensional feature map from the motion history data. For the pattern classification, we have adopted the weighted FMM model [9] which can provide a feature analysis facility using a feature relevance measure.

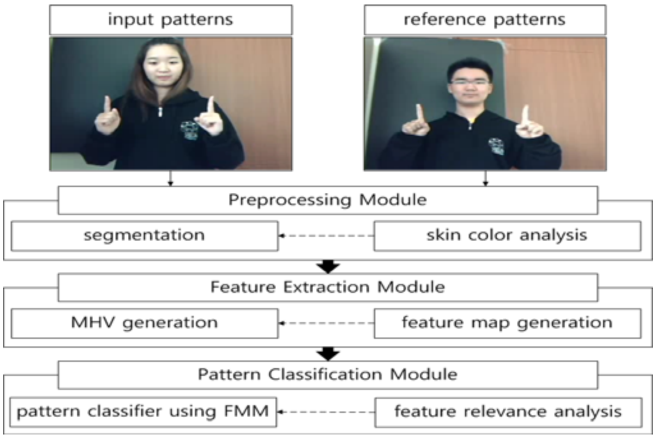

Fig. 1. The underlying face detection system

Manuscript received September 17, 2012; revised November 12, 2012.

Ho-Joon Kim is with the School of Computer Science and Electric Engineering, Handong Global University, Pohang, Kyungbuk, Korea (e-mail: hjkim@handong.edu).

So-Jeong Park and Seung-Kang Lee are with the Dept. of Information and Communication at Handong Global University, Pohang, Kyungbuk, Korea (email: sojeongmon@gmail.com; lsk0704@gmail.com)

10.7763/IJMLC.2012.V2.229

III. Feature Extraction

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 this paper, we introduce 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.

The size of the initial feature map used in this research is (23 × 23 ×23). After the input data is processed through the feature extractor, a final feature map of size (3 × 3 × 3) is generated. This feature map becomes the input of the pattern classifier.

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[4]. Fig. 2 shows an example of the data representations of sign patterns in video 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.

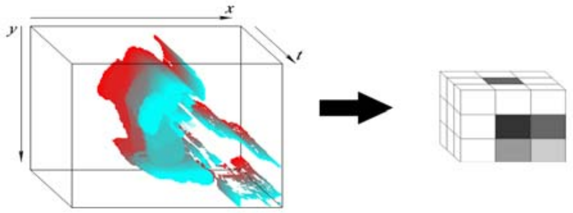

Fig. 2. An example of the spatio-temporal volume and the feature map

IV. Pattern Classification

The weighted FMM neural network model[9] has been used for the pattern classifier of the system. In the model a hyperbox is defined as:

$$B_j = \{X, U_j, V_j, C_j, F_j, f(X, U_j, V_j, C_j, F_j)\} \;\; \forall X \in I^n \qquad (1)$$

In the equation, U_j and V_j mean the vectors of the minimum and maximum values of hyperbox *j*, respectively. C_j is a set of the mean points for the feature values, and F_j means a set of frequency of feature occurrences within a hyperbox. The activation function of the network is defined as following fuzzy membership function.

$$b_j(A_k) = \frac{1}{\sum_{i=1}^{n} w_{ji}} \cdot \sum_{i=1}^{n} w_{ji} \Big[max\big(0, 1\text{-}max\big(0,\ \gamma min\big(1,\ a_{ji} - v_{ji}\big)\big)\big)$$
$$+ max\big(0, 1\text{-}max\big(0,\ \gamma \min\big(1,\ u_{ji} \text{ - } a_{hi}\big)\big)\big)\Big] \qquad (2)$$

$$\begin{cases} \gamma_{jiU} = \dfrac{\gamma}{R_U} & R_U = \max(s, u_{ji}{}^{new} - u_{ji}{}^{old}) \\ \gamma_{jiV} = \dfrac{\gamma}{R_V} & R_V = \max(s, v_{ji}{}^{old} - v_{ji}{}^{new}) \end{cases} \qquad (3)$$

As shown in the equation, the membership function in the model has the weight factor to consider the relevance of each feature as different values. In the equation, w_{ij} is the connection weight between *i*-th feature and *j*-th hyperbox. The parameter γ_{ijU} and γ_{ijV} control the slope of the fuzzy membership function at the right and left boundaries of the feature range, respectively.

The learning process of the model consists of two processes: hyperbox creation and hyperbox expansion. When a hyperbox is created, the initial weight value of that is set by 1.0. During the hyperbox expansion, the weight values are adjusted. If the expansion criterion has been met for hyperbox B_j , f_{ji}, u_{ji}, v_{ji} and c_{ij} are adjusted using the following equations.

$$\begin{cases} f_{ji}^{new} = f_{ji}^{old} + 1 \\ u_{ji}^{new} = \min(u_{ji}^{old}, x_{ki}) \\ v_{ji}^{new} = \min(v_{ji}^{old}, x_{ki}) \end{cases} \qquad (4)$$

$$c_{ji}^{new} = (c_{ji} * f_{ji}^{old} + x_{hi}) / f_{ji}^{new} \qquad (5)$$

The frequency factor increases in proportion to the relative size of the feature range, and the mean point value is adjusted by considering the expanded feature range.

V. Feature Analysis

This section describes a feature analysis technique for the sign pattern recognition. We have defined two kinds of relevance factors using the proposed FMM model as follows:

RF1(x_j, C_k): the relevance factor between a feature value x_j and a class C_k

RF2(X_i, C_k): the relevance factor between a feature type X_i and a class C_k

The first measure *RF1* is defined as Equation (9). In the equation, constant N_B and N_k are the total number of hyperboxes and the number of hyperboxes that belong to class k, respectively. Therefore if the *RF1(x_j*, *C_k)* has a positive value, it means an excitatory relationship between the feature x_i and the class k. But a negative value of *RF1*(x_j, C_k) means an inhibitory relationship between them. A list of interesting features for a given class can be extracted using the *RF1* for each feature.

$$RF1(x_i,C_k)=(\frac{1}{N_k}\sum_{B_j\in C_k} S(x_i,(u_{ji},v_{ji}))\cdot w_{ij}$$

$$-\frac{1}{(N_B-N_k)}\sum_{B_j\notin C_k} S(x_i,(u_{ji},v_{ji}))\cdot w_{ij})\,/\sum_{B_j\in C_k} w_{ij} \qquad (6)$$

In Equation (6), the feature value x_i can be defined as a fuzzy interval which consists of min and max values on the *i*-th dimension out of the n-dimension feature space. For an arbitrary feature x_i, let x_i^L and x_i^U be the min and max value, respectively, then the similarity measure *S* between two fuzzy intervals can be defined as Equation (7).

$$S(x_i,(u_i,v_i)) = S((x_i^L,x_i^U),(u_i,v_i))$$

$$=\frac{Overlap((x_i^L,x_i^U),(u_i,v_i))}{Max(x_i^U-x_i^L,v_i-u_i)} \qquad (7)$$

In Equation (10), if two fuzzy intervals are all point data, then the denominator part of the equation $Max(x_i^U-x_i^L,v_i-u_i)$ becomes zero. Therefore we define the similarity measure in this case as Equation (8). As shown in the equation, the similarity value is 1.0 when two intervals are an identical point, and 0 when they indicate two different points.

$$S((x_i^L,x_i^U),(u_i,v_i)) \;=\; \begin{cases} 1 & if \quad (x_i^L = x_i^U = u_i = v_i) \\ 0 & \quad Otherwise \end{cases} \qquad (8)$$

But if $Max(x_i^U-x_i^L,v_i-u_i)$ is greater than zero, the value is determined as described in Equation (9).

$$Overlap((x_i^L,x_i^U),(u_i,v_i)) \;=\; \begin{cases} x_i^U-u_i & if\,(x_i^L\le u_i\le x_i^U\le v_i) \\ v_i-\;u_i & if\,(x_i^L\le u_i\le v_i\le x_i^U) \\ x_i^U-x_i^L & if\,(u_i\le x_i^L\le x_i^U\le v_i) \\ v_i-x_i^L & if\,(u_i\le x_i^L\le v_i\le x_i^U) \\ 0 & Otherwise \end{cases} \qquad (9)$$

The second measure *RF2* can be defined in terms of *RF1* as shown in Equation (10). In the equation, *Li* is the number of feature values which belong to *i*-th feature.

$$RF2(X_i,C_k)=\frac{1}{L_i}\sum_{x_l\in X_i} RF1(x_l,C_k) \qquad (10)$$

The *RF2* shown in Equation (10) represents the degree of importance of a feature in classifying a given pattern class. Therefore it can be utilized to select a more relevance feature set for the sign language recognition.

VI. Experimental Results

Six types of sign pattern classes (*greeting, meet, depart, glad, thank you*, and *very*) have been considered for the experiments. As shown in Fig. 3, three types of features motion energy data, motion history volume, and hand-shape features, are used for the system.

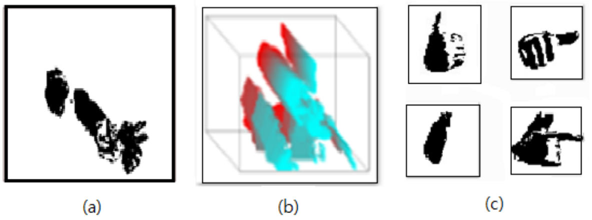

Fig. 3. The feature sets used in the experiments: (a) motion energy data, (b) motion history volume, (c) hand-shape data

The first experiment is to evaluate the recognition rate. A total of 120 patterns, 20 patterns for each class, have been used for the experiment. Table I shows sign language recognition results of the 120 sign patterns. The recognition rates are compared for three different models, FMM, MLP and the proposed model. The same feature sets have been used each model. The proposed model have got better recognition results for the given data than the other models by 1.6% ~4.9%.

TABLE I: Recognition Rates for Three Models: FMM, MLP and WFMM

Classifier	FMM	MLP	WFMM
P1	80	75	80
P2	65	60	70
P3	75	75	80
P4	75	75	75
P5	90	80	90
P6	80	80	80
average	77.5	74.2	79.1

The second experiment is to evaluate the usefulness of the feature analysis technique. We have named the pattern set as P = {p1, p2, p3, p4, p5, p6}. We have used 25 motion energy features , 27 motion history features and 30 hand shape features which are named as E = {e1, e2, …. , e25}, M = {m1, m2, …. , m27}, and H = {h1, h2, …. , h30}, respectively. Table II shows the feature analysis results. As shown in the table, 3 most relevant features are listed for each class.

TABLE II: Relevance Factors between Features and Pattern Classes

Sign pattern class	Relevant features
p1 (greeting)	RF(e12, p1)=0.31, RF(m24,p1)=0.37, RF(h20,p1)=0.34
p2 (meet)	RF(e13, p2)=0.29, RF(m23,p2)=0.39, RF(h03,p2)=0.41
p3 (depart)	RF(e20, p3)=0.24, RF(m27,p3)=0.41, RF(h30,p3)=0.38
p4 (glad)	RF(e09, p4)=0.25, RF(m22,p4)=0.33, RF(h11,p4)=0.44
p5 (thank you)	RF(e24, p5)=0.33, RF(m24,p5)=0.37, RF(h12,p5)=0.37
p6 (very)	RF(e12, p6)=0.25, RF(m22,p6)=0.39, RF(h08,p6)=0.45

We have conducted another experiment to evaluate the usefulness of the feature analysis technique. We have compared the pattern recognition rate as varying the number of selected features. Fig.4 shows the result of the experiment. As shown in the figure, 38 less relevant features out of the 82 total features can be removed with less than 3% degradation of recognition rate. This result shows that we can reduce the number of features to improve the computation speed without performance degradation.

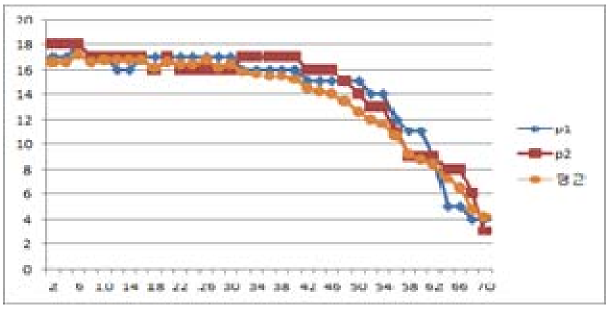

Fig. 4. Pattern recognition rate as varying the number of selected features

VII. Conclusion

A hybrid neural network for sign language recognition is described. The proposed system is implemented using two types of neural networks, a modified convolutional neural network and a weighted fuzzy min-max network. The basis of the data representation is a motion history volume which is built by stacking regions of motion information. The modified CNN is a bio-inspired feature extractor with three-dimensional receptive fields that can accept a spatiotemporal template. One of the major advantageous features of convolutional neural network is invariant detection capability for distorted patterns in the pattern data. The suggested model tolerates partial variation of feature points among the spatial and time dimension.

The weighted FMM neural network model is capable of utilizing the feature distribution and the frequency factor in the learning process as well as the classification process. Since the weight factor effectively reflects the relationship between feature range and its distribution, the system can prevent undesirable performance degradation which may be caused by noisy patterns. Consequently the proposed model can provide more robust performance of pattern classification in case that the training data set in a given problem includes some noise patterns or unusual patterns.

Acknowledgment

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation

References

[1] M. M. Zaki and S. I. Shaheen, “Sign language recognition using a combination of new vision based features,” *Pattern Recognition Letters*, vol.32, pp.572-577, April 2011

[2] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp.873-891, June 2005.

[3] R. Yang and S. Sarkar, “Coupled grouping and matching for sign and gesture recognition,” *Computer Vision and Image Understanding* vol.113, pp.663-581, Dec. 2009.

[4] 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.

[5] A. Yilmaz and M Shah, “Actions sketch: A novel action representation,” *The Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 984-989, 2005.

[6] C. Garcia, M. Delakis, “Convolutional face finder: A neural architecture for fast and robust face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp.1408-1423, Nov. 2004

[7] P. K. Simpson, “Fuzzy min-max neural network Part1: Classification.” *IEEE Transaction on Neural Network*, vol.3, pp.776-786, May 1992.

[8] B. Gabrys, A. Bargiela, “General fuzzy min-max neural network for clustering and classification,” *IEEE Transactions on Neural Networks*, vol. 11, pp. 769-783, March 2000.

[9] H. Kim, J. Lee, H. Yang, “A weighted FMM network and its application to face detection,” *Lecture Notes in Computer Science*, vol. 4233. pp. 177-186, Oct. 2006.

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 Technology 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.

So-Jeong Park was born in Pohang, Korea, May 6, 1986. She received her B.S. degree from Uiduk University, Gyeongju, Korea in 2009.

She is currently doing master’s degree in computer science at Handong Global University, Pohang, Korea. Her research interests include image processing, multimedia communication and computer vision.

Seung-Kang Lee was born in Gunpo, South Korea, July 4, 1987. Seung-Kang received a bachelor’s degree of in computer engineering from Handong Global University, Pohang, Kyeongbuk, Korea in 2011.

He is currently studying for a master’s course in Handong Global University, Pohang, Kyeongbuk, Korea. He is interested in computer vision and image processing.