A Chinese Minority Script Recognition Method Based on Wavelet Feature and Modified KNN

Hai Guo
Dalian Nationalities University, Dalian, China
Email: guohai@dlnu.edu.cn

Jing-ying Zhao*
Dalian Nationalities University, Dalian, China
Email: dalianzjy@gmail.com

Abstract—In recent years, K-Nearest Neighbor (KNN) has demonstrated excellent performance in a variety of pattern recognition problems. In this paper, we proposed modified KNN for the Chinese minority scripts classification, using wavelet energy distribution and wavelet energy proportions features generated from the discrete wavelet multi-resolution decomposition. The experimental results show that the method achieves high accuracy in testing dataset that consists of Chinese, English and Chinese minority scripts such as Tibetan, Tai Lue, Naxi Pictographs, Uighur, Tai Le, Yi. The results also show that this method is feasible and reasonable, and the recognition rate is up to 96%.

Index Terms—Chinese minority script; scripts recognition; wavelet analysis; Modified k-Nearest Neighbor (mk-NN)

I. INTRODUCTION

With the development of computer technology, the information processing on Chinese Minority Script grows up gradually. Such as the collaborative research by Northwest University for Nationalities and Tsinghua University on Tibetan character recognition[1], the multi-speech (Uyghur, Kazak and Kyrgyz ) recognition by Xinjiang University[2], the Mongolian recognition by Inner Mongolia University[3] and the NaXi pictographs information processing[4-6], all of these have made great progress. Script recognition is to estimate the type of script in the text image—one important research in automatic document analysis and information processing. The existing Chinese Minorities OCR system is mainly oriented in the "literacy" level, the script recognition has not attracted the attention it deserves, this paper fills a domestic gap in the field. This paper presents a method of recognizing the kinds of Chinese minority scripts based on wavelet analysis[7-8] and K-Nearest Neighbor (KNN)[9-11].

II. DISCRETE WAVELET TRANSFORMATION

The Discrete Wavelet Transform (DWT), which is based on sub-band coding is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required.

The foundations of DWT go back to 1976 when techniques to decompose discrete time signals were devised. Similar work was done in speech signal coding which was named as sub-band coding. In 1983, a technique similar to sub-band coding was developed which was named pyramidal coding. Later many improvements were made to these coding schemes which resulted in efficient multi-resolution analysis schemes [12-14].

In CWT, the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales [15].

Filters are one of the most widely used signal processing functions. Wavelets can be realized by iteration of filters with rescaling. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by up sampling and downsampling (subsampling) operations [16-17].

The DWT is computed by successive lowpass and highpass filtering of the discrete time-domain signal as

\[ X[n] \xrightarrow{H_0} \{ d_2[n], c_2[n] \} \xrightarrow{H_2} \{ d_4[n], c_4[n] \} \]

\[ X[n] \xrightarrow{G_0} \{ d_2[n], c_2[n] \} \xrightarrow{G_2} \{ d_4[n], c_4[n] \} \]

\[ X[n] \xrightarrow{H_0} \{ d_2[n], a_2[n] \} \xrightarrow{H_2} \{ d_4[n], a_4[n] \} \]

\[ X[n] \xrightarrow{G_0} \{ d_2[n], a_2[n] \} \xrightarrow{G_2} \{ d_4[n], a_4[n] \} \]

Figure 1 Three-level wavelet decomposition tree

*Corresponding author: Jing-ying Zhao Email: dalianzjy@gmail.com

© 2010 ACADEMY PUBLISHER
doi:10.4304/jsw.5.2.251-258
shown in figure 1. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time multiresolution to discrete-time filters. In the figure, the signal is denoted by the sequence $x[n]$, where $n$ is an integer. The low pass filter is denoted by $G_0$ while the high pass filter is denoted by $H_0$. At each level, the high pass filter produces detail information, $d[n]$, while the low pass filter associated with scaling function produces coarse approximations, $a[n]$.

At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. In accordance with Nyquist’s rule if the original signal has a highest frequency of $\omega$, which requires a sampling frequency of $2\omega$ radians, then it now has a highest frequency of $\omega/2$ radians. It can now be sampled at a frequency of $\omega/2$ radians thus discarding half the samples with no loss of information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. Thus, while the half band low pass filtering removes half of the frequencies and thus halves the resolution, the decimation by 2 doubles the scale. With this approach, the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients, $a[n]$ and $d[n]$, starting from the last level of decomposition.

The decomposition of an image using discrete wavelet transform comprises of a chosen low pass and a high pass filter, known as Analysis filter pair. The low pass and high pass filters are applied to each row of data to separate the low frequency and the high frequency components. These data can be sub-sampled by two. The filtering is then done for each column of the intermediate data finally results in a two dimensional array of coefficients containing four bands of data, known as low-low (LL), high-low (HL), low-high (LH) and high-high (HH). Each coefficient represents a spatial area corresponding to one-quarter of the original image size. The low frequencies represent a bandwidth corresponding to $0<|\omega|<\pi/2$, while the high frequencies represent the band $\pi/2<|\omega|<\pi$. It can be possible to decompose the LL band in the same way up to any level, resulting in pyramid-structured decomposition as shown Fig 2. The LL band at the top of the pyramid containing approximate coefficients holds the most significant information and the other bands containing details coefficients have lesser significance. Thus the degree of significance is decreasing from the top of the pyramid to the bands at the bottom.

III. FEATURE EXTRACTION BASED ON WAVELET ANALYSIS

In digital image processing, after dispersing the wavelet transform, the multi-resolution analysis is done on the images, and then the texture features are obtained by means of extracting high frequency information.

Chinese minority script recognition mainly utilizes two feature vectors, the wavelet energy feature and the wavelet energy proportion feature. The definitions of average energy of images, the wavelet energy feature and the wavelet energy proportion feature are introduced firstly.

A. Average Energy of Images

Average energy of an $N\times N$ image is defined as:

$$f = \sum_{m=1}^{N} \sum_{n=1}^{N} f(m,n) \frac{f^2(m,n)}{N^2}, \quad m, n = 1, \ldots, N \quad (1)$$

Before the wavelet decomposition the image energy is necessary in normalization processing by Eq. (2).

$$f(m,n) \leftarrow f(m,n) \left(\text{energy} \right)^{1/2}, \quad m, n = 1, \ldots, N \quad (2)$$

B. Wavelet Energy Feature

After frequency domain resolving on image we can get four detail subgraphs, LL , LH , HL and HH. The average energy distributing of he subgraphs on the k-order wavelet decomposition is defined (3)—(6), where $\text{ELH}$, $\text{EHH}$ and $\text{EHL}$ are the multi-scale wavelet features of the images. The $F_d$ in the Eq. 15 is defined as wavelet energy feature vector[18-19].

![Two-dimensional wavelet transform](image_url)
Wavelet energy feature include three chief apartments, \( \text{ELH}^{(k)} \), \( \text{EHH}^{(k)} \), \( \text{EHL}^{(k)} \) : through the formula 7 - 9 , we define the \( \text{EPHL}^{(k)} \), \( \text{EPHH}^{(k)} \), \( \text{EPLH}^{(k)} \) , three energy distribution proportions , respectively denoting the energy distribution proportion on k-order wavelet decomposition in the subgraph HL, HH and LH . The above three features consist of \( F_{dp} \) in formula 10, wavelet energy feature [20-23].

\[
\begin{align*}
\text{ELH}^{(k)} &= \frac{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{LH}^{(k)}(m,n) \right)^2}{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{N/2} \right)^2}, \; k=1, \ldots, 1bN \\
\text{EHH}^{(k)} &= \frac{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{HH}^{(k)}(m,n) \right)^2}{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{N/2} \right)^2}, \; k=1, \ldots, 1bN \\
\text{EHL}^{(k)} &= \frac{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{HL}^{(k)}(m,n) \right)^2}{\sum_{m,n=1}^{N/2} \sum_{m,n=1}^{N/2} \left( \text{N/2} \right)^2}, \; k=1, \ldots, 1bN \\
\end{align*}
\]

\[
\begin{align*}
F_{dp} &= \begin{bmatrix} \text{EHL}^{(k)} \\ \text{EHH}^{(k)} \\ \text{ELH}^{(k)} \end{bmatrix}, \; k=1, \ldots, 1bN \\
\end{align*}
\]

C. The Ratio of Energy Distribution

After the wavelet decomposition and the energy extraction in six kinds of Chinese Minority Script (Tibetan, Tai Lue, Naxi Pictographs, Uighur, Tai Le and Yi), English and Chinese , we can obtain each type of scripts’ multi-resolution wavelet energy feature and energy distribution proportion. It is vital to choose the k value in the wavelet decomposition. The classification
effect tends to be bad if the k value is too low, and the feature dimensions greatly increase if the k value is too high so as to reduce the recognition speed. By testing, we found that when k=2, in other words when we extract $F_d$ and $F_{dp}$ on two-level decomposition, the efficiency and accuracy is appropriate. Fig.3a shows the eight different types of scripts’ horizontal characteristic curve in the one-level decomposition.

IV. DESIGN OF THE KNN CLASSIFIER

A. K-Nearest Neighbor Algorithm

K-NN is one of the most widely known and used nonparametric estimation methods [25]–[27], [30]. Because of its conceptual simplicity and relatively easy applicability, $k$-NN is routinely utilized by Finnish and Swedish forest experts for updating their national inventories [24], [27]. A complete description of the procedure can be found in [25] and [26]. Thus, only the main features of the configuration used are described here.

In the $k$-NN estimation procedure, the value of an environmental parameter for a specified pixel $k$-NN is predicted as the weighted average of the parameter’s measurements at the $k$ most spectrally similar reference points $Y_1, \ldots , k$ according to the formula The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its $k$ nearest neighbors. It will be useful to weigh the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

$$Y_{k-NN} = \frac{\sum_i W_i Y_i}{\sum_i W_i}$$

(11)

where $W_i$ are the weights of the $k$ reference points, which may be kept constant or may be computed as inversely proportional to the spectral distance. Recent studies have demonstrated that the two options provide almost equivalent accuracies [25], so that only the simplest strategy (i.e., the computation of simple averages of the $k$ nearest points) was currently implemented and tested.

Other factors that influence the performance of $k$-NN procedures are the consideration of horizontal and/or vertical reference areas and the application of environmental stratifications for the selection of the $k$ points [27], [29]. In the present situation, only the first factor was taken into consideration, looking for the horizontal reference area that was optimal for each $k$-NN application. On the contrary, no digital elevation model or auxiliary thematic map was used not to alter the comparison to the other estimation methods by considering external information. Several options also exist regarding the form of spectral distance that is used to identify the optimal $k$ [25], [28]. In general, however, the use of more complex spectral metrics only brings marginal and occasional improvements over that of the simple Euclidean distance [30], [31], whereas consistent benefits can only be obtained by much more sophisticated and computer-intensive procedures [28]. For this reason, only the basic Euclidean distance was currently considered. As regards the possibility of obtaining a per-pixel indicator of the possible error made by (11), it can be computed as the variance of the $k$-NN estimator. More particularly, assuming that all sample weights are constant and $k$ is greater than 1, such variance $\sigma_{k-NN}^2$ can be computed as

$$\sigma_{k-NN}^2 = \frac{\sum (Y_i - Y_{k-NN})^2}{k-1}$$

(12)

This variance is actually the likely dispersion of the estimated values (error variance), in analogy with what is done by other estimation methods (see below). It is intuitive that low values of such variance indicate that the spectral data exactly define the estimated parameter, i.e., are informative on this (in a nonparametric way). The contrary is true when the estimated variance is high, which means that the selected $k$ points are the most spectrally similar to the sample pixel are dissimilar for the parameter’s values. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known.

This can be thought of the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data. The training examples are vectors in a multidimensional feature space. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class $c$ if it is the most frequent class label among the $k$ nearest training samples. Usually Euclidean distance is used as the distance metric, however this will only work with numerical values. In these cases, such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the actual classification phase, the test sample (whose class is not known) is represented as a vector in the feature space. Distances from the new vector to all stored vectors are computed and k closest samples are selected. There are a number of ways to classify the new vector to a particular class, one of the most used techniques is to predict the new vector to the most common class amongst the K nearest neighbors. A major drawback to using this technique to classify a new vector to a class is that the classes with the more frequent examples tend to dominate the prediction of the new vector, as they tend to come up in the K nearest neighbors when the neighbors are computed due to their large number. One of the ways to overcome this problem is to take into account the distance of each K nearest neighbors with the new vector that is to be classified and predict the class of the new vector based on these distances.

© 2010 ACADEMY PUBLISHER
The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when k = 1) is called the nearest neighbor algorithm. The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to evaluate features based on their distance from the training data.

### TABLE I. CLASSIFIED DATA OF EIGHT TYPE OF SCRIPTS

<table>
<thead>
<tr>
<th>K value</th>
<th>Tibetan</th>
<th>Tai Lue</th>
<th>Chinese</th>
<th>Naxi Pictographs</th>
<th>Uighur</th>
<th>Tai Le</th>
<th>Yi</th>
<th>English</th>
<th>average recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=15</td>
<td>92%</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
<td>90%</td>
<td>98%</td>
<td>92%</td>
<td>100%</td>
<td>96.25%</td>
</tr>
</tbody>
</table>

By testing with different values of k, we found that k=15 gave a classification accuracy better than any other value.

### A. Experimental Design

To prove the effectiveness of the proposed method in this paper, a series of experiments have been carried out in Chinese, English, and other six different kinds of frequently-used minority scripts such as Tibetan, Tai Lue, Naxi Pictographs, Uighur, Tai Le, Yi, 8 kinds of scripts all together. Because the scripts that need to identify are always changeable in the real world, we programmed an automatic generation procedure of samples by means of random function to generate every kinds of script images mentioned above, as shown in Fig. 4. Every kind of script images is generated 200, 100 images as samples dataset A, the others as dataset B. Finally dataset A and dataset B respectively contain 800 images that consist of 8 kinds of script.

### B. Experimental Results

Dataset A is for training the classifiers, dataset B is for testing. The experimental results show that the best performance was achieved by K=15. The performance results are showed in Table 1. The experimental results show that Chinese, English, and Tai Lue can achieve the best performance, at the rate of 100%, using the KNN as classifier. But among Tibetan, Naxi Pictographs, Uighur, Tai Le, Yi, there are some wrong classification, including 8 Tibetan script images were classified as Naxi Pictographs, 2 Naxi Pictographs script images were classified as Tibetan, 5 of Yi script images were classified as Chinese and 3 as Tibetan, 2 Tai Le script images were classified as Uighur, and 10 Uighur script images were classified as Tibetan. The above results show that the method presented in this paper has an excellent performance on recognizing the kinds of Chinese minority scripts, the average accuracy is up to 96%.

### VII. CONCLUSION

Recognizing the kinds of Chinese minority scripts is a very significant and difficult work. This paper studied this issue and presented a method utilizing wavelet decomposition on every kind of script images, extracting wavelet energy feature and the wavelet energy proportion feature from wavelet images, and then using the modified K-Nearest Neighbor (K-NN) as classifier to recognize the different types of script images. The experimental results show that our approach can effectively identify the types of Chinese minority scripts, with the average recognition rate at 96%.

### ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 60803096).
Figure 4 Eight type of script image

REFERENCES


Hai Guo received the M.Sc. degree in pattern recognition and intelligent system from the Kun Ming University of science and technology, China, in 2004. He has been teaching at Dalian Nationalities University since graduation. For many years, he has done research in image processing, pattern recognition. He has published more than 10 papers.

Jing-ying Zhao received the M.Sc. degree in computer application from the Chang Chun University of science and technology, China, in 2003. She has been teaching at Dalian Nationalities University since 2003. For many years, she has done research in image processing, pattern recognition. He has published more than 10 papers.