A Hierarchical Coarse Classification of Handwritten Chinese Characters by Distance Threshold

Fu Jiacai, Sun Guangling and Shi Guangfan

Abstract—A novel hierarchical coarse classification method for handwritten Chinese characters by distance threshold is proposed in this paper. Certain statistical features are detected and extracted first, and then a hierarchical distance between a tested sample and a template is calculated. In hierarchical distance, each level corresponds to each dimension. Each dimension distance is accumulated into pre-dimensions’ distance. After accumulated distance is computed, whether a character will enter next level is determined according to the accumulated distance and current level distance threshold of the template. The distance threshold of every character’s template is learned from samples of the character. The speed of the coarse classification is very fast, while the error rate is only 1.4% and the number of candidate characters has been reduced from 3,755 to only about 15 in average.

Index Terms -- Handwritten Chinese character; Hierarchical coarse classification; Statistical features; Distance threshold.

I. INTRODUCTION

Handwritten Chinese character recognition is a well-known difficult task in pattern recognition with one of the most important reasons is that the number of categories of Chinese character is considerably large. The strategy that deals with it is referred as coarse classification. A good coarse classification method should try to meet with three requirements: low error rate, greatly reduced number of candidates and fast speed.

Up to now, many coarse classification schemes have been proposed, which can be classified as two categories roughly according to the features used: structural feature based coarse classification and statistical feature based coarse classification. The examples of former are [1,2,3,4], and the examples of latter are [5,6,7]. The success of coarse classification using structural features heavily depends on the stroke’s extraction. It is regretted that it will suffer from problems of stroke ambiguity when strokes are connected or intersected. Once the stroke’s extraction fail, the coarse classification will fail. So, a low error rate cannot be guaranteed in this situation. On the contrary, the statistical features are extracted from the character’s image to avoid the stroke’s ambiguity. According to ways in which features are employed, statistical feature based coarse classification can be further classified into two categories: cluster-based and matching-based. In cluster-based method [5,6], the feature vectors of learning samples are clustered first based on a cluster criterion, and then a distance of feature vector of a test sample from every cluster center is calculated. The characters belonging to a cluster or some clusters that are nearest to the tested sample’s feature vector are chosen as candidates. The drawback of this method mainly rests on the non-reasonable assumption of Gaussian distribution of learning feature vectors of all the classes. Though the speed is sufficiently fast and the error rate is low enough, the average number of characters in a cluster is still very large, for example, 516 characters per cluster reported in [5]. Though the candidate number has been reduced greatly by [6], there are many parameters that have to be set empirically. In matching-based method [7], a template vector is generated from the learning samples of a character. When a sample is recognized, the
distance between feature vector of the sample and every template vector is calculated. The $n$ characters of which distances from the tested sample are listed top $n$ positions in ascending order are selected as candidates. The accuracy rate is very high while the number of candidates is greatly reduced. One problem of this method is that $n$ is set a fixed value that is difficult to set, and the other problem is the much slower speed compared with the cluster-based coarse classification.

In this paper, a novel hierarchical coarse classification scheme is presented. The coarse classification problem is decomposed into a serial of two-class classification problems. Specifically, based on a certain distance metric, all the distances between a tested sample and a template of every character are calculated first. Then every character will be determined weather it should be candidate according to the distance between the tested sample and its template. The determination rule is based on the character’s distance threshold. If the distance is less than the threshold, the character will be candidate, otherwise not. Every character’s distance threshold is obtained by learning samples of the character. In addition, to improve the speed of the coarse classification, a hierarchy scheme is adopted. Each level in the hierarchy corresponds to each dimension of feature vector.

The remainder of this paper is organized as follows. Section 2 describes our coarse classification method and statistical features used. In addition, feature extractions are conducted to improve classification accuracies and classification speed. Section 3 explains the hierarchical version of the proposed method. Experiment results are provided in section 4. Finally in section 5, conclusions and discussions are given out.

II. DISTANCE THRESHOLD BASED COARSE CLASSIFICATION

The coarse classification has great impact on the speed and accuracy of a character recognition system, especially for large set recognition system. So, it is very important for character recognition system. The position of coarse classification in a character recognition system has been illustrated in Fig.1. From Fig.1, we can see that the four modules, preprocessing, feature’s detection and extraction, coarse classification and fine classification compose a complete system. Because preprocessing, feature detection and extraction precede coarse classification, they will affect the performance of coarse classification directly. So, we will briefly discuss the two modules first.

A. Preprocessing

Preprocessing includes normalization and smooth. Our coarse classification method requires for a uniform size of character. An original character’s image has an arbitrary size, so it must be normalized. A nonlinear normalization algorithm proposed in [8] is applied and the normalization size is 64*64 pixels. After normalization, smooth based on Chamfer distance transform [9] and a 3*3 smooth window is conducted to patch holds and smooth contours of the character. Fig.2 shows an example of normalization and smooth of a character image.
B. Feature’s Detection and Extraction

Our coarse classification method uses statistical features. To compare different coarse classification results by different features, we introduce three generally used statistical features; they are PSF (Peripheral Shape Feature), SDF (Stroke Density Feature) and DEF (Directional Element Feature) respectively. The method of detecting these features has been explained in [6,7]. As far as how to extract desired features based on the detected features, it will be discussed in section 2.3.1

C. Distance Threshold based Coarse Classification

As mentioned above, coarse classification should try to satisfy three requirements: low error rate, greatly reduced number of candidates and fast speed. Though the three factors are conflicted mutually, a coarse classification algorithm should compromise among the three factors.

Let \( n \) stands for the number of all characters; \( m \) stands for the number of after coarse classification; \( C_i \) stands for the \( i \)th character; \( x \) stands for an input feature vector; \( R_i \) stands for the template feature vector of \( C_i \) and it is a mean of feature vectors of all learning samples of \( C_i \); \( T_h \) is a distance threshold of \( C_i \). Our coarse classification method is described in Fig. 3. For convenience, we call the algorithm DTHCC (Distance Threshold based Coarse Classification).

```
Input: x;  
Output: m candidate characters determined by coarse classification;  
for i = 1 to n  
if D(x, R_i) < T_h, C_i will be accepted by coarse classification, otherwise rejected.  
end
```

Fig. 3 DTHCC algorithm

The method itself is very simple, but it introduces a new idea that a coarse classification is decomposed into \( n \) two-class classification problems. In \( i \)th two-class classification problem, \( C_i \) is one class and all other characters except for \( C_i \) form the other class. Therefore, weather \( C_i \) will be received by coarse classification depends on the relationship between its distance threshold \( T_h \) and the distance \( D \) of \( x \) and \( R_i \). This is a commonly used scheme in two-class classification problem. The advantage of this method for coarse classification is that a very high accuracy rate and a much less number of candidates can be obtained simultaneously.

There are two key problems worthwhile thinking over: One is the distance between input character and a template and the other is the distance threshold of the template. Specifically, the distance between input character and the template of its class should be as small as possible but the distance between input character and the template of other class should as large as possible. As far as for distance threshold of a template, under keeping a very high true accuracy rate, it should try to accept false samples as less as possible. The solving method is further explained in two subsections below:

1) Feature Extraction

Feature extraction is an effective means to deal with the former problem. Intra-class and inter-class distance based feature extraction criteria are able to extract desired features that make distances of intro-class as small as possible and distances of inter-class as large as possible. Of these criteria, Devijer’s \( J_2 \) criterion [10] is most commonly applied for its simple computation. So in our paper, we use \( J_2 \) criteria to extract features from DEF, PSF and SDF respectively. \( J_2 \) can be obtained as in equation (1):

\[
J_2 = tr(S_w^{-1}S_b) = \sum_{i=1}^{d} \lambda_i \tag{1}
\]

\[
\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \quad W = [u_1, u_2, \cdots u_d]
\]

Where \( D \) is the dimension of original feature vector and \( d \) is the dimension of extracted feature vector, which \( D \geq d \). \( \lambda_i \) and \( u_i \) is the eigenvalue of matrix \( S_w^{-1}S_b \) and corresponding eigenvector respectively. As far as \( S_w \) and \( S_b \), they can be obtained according as the following equations [cf.8]:

\[
m_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i}^{(j)} \tag{2}
\]

\[
m = \sum_{i=1}^{n} P_i m_i \tag{3}
\]

\[
S_w = \sum_{i=1}^{n} P_i (\frac{1}{n_j - 1} \sum_{j=1}^{n_j} (x_{i}^{(j)} - m_i)(x_{i}^{(j)} - m_i)^T) \tag{4}
\]

\[
S_b = \sum_{i=1}^{n} P_i (m_i - m)(m_i - m)^T \tag{5}
\]

Where \( x_{i}^{(j)} \) is a feature vector of \( j \)th sample of \( i \)th character; \( P_i \) is a probability of \( i \)th character and it can be estimated by a frequency of \( i \)th character.

If not declared specially, either \( x \) or \( R \) that appears in algorithms always refers to an extracted feature vector.

2) Distance Threshold Learning Algorithm

Obviously, distance threshold is a crucial parameter for the performance of coarse classification. In order to fully use limited samples, we propose leave-one-out algorithm to learn the distance threshold. The procedure is explained in Fig. 4.
According to leave-one-out algorithm, the threshold-learning algorithm includes each loop, a different sample is excluded as a test sample and remaining samples as learning samples. A maximal distance is speed problem. For in our problem, obtained based on the learning samples. After comparatively large value; Assume that get feature vector. Then even if based on the simplest distance distance threshold is a value that the mean plus the deviation metric—City distance, there will be $(2^*2$ multiplied by $2$ of the distribution. At the same time, we get a subtraction operations. It is a rather heavily calculation burden leading to a slow coarse classification. To speed up coarse classification, a hierarchical strategy is employed. The hierarchical version of DTHCC is described in Fig.5. For convenience, we call it HIE-DTHCC.

**Input:** $X = \{x_1, x_2, \ldots, x_n\}$ 

$X$ is a feature vector set that is extracted from samples of a character; $n$ refers to the sample’s number of the character. 

**Output:** distance threshold of a character. 

For $i = 1$ to $n$ 

$$R_i = \frac{1}{n-1} \sum_{k=1, k\neq i}^{n} x_k$$ 

$$d_i = \max D(x_i, R_i)$$ 

end 

$$m_d = \frac{1}{n} \sum_{i=1}^{n} d_i, \sigma_d^+ = \sqrt{\frac{1}{N^*} \sum_{d_i > m_d} (d_i - m_d)^2}$$ 

$N^*$ is the number of $d_i$ larger than $m_d$ 

$$Th = m_d + 2^* \sigma_d^+$$

Fig.4 Distance threshold learning algorithm

According to leave-one-out algorithm, the threshold-learning algorithm includes $n$ loops which $n$ is the number of samples. In each loop, a different sample is excluded as a test sample and remaining samples as learning samples. A maximal distance is obtained based on the learning samples. After $n$ loops, we will get $n$ maximal distances that form a distribution. Finally, the distance threshold is a value that the mean plus the deviation multiplied by 2 of the distribution. At the same time, we get a template feature vector $R$: 

$$R = \frac{1}{n} \sum_{i=1}^{n} R_i$$ \hspace{1cm} (6)$$

III. HIERARCHICAL COARSE CLASSIFICATION

The most disadvantage of our coarse classification is its speed problem. For in our problem, $n$ is 3,755 that is a comparatively large value; Assume that $k$ is the dimension of feature vector. Then even if based on the simplest distance metric—City distance, there will be $(2^*k-1)*3,755$ addition or subtraction operations. It is a rather heavily calculation burden leading to a slow coarse classification. To speed up coarse classification, a hierarchical strategy is employed. The hierarchical version of DTHCC is described in Fig.5. For convenience, we call it HIE-DTHCC.

**Input:** $x$; 

**Output:** $m$ candidates determined by coarse classification; 

$$n_0=n, S_0=\{R_1, R_2, \ldots, R_n\}$$ 

for $j = 1$ to $k$ 

$$n_j=0, S_j=\{\Phi\}$$ 

for $i = 1$ to $n_{j-1}$, 

$$D_{i,j}(x, R_i) = D_{i,j-1}(x, R_i) + d_{i,j}(x, R_i), R_i \in S_{j-1}$$ 

if $D_{i,j}(x, R_i) < Th_{i,j}$, $S_j = S_j \cup R_i$, $n_j = n_j + 1$; 

end 

end

Fig.5 HIE-DTHCC algorithm.

Where $n_i$ and $S_i$ refer to the candidate number and the set of corresponding template feature vectors of candidates selected after $j$ dimensions distances are calculated. They are initialized the number and template feature vectors of all characters respectively; $k$ refers to the feature’s dimension; $d_{i,j}$ refers to the distance between $j$th dimension of $x$ and $j$th dimension of $R_i$; $D_{i,j}$ refers to accumulated distance up to $j$th dimension between $x$ and $R_i$; $Th_{i,j}$ refers to accumulated distance threshold up to $j$th dimension of $C_i$.

In fact, each dimension of feature vector is capable of discarding a lot of candidates. Therefore the essential of HIE-DTHCC is its fully using each dimension’s discriminative ability, so as to make the coarse classification efficiency improved notably.

The high efficient HIE-DTHCC algorithm is at the cost of decreasing accuracy rate, since any error in $k$ levels of coarse classification will result in failure of coarse classification. So the number of levels in HIE-DTHCC algorithm should as little as possible. This will not decrease the efficiency of coarse classification, in that the candidate number will be reduced distinctly with the level increases.

IV. EXPERIMENTS AND RESULTS

A. Experiment Database

Our experiment is conducted based on the database collected by Harbin Institute of Technology and Hong Kong Polytechnic University. The database comprises a collection of 751,000 loosely constrained handwritten Chinese characters, including 3,755 categories written by 200 different writers. That means the sample’s number of every character is 200. Since we adopt leave-one-out algorithm, the number of tested sample of every character is 200.

B. Empirical Results

1) The Different Impacts on Coarse Classification Using Different Features

As discussed in section 2.2, to test different results of coarse classification because of different features, three statistical
features: PSF, SDF and DEF are detected. As the original dimension of PSF and SDF are 32, the three features are all extracted to 32 dimensions. Note that the dimension of DEF has been reduced from 196 to 32, which means much discriminative information has been lost. Table 1 lists the statistics about these features.

Table 1 Statistics of PSF, SDF and DEF

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>ACCURACY RATE</th>
<th>NO. OF CANDIDATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSF</td>
<td>99.5%</td>
<td>95</td>
</tr>
<tr>
<td>SDF</td>
<td>99.7%</td>
<td>110</td>
</tr>
<tr>
<td>DEF</td>
<td>99.2%</td>
<td>45</td>
</tr>
</tbody>
</table>

The statistics of table 1 show that PSF and SDF get a little higher accuracy rate than DEF. However, the number of candidate of PSF and SDF are much more than that of DEF even if the dimension of DEF has been compressed seriously. The fact indicates that DEF has a stronger distinguished character than PSF and SDF. So in following experiments, we use DEF to obtain other results.

2) The Different Impacts on Coarse Classification Using Different Dimensions

To show the impact of same feature but having different dimensions on coarse classification, different dimensional features of DEF have been extracted and tested. The results are reported in table 2.

Table 2 Accuracy rate and No.of Candidate of different dimensions

<table>
<thead>
<tr>
<th>Accuracy Rate</th>
<th>No.of Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
</tr>
<tr>
<td>16</td>
<td>56</td>
</tr>
<tr>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td>48</td>
<td>40</td>
</tr>
</tbody>
</table>

The accuracy rate goes up and candidate number drops down with the dimension increases. The law recovered by data in table 2 is coincident with the feature extraction criterion $J_2$, for with the dimension increase, $J_2$ also increases, which means the distinctness among characters will be more intensive while the samples of a character will be more concentrated simultaneously. However, since $J_2$ increase slower and slower, the number of candidate decreases slower and slower.

3) Performance of HIE-DTHCC

To speed up the coarse classification, hierarchy scheme must be adopted. The data of table 3 illustrated the performance of the hierarchical coarse classification.

Table 3 Performance of HIE-DTHCC

<table>
<thead>
<tr>
<th>Accuracy Rate</th>
<th>No.of Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>99.7</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
</tr>
<tr>
<td>12</td>
<td>98.6</td>
</tr>
</tbody>
</table>

From table 3, we can see that the candidate’s number has decreased steeply, but the accuracy rate also decrease at the same time. Therefore, we choose 12 as the dimension of extracted features.

V. CONCLUSIONS AND DISCUSSIONS

Chinese character recognition is a typical large set recognition problem in which there are 3,755 characters according as GB2312-80. Therefore, the success of Chinese character recognition system is heavily depends on the coarse classification. In this paper, a new coarse classification scheme is proposed which has mainly three advantages compared with existing methods: (1) The statistical feature is easier to be obtained than structural features; (2) The much less candidate number than [5,6] while keeping a sufficiently low error rate; (3) The parameters of our method, the distance threshold of every character, can be controlled better and more reasonable than the general and global parameter $n$ in [7].

Our method can also be easily extended to other large-scale recognition problem, such as Japanese Kanji character recognition.

Future work includes to study how to combine multi-features and how to select and extract ones from the features to improve the performance of our coarse classification further.

REFERENCES