Off-line Recognition of a Handwritten Chinese Zither Score

Yi-Hung Liu[**] and Han-Pang Huang[*]

Robotics Laboratory, Department of Mechanical Engineering
National Taiwan University, Taipei 10660, TAIWAN
TEL/FAX: (886) 2-23633875, E-mail: hphuang@w3.me.ntu.edu.tw
[*] Professor and correspondence addressee, [**] Graduate student

Abstract

A Chinese zither score is different form a western staff. The Chinese zither score is handwritten, and is a combination of fingerings, scales, and several different types of notes. In this paper, we first construct pattern classes for fingerings and scales we frequently play. A specific segmentation method is derived in accordance with the zither score. After segmentation, all meaningful individuals can be found out and the weighted cross counting feature is used to extract features. A cascaded architecture of neural network with feature map (CANF) is proposed to obtain high recognition rates. The CANF cascades a supervised neural network trained by back propagation (BPNN) with an unsupervised neural network, Kohonen's self-organized feature map (SOFM). The SOFM can reduce the dimension of feature space and remove the redundancy of features in transformation such that the learning time of BPNN can be speeded up and the recognition rate can be improved. In our experiment, a real Chinese zither score is segmented, and the CANF shows a 100% perfect recognition rate.

Keywords: Chinese zither score, recognition, SOFM, fuzzy segmentation, neural network.

1. Introduction

Handwritten recognition is a popular and important topic in research fields of pattern recognition. Plamondon and Srihari [1] gave an overall survey for on-line and off-line handwriting recognition. Gader successfully used the hybrid fuzzy-neural method to overcome the word recognition [2][3]. These handwritten applications can roughly be classified into several fields: zip code recognition, Arabic numerals recognition, English character recognition, Chinese character recognition, and Japanese character recognition. This paper aims to recognize the Chinese zither score.

The Chinese zither, also called zheng, has existed for two thousand years in China since the Chin Dynasty. The score of Chinese zither is in a special form [4]. It is different from a western music staff. In the past, the scores were handwritten. Although some scores are printed in recent years, most Chinese zither scores are still handwritten. A zither score consists of many symbols, such as fingerings, scales, meter lines, bar lines, extended lines octachord notes,.., etc. For example, a fingering is usually placed on the scale, and a meter line is located below the scale. A bar contains 2, 3, or 4 beats, and a row contains 4 bars or more. Since only fingerings and scales will be recognized, other symbols should be filtered out in the segmentation.

Segmentation is an important process in the general document analysis since unsuitable segmentation will cause a wrong recognition result especially in handwritten symbols [5]. The first task of recognizing a Chinese zither score is to segment the score into meaningful individuals containing fingerings and scales and then pass them to the recognition process. In the segmentation process, the fuzzy set theory is incorporated into the column segmentation so that undesired patterns can be filtered out before classification.

Another important process is the feature extraction. The weighted cross counting features (WCCF) are selected as features in this paper. The WCCF is robust to variations in handwritten symbols [6]. Finally, a cascaded architecture of neural network with feature map (CANF) is proposed to classify the fingerings and scales. The CANF is composed of two neural networks: the Kohonen's self-organized feature map (SOFM) and the supervised multi-layer feedforward neural network trained by standard back propagation (BPNN). It will be shown that the proposed algorithm can achieve perfect recognition in the experiments.

2. Construction of Pattern Class

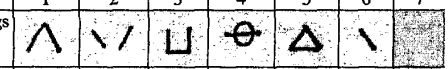
The numbers of fingerings in a real zither score are more than 20 since a Chinese zither has 16 or 21 strings [7]. For a simple zither score, there are few fingerings. Six kinds of fingerings, which are most frequently played by beginners, are selected in this paper. In the Chinese zither score, the scales are represented in handwritten Arabic numerals. For instance, the scales do, re, mi, fa, so, la, si, are represented by 1, 2, 3, 4, 5, 6, and 7, respectively. The pattern classes of fingerings and scales used for recognition in this paper are arranged in Table 1.

3. Segmentation of the zither score

In the document recognition field, the segmentation of a word into a string of characters has been well

Table 1. Classes of fingerings and scales.

0-7803-7087-2/01/$10.00 © 2001 IEEE

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Fingerings							
Scales	1	2	3	4	5	6	7

developed. The signature segmentation is used to do this work. The signature, which is a projection, is the histogram of nonzero pixels of the resultant masked image. Signature analysis was first used in printed character recognition. Sanz and Dinstein [8] discussed one kind of pipeline architectures to compute projections and projection-based geometric features.

However, such a segmentation algorithm cannot be used very well in the handwritten Chinese zither score since the following several problems may occur during the segmentation procedure of a handwritten Chinese music score.

1. The length of each segment of the projection may be long, short or even a unit, i.e., a pixel long, no matter the projection is vertical or horizontal projections. A traditional signature segmentation is not suitable. In fact, the segments of projections whose lengths are nonzero need not be segmented. For example, a bar line "|" and " ¦ " have similar segments in the vertical projection. However, "|" is not a desired pattern since it is not a fingering. But " ¦ " is a desired pattern for the recognition. If we can decide what regions or what patterns we need in the recognition procedure, then we can separate them efficiently during the segmentation procedure.
2. Noise usually causes a problem in the recognition process regardless printed or handwriting documents. Noise may appear in several ways: an isolated point, line-like noise, arc-like noise, ..., etc. These noises must be rejected finally in the recognition results if they had been segmented in the segmentation procedure. But, noises may result in wrong recognition due to ambiguous shapes. There are several types of filters for a spatial image, such as smoothing filter, median filter, temporal filter and high pass filter. Noise removal is usually done in the image preprocessing after the thresholding of the image. But it takes time and may not clean all the noise. In order to remove undesired noise, a good algorithm should be used.

In this paper, the fuzzy set theory is incorporated into the step of segmentation to solve the above problems. In the fuzzy segmentation process, those unwanted symbols and noise can be eliminated. A set of fuzzy rules is designed to decide which segment is preserved (separated) or canceled. In the meantime, some regions, such as noises, in the spatial image will not be segmented. This decreases a lot of time of noise removal in the preprocessing. The segmentation procedures are illustrated below.

Step 1. Row segmentation: A Chinese zither score may have several rows in a page, and these rows are nearly parallel. The first step in the segmentation is to segment these rows. The method of row segmentation in this paper is based on the horizontal projection. Suppose that there are n rows. The rule of the row segmentation is defined in terms of the length of a segment along the histogram as:

$$\textit{The row is} \begin{cases} \textit{preserved} & \textit{if } SLR_i \geq \frac{1}{2}\max_i SLR_i \\ \textit{canceled} & \textit{otherwise} \end{cases} \quad (1)$$

where the SLR_i denotes the segment length of the row i.

Step 2. Fuzzy Column Segmentation: After a row has been segmented through row segmentation, the fuzzy column segmentation is followed. It segments all the individuals that have meaning for playing the zheng in the row. Suppose that there are j individuals in the row i. These j individuals include bar lines and noise except meaningful individuals. Let SL_{ij} and SH_{ij} be the j_{th} segment length and maximum height along the histogram of the vertical projection of the row i, respectively. The following linguistic rules are used for determining which can be preserved or canceled:

If $SL_{ij}$is large and SH_{ij} is large, then S_{ij} is preserved.
If $SL_{ij}$is large and SH_{ij} is medium,then S_{ij} is preserved.
If SL_{ij} is large and SHij is small, then S_{ij} is preserved.
If $SL_{ij}$is medium and SH_{ij} is large,then S_{ij} is preserved.
If $SL_{ij}$is medium and SH_{ij} is medium,then S_{ij} is preserved.
If SL_{ij} is medium and SH_{ij} is small, then S_{ij} is canceled.
If SL_{ij} is small and SH_{ij} is large, then S_{ij} is canceled.
If SL_{ij} is small and SH_{ij} is medium, then S_{ij} is preserved.
If SL_{ij} is small and SH_{ij} is small, then S_{ij} is canceled.

Note that S_{ij} denotes the j_{th} segment on the vertical projection of the row i. The membership functions are shown in Fig. 1.

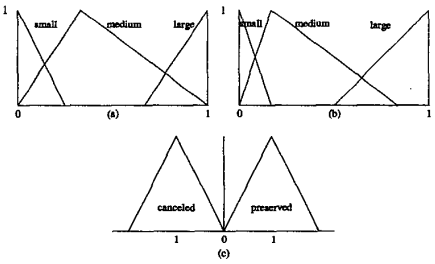


Fig.1 (a) SL, (b) SH, (c) output membership functions for column segmentation

Before fuzzification, the normalized factors for SL and SH are defined according to their maximum length and maximum height for all segments in the vertical projection of the row i as

$$NSL_i = \max_j \{SL_{ij}\}$$
$$NSH_i = \max_j \{SH_{ij}\} \quad (2)$$

In the inference process, min operation is used as the

conjunction of the proposition. The output y_o of the defuzzifier is obtained from the centroid of the weight as

$$y_o = \frac{\int \mu(y) \cdot y \, dy}{\int \mu(y) dy} \quad (3)$$

Whether the segment (or individual) S_{ij} is preserved or not, it can be determined by the decision rule

$$S_{ij} \text{ is } \begin{cases} preserved, & \text{if } y_o \geq 0 \\ canceled, & \text{if } y_o < 0 \end{cases} \quad (4)$$

However, a discontinuous region on the spatial image may cause two or more sub-segments on the vertical projection. For example, an unsuitable thresholding value may divide an extended line into two horizontal lines from the vertical direction after binarizing. Such two sub-segments are not the expected results. One of the two segments or both may be canceled by the fuzzy column segmentation since their lengths are shorter than the length of the extended line. For avoiding the discontinuous break, a connectivity rule is defined as follows

$$\text{if } SP_{i(j+1)} - EP_{ij} \leq \varepsilon, \text{ then } S_{ij} = SP_{ij}, ..., EP_{ij}, SP_{i(j+1)}, ...EP_{i(j+1)} \quad (5)$$

where SP_{ij} and EP_{ij} are the starting point and the end point of the segment j on the row i, respectively. The $SP_{i(j+1)}$ and $EP_{i(j+1)}$ are the starting point and the end point of the segment $j+1$ on the row i, respectively.

Step 3. Decomposition of Half Notes: Those individual sets segmented by the fuzzy column segmentation usually include several notes with one beat, couples of half beat and extended lines. This step is to segment a couple of half notes that are connected by a meter line into sub-individuals so that each individual contains only one scale. The histogram distribution of the vertical projection of a couple of half notes is represented by a shape which has twin peaks. A couple of half notes can be found by the following rule

$$S_{ij} \text{ is a couple of half notes if } SL_{ij} \geq \frac{3}{2} \min_j SL_{ij} \quad (6)$$

Upon finding the couple of half notes, these two half notes can be separated (or segmented) by a cut-off point (COP). Suppose that the length of S_{ij} is $1,...,k,...,SL_{ij}$

$$COP = \arg\min_k SH_{ij}(k), \quad k = \frac{1}{4} SL_{ij}, ..., \frac{3}{4} SL_{ij} \quad (7)$$

where $SH_{ij}(k)$ is the height of segment S_{ij} at point k.

After segmenting all the individuals by the proposed segmentation method, a top-down search, instead of traditional connected component method [9], is used to extract fingerings, scales, meter lines, and octachord notes since the connected component may extract wrong regions. For example, a fingering of class 2 may become two regions after the connected component search no matter 4-connected or 8-connected is used.

4. Feature Extraction

Before selecting and designing the algorithm for feature extraction, we must realize the characteristics of Chinese music scores. Usually, they have an important property: They are handwriting. It implies the variation characteristics of a stroke. Based on this fact, the weighted cross counting feature (WCCF) is proposed for feature extraction. In WCCF , the weight is added to the traditional stroke cross counting features. The dimension of the feature vectors in BCCF is the same as in SCCF but more robust against the variations of handwritten Chinese music notes.

Let $f(x,y)$ denote the pixel value at position (x,y) on a 40× 40 binary image which has been segmented.

$$f(x,y) = \begin{cases} 1, & \text{if } (x,y) \text{ is a black pixel} \\ 0, & \text{if } (x,y) \text{ is a white pixel} \end{cases} \quad (8)$$

Assume that several rasters scan these score notes in four directions: vertical, horizontal and two orthogonal diagonals. The cross counting along a raster is defined as the number of times of the raster crosses the strokes of the music notes. Since there are many variations among different handwritings, the features extracted along the same fixed raster may not be invariant. A boundary search concept can resolve the variation problem of a stroke in a pixel window because little variation can occur in a small range. Therefore, while keeping the rasters in SCCF as main rasters, the neighbors of each raster are also taken into consideration. A weighted vector is used to include the contributions of the eight neighbors of each main raster so that the extracted feature can represent more information of a music note. The idea is similar to the notion of the "membership grade" in a fuzzy set, but it is 1-dimension input. The weighted vector, which is a 1× 9 row vector, is defined as :

$$W = [0.2, 0.4, 0.6, 0.8, 1.0, 0.8, 0.6, 0.4, 0.2]$$

The weighted cross counting features along a raster I_i at the horizontal and vertical directions are defined as

$$HWCCF = \frac{1}{10} \sum_{K=-4}^{4} W[k+5] \sum_{j=1}^{N} f(I_i + k, j) \cdot \overline{f(I_i + k, j+1)} \quad (9)$$

$$VWCCF = \frac{1}{10} \sum_{k=-4}^{4} W[k+5] \sum_{i=1}^{N} f(i, J_j + k) \cdot \overline{f(i+1, J_j + k)} \quad (10)$$

where N=40, I_i=8i for i=1,...,4, and J_j=8j for j=1,...,4. $\overline{f(I,j)}$ is the negative function, i.e.

$$\text{If } \quad f(I,j) = 1, \text{ then } \overline{f(I,j)} = 0 \quad (11)$$

Similarly, the weighted cross counting features taken in the direction of the two diagonal (right slant and left slant) are given by

$$RSWCCF_i = \frac{1}{12} \sum_{k=-4}^{4} W[k+5] \sum_{j=1}^{N-8i} f(I_i + k + j, j) \cdot \overline{f(I_i + k + j + 1, j+1)} \quad (12)$$

$$RSWCCF_j = \frac{1}{12}\sum_{k=-4}^{4} W[k+5]\sum_{i=1}^{N-8j} f(i+1, i+J_j+k)\cdot \overline{f(i, i+J_j+k-1)} \quad (13)$$

where I_i=8i, i=1,...,4 and J_j=8j, j=1,1...4. Another weighted cross counting taken in the left slant diagonal direction (LSWCCF) can be defined in the same way. The numbers of features in horizontal, vertical, right slant diagonal, and left slant diagonal, are 4, 4, 8, 8, respectively. Therefore, there are total 24 features after taking the feature extraction to a 40 by 40 pixel matrix.

5. Cascaded Architecture of Neural Network with Feature Map (CANF)

After the features are obtained from the four directional WCCF, these features will be sent to the Cascaded Architecture of Neural Network with Feature Map (CANF). It is a cascaded architecture composed of two neural networks. One is the unsupervised Kohonen's SOFM, the other is the supervised multi-layer BPNN which is fully connected. Figure 2 shows detailed information of the architecture. There are 4 channels in the CANF. Each channel represents different directional WCCF, i.e., HWCCF, VWCCF, RSWCCF, and LSWCCF. For example, channel 1 is the HWCCF, which has 4 features as the input of the Kohonen's SOFM and the output is the 2-D coordinates (on the x and y axes) in the 2-D topological net. After the feature map, the 2-D coordinates are the input of the BPNN. In this way, there are total 8 input nodes of the BPNN due to 4 channels. The cascaded architecture can compensate for the inaccuracy of unsupervised neural network and long training time of supervised neural network. Another important contribution of the proposed cascaded architecture is to reduce the input dimension (feature space) of the classifier. Furthermore, the input selection is done by the Kohonen's 2-D SOFM. Although the concept of cascading a supervised neural network with an unsupervised neural network is not a new concept to reduce the training time of a supervised neural network [10], the proposed architecture is a new concept for dealing with the reduction of independent feature spaces in handwritten recognition research field. In addition, the Kohonen's SOFM can reduce the feature dimension. It can also map a space with variations onto a cluster in the 2-D map by the concept of neighborhood [11]. This offers us a great benefit for handwritten recognition.

5.1 Reduction of Input Dimension by SOFM

The 2-D Kohonen's SOFM can keep the neighborhood relations of the input pattern since it learns under the topology-preserving map [12]. Therefore, the 2-D SOFM is also called a topology-preserving map. The merit is useful for clustering handwritten fingerings and scales. In addition, the Kohonen's SOFM can be used to remove the redundancy in feature space or noise information in the input pattern, and avoid the trial and error approaches, such as genetic algorithm or KL expansion. In other words, the 2-D Kohonen's SOFM reduces the input dimension from a high dimension to a 2-D coordinates.

The supervised SOFM can find a winner on the 2-D map to represent the original pattern which is in high dimension. The winner c can be determined by finding minimum of the Euclidean distance, i.e.,

$$y_c = \begin{cases} 1, \textit{if } \| X(t) - W_c(t) \| \\ \quad = \min_{i=1}^{k} \| X(t) - W_i(t) \| \\ 0, \textit{otherwise} \end{cases} \quad (14)$$

where $X(t)$ is the current input pattern vector, and $W(t)$ is the weight vector.

5.2 Classification by Supervised neural network

After reduction of the input space using Kohonen's SOFM, the four sets of 2-D coordinates (8 newly features) are fed into the BPNN for classification. There are two CANF modules. One is for fingerings' recognition, the other is for scales' recognition. They have the same architecture but different system parameters.

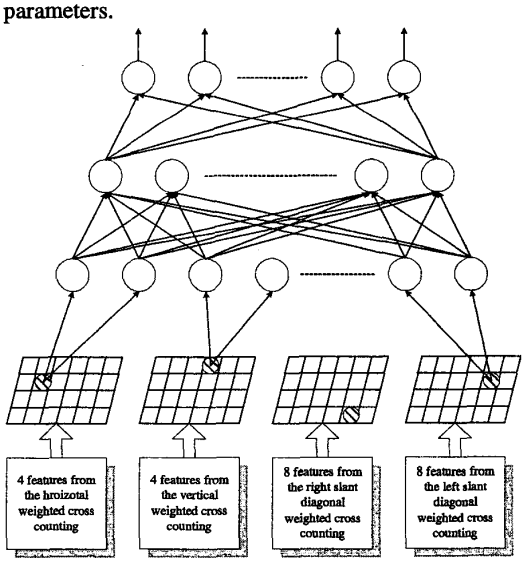


Fig. 2. System architecture of the CANF

6. Experimental Results

6.1 Segmentation results

In our experiment, a real handwritten Chinese zither score, 笑傲江湖, is considered. The digital camera is used to get the bitmap file of the score. The proposed segmentation algorithm is employed to separate all the individuals that are need to be recognized from the score. The bitmap of the

handwritten Chinese zither score is shown in the middle of Fig. 3. After binarization, the segmentation results of each step of the score are shown in Fig 3. For verifying the validity of the segmentation methodology, some noises are deliberately added on the score. The segmentation results show that step 1 and step 2 had separated all the rows and individuals successfully regardless man-made noises. In the step 2, twenty three meaningful individuals are found by the fuzzy column segmentation, and those meaningless individuals for playing the Chinese zither are filtered out. The task of segmentation is not finished yet because there are still 2 and 4 individuals which contain two sub-individuals in row 1 and row 2, respectively. Therefore, the step 3 is performed to accomplish the task. From Eq.(7), the COPs of the six individuals are determined and then the six individuals are segmented into 12 sub-individuals. Finally, there are totally 29 individuals are found out after the three segmentation steps.

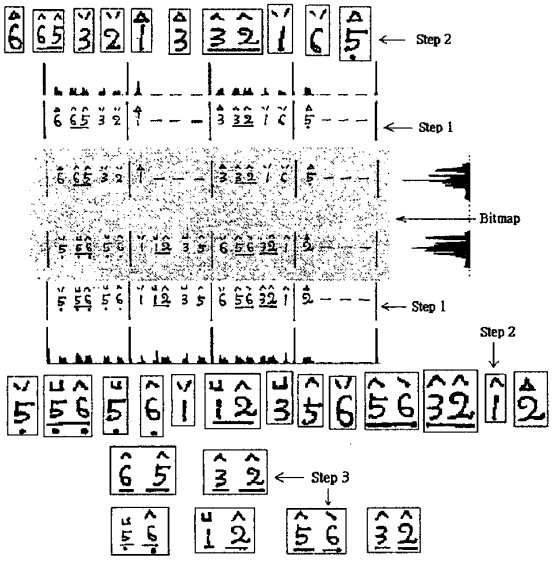


Fig.3. Segmentation results

6.2 Fingerings and scales recognition results

Before the experiments of recognition, we prepare 20 training patterns and 20 testing patterns for each class with variations. These patterns are obtained from digital camera and saved as bitmap files. After binarizing and thinning, these binary pixel matrices are normalized to 40 by 40 pixel matrices and then perform the feature extraction to obtain WCCF of each of them.

Table 2 shows recognition results of the *k*-nearest neighborhood (KNN) method and BPNN. For avoiding a tie among classes [13], an odd value k=7 is chosen in the KNN. The inputs of BPNN and KNN are all features (24 WCCFs) without going through the Kohonen's SOFM. In this experiment, the BPNN has one hidden layer with 15 hidden nodes, and the learning rate is set as 1.0. The good results from Table 2 give us an information: although the input dimension is high, the WCCFs are remarkable for classifying all the classes with variations when a BPNN or a KNN is used. Table 3 shows the comparison of recognition results between the KNN+SOFM and the CANF. In this experiment, the output nodes are arranged into a 2-D net, i.e., a 20 by 20 neuron matrix. The initial neighborhood size is set as 10, and the reduction rate of the neighborhood size is 0.01/cycle. The feature dimension is reduced by the 2-D SOFM, i.e., inputs of KNN and CANF are 4 sets of 2-D coordinates.

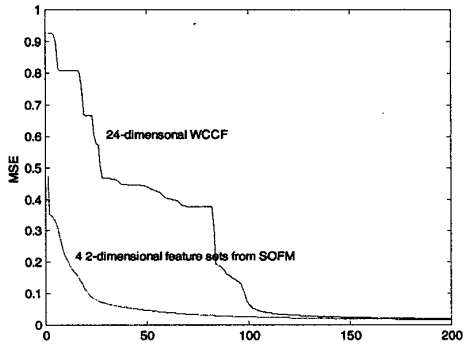


Fig 4. MSE of BP against CANF for scales.

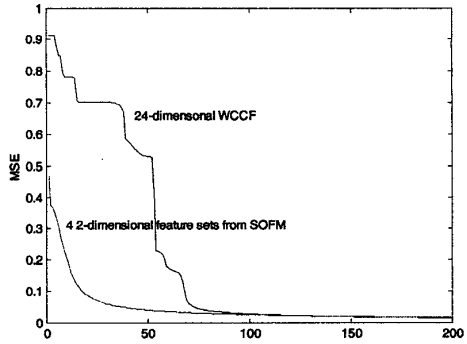


Fig. 5. MSE of BP against CANF for fingerings.

Table 2. Comparison of recognition results between BP and KNN for all 24 WCCFs input (%)

		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Average
BPNN	*Fingerings*	100	100	95	100	100	85		96.67
	Scales	95	100	95	90	90	95	95	94.29
KNN	*Fingerings*	90	95	90	90	90	80		89.17
	Scales	90	90	80	80	80	85	90	85

Table 3. Comparison of recognition results between SOFM+KNN and CANF (%)

		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Average
SOFM	*Fingerings*	95	100	100	100	95	85		95.83
+ KNN	*Scales*	100	100	95	95	90	90	95	95
CANF	*Fingerings*	100	100	100	100	100	100		100
	Scales	100	100	100	100	100	100	100	100

The hidden layer of BPNN in CANF contains 8 hidden nodes only, which is less than the pure BPNN. The results in Table 3 show that the recognition rates of SOFM+KNN become higher than the results of KNN in Table 2. In addition, the average recognition rate of CAFN is up to 100%. Fig. 4 and Fig. 5 show the MSE of BPNN and CANF. Obviously, no matter for fingerings or for scales, the time of convergence of CANF are much shorter than BPNN. It proves that Kohonen's SOFM can actually speed up the learning time of the BPNN after transforming the feature dimension and maintaining the relations of neighborhood of input patterns since the recognition rates for all classes are 100%. By the comparison of recognition rates of BPNN and CANF, it shows that the redundancy and noise have been removed in the learning period of the unsupervised SOFM since recognition rates in some classes are not 100% in the BPNN, but 100% for each class in the CANF.

7. Conclusions

In this paper, an off-line recognition algorithm for a handwritten Chinese zither score is proposed. The algorithm consists of fuzzy segmentation, WCCF feature extraction, and CANF recognition. A real Chinese zither score is conducted in the experiment. The experimental results show that the proposed algorithm can achieve a 100% recognition rate.

References

[1] R. Plamondon, and S.N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no.1, pp. 63-84, 2000.

[2] J. H. Chiang, P. D. Gader, "Hybrid fuzzy-neural systems in handwritten word recognition," *IEEE Trans. on Fuzzy Systems*, vol. 5, no. 4, pp. 497-510, 1997.

[3] P. Gader, M. Mohamed, and J.H. Chiang, "Comparison of crisp and fuzzy character neural networks in handwritten word recognition," *IEEE Trans. on Fuzzy Systems*, vol. 3, no. 3, 1995.

[4] T. R. Huang, *Collection of Chinese zither scores*, Taipei: T. R. Huang, 1975.

[5] S.W. Lee, S.Y. Kim, "Integrated segmentation and recognition of handwritten numerals with cascade neural network," *IEEE Trans. on System, Man, and Cybernetics*, vol. 29, no. 2, pp. 285-290, 1999.

[6] Y.H. Liu, H.P. Huang, "Handwritten Chinese music score recognition using a two-stage neural network architecture," *Proceedings of Sixth International Conference on Automation Technology*, vol.2, pp. 861-866, 2000.

[7] P. Chao, *Chinese Musical Instrument: Topic on Chinese Zither*, Taipei: Department of Culture Development, Executive Yuan, 1985.

[8] J. Sanz, I. Dinstein, "Projection-based geometrical feature extraction for computer vision: Algorithms in pipeline Architecture," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 160-168., 1987.

[9] R. M. Haralick, L. G. Shapiro, *Computer and Robert Vision: Volume I*, New York: Assidon-Wesley Publishing, 1992.

[10] C.T. Lin, Y.C. Lee, and H.C. Pu, "Satellite sensor image classification using cascaded architecture of neural fuzzy network," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 38, no. 2, 2000.

[11] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[12] C.T. Lin, C.S.G. Lee, *Neural Fuzzy Systems: A Neural-Fuzzy Synergism to Intelligent System.* Englewood Cliffs, NJ: Prentice-Hall, 1996.

[13] J.M. Keller, M.R. Gray, and J.A. Givens, JR., "A fuzzy *k*-nearest neighbor algorithm," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 15, no. 4, pp. 580-585, 1985.