Finger-Written Chinese Characters Recognition using Hierarchical LDA

Duanduan Yang[1,2], Lianwen Jin[1], Li-Xin Zhen[2], Jiang-Cheng Huang[2]
[1]*School of Electronics and InformationSouth China University of technology,*
Guangzhou, 510640, P.R.China
[2]*Motorola China Research Center, Shanghai, 210000, P.R.China*
ddyang@scut.edu.cn , eelwjin@scut.edu.cn

Abstract

In our camera based user interface, user inputs Chinese characters by moving fingertip with his/her fingers. Therefore, we call those characters "finger–written Chinese characters" (FWCCs). To recognize FWCCS with high recognition accuracy, we propose a new method to discriminate similar FWCCs based on a two stages Linear Discriminant Analysis (LDA) architecture. Experimental results show the proposed method is useful for the improvements of recognition rate and outperforms our previous hybrid recognition method. Comparing with MQDF-based method, the proposed method can reach similar classification performance but with much less storage requirement.

1. Introduction

Since the first paper about Chinese character recognition was published in 1966 [1], great advances have been achieved in the filed. The current systems are already able to recognize regular handwritten characters with very high accuracy. But finger–written Chinese characters without ink information are different from the regular handwritten characters. In our camera-based system [2] (Figure 1), FWCCs are reconstructed using the fingertip location frame by frame. Since it is difficult for the system to distinguish the state of up and down of fingertip when writing with finger, all strokes of a FWCC are connected. So they are one-stroke style of characters [2] as figure 2 shows. We previously used a hybrid method [3] to recognize FWCCs. The recognition rate can reach 90.8%.

There are about 10% samples could not recognized correctly. The reasons are maybe multifold. By analyzing the recognition results, we found one of them is that the classifier is not good enough to discriminate the similar characters of different categories. The. discrimination is particularly important for FWCC recognition. For example, it is not difficult to distinguish the character "海"and the character"诲" when both of them are written as a regular style figure 3(a). However, when they are written as the style of FWCCs, their appearances are very similar as the figure 3 (b) shows. Therefore, it is significant for FWCC recognition to design a good classifier, which is able to distinguish the similar characters to certain extent.

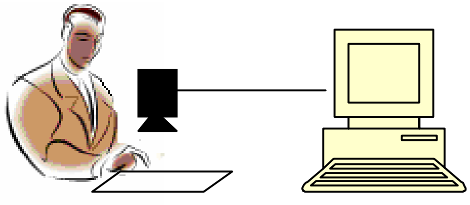


Figure. 1 Our camera-based system.

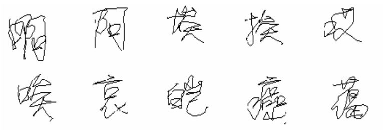
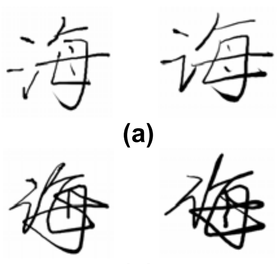


Figure 2. Some finger-written Chinese characters



(a)

(b)

Figure 3. Two writing style of the character "海"and the character"诲". (a)Handwritten

Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)

characters (regular style); (b) Finger-written characters (one-stroke style).

LDA can maximize the between-class measure while minimizing the within-class measure, it is useful for the different categories discrimination. If we apply LDA on all character categories, it contributes less to discriminating similar categories. Therefore, in this paper, we generate a series of *similar category sets* using the recognition results on training samples and apply LDA on each set for similar characters discrimination. By the method, some similar characters of different categories can be distinguished correctly and the recognition rate can be improved to about 93%. To prove its effectiveness, we compare the proposed method with MQDF [7], which is one of the best discriminant methods for handwriting recognition. Experiments show the two methods can reach the similar classification performance but our method requires much less storage space.

2. FWCCs reconstruction

In our camera based user interface [2], user inputs Chinese characters by moving fingertip. The Chinese input system can figure out the location of user’s fingertip when it is moving. Therefore, FWCCs can be reconstructed using the trajectories of the user’s fingertip. The reconstruction of a typical FWCC “中” is shown in Fig. 1. More details about FWCCs reconstruction can reference the [2].

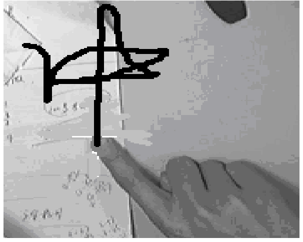


Figure 4.A typical FWCC “中”.

3. Feature extraction

Feature extraction is of particular importance for handwriting recognition since the performance of recognizer depends largely on it. In this paper, we use 8-directional features to represent the finger-written Chinese character pattern. Directional features have been widely used for Chinese character recognition with great success. The 8-directional features have been demonstrated to be very effectiveness for online Chinese character recognition in [4]. Details of the 8-directional feature extraction can be found in [4].

4. Classifier design

The main idea of the classifier design is to generate a series of *similar category sets* (SCSs) and then to apply LDA on each set. We expect the method can discriminate the similar characters effectively. Suppose the n SCSs , $S_1,...S_n$, had been constructed, the new pattern x can be recognized by two steps. Step 1, the *similar category set* S_i which is expected to contain the category of x is determined. Step 2, the pattern x is recognized from S_i using a LDA Euclidean distance classifier.

The generation of the SCSs is the key of the classifier design. If we use $\Pi = \{c_1,...c_m\}$ to denote the universe set of Chinese character categories, where $c_1,...c_m$ are the elements of Π and denote the m categories for Chinese characters, any *similar category set* is a subset of Π. To generate the SCSs, $S_1,...S_n$, we firstly design a standard LDA Euclidean distance classifier f_1 using the mean of training samples of each category contained in Π and LDA algorithm. The classifier is used to recognize all training samples. If the recognition results of training samples from different categories are same, those categories are considered similar. Concretely, the elements of the SCSs, $S_1,...S_n$, are defined by the following steps.

Step 1, we initially construct m *similar category sets*, $S_1,...S_m$. The elements of S_i ($i = 1,...m$) are determined by the following rule. If there exist a training sample labeled with the category c_j which is recognized as the category c_i by the classifier f_1 , the category c_j is one of the elements of S_i . In another word,

$S_i = \{c_j$ | there are training samples of c_j is recognized as c_i by f_1 }.

To determine the *similar category set* for a new sample, we need a index set L_i ($i = 1,...m$) to index S_i . At step 1, the index set L_i has only one element, $L_i = \{c_i\}$, $i = 1,...m$.

Step 2, traversing all SCSs, the *similar category set* S_k with the smallest size is determined.(S_k contains the least number of elements of all SCSs.)

Step 3, Considering the SCSs whose size is less than *T* (*T* is a constant), we can find the SCS S_u, which contains the most number of elements that also belong to S_k .(The cardinality of the intersection of S_k and S_u is the largest.)

Step 4, S_k and S_u are replaced by their union and the new union set is indexed with $L_u \cup L_k$. By this step, the number of the *similar category sets* becomes $m-1$.

Step5, Repeat step2~step4 until the number of the *similar category sets* is equal to the predefined value, n.

As we can see, the categories are considered similar and are put in a *similar category set*, if the training samples labeled with them are recognized as the same category by f_1.

After the elements of each *similar category set* are defined (Fig.5), we can train a LDA Classifier for each similar category set. Corresponding to S_i, the LDA Classifier f_{2i}, is trained using the samples labeled with the elements contained in S_i. Comparing with the standard LDA Euclidean distance classifier f_1, f_{2i} is generated by the same algorithm but with different training samples. To distinguish with f_1, we call $f_{1i} \cdots f_{2n}$ similar category classifiers.

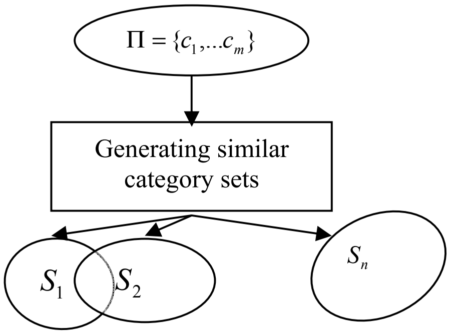


Figure 5. Similar category sets construction.

The final classifier consists of a standard LDA classifier f_1 and a series of similar category classifiers $f_{1i} \cdots f_{2i} \cdots f_{2n}$ (see figure 6). When a new sample, x, is input, it is recognized by f_1 firstly and the recognition result is denoted with $r1$. In the index sets $L_1, \dots L_n$, there is exactly one index set L_i satisfied $r1 \in L_i$. The *similar category set* S_i indexed with L_i is considered to contain the recognition result of x. Therefore, the classifier f_{2i} corresponding S_i is chosen to recognize x finally.

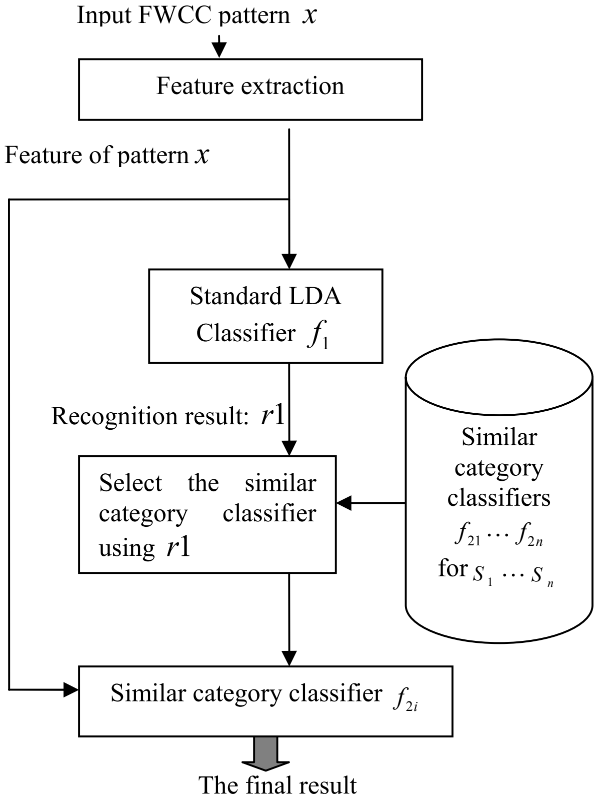

Input FWCC pattern x
Feature extraction
Feature of pattern x
Standard LDA Classifier f_1
Recognition result: $r1$
Select the similar category classifier using $r1$
Similar category classifiers $f_{21} \cdots f_{2n}$ for $S_1 \cdots S_n$
Similar category classifier f_{2i}
The final result

Figure 6.the overview of the final classifier

By using the recognition results on training samples to generate the similar character category sets, a set of LDA corresponding classifiers are designed. By applying LDA algorithm one each similar character category set, the final classifier is expected to discriminate the similar characters more effectively.

5. Experiments

Some experiments are performed to evaluate the proposed method in this section. 300 sets of handwritten Chinese characters (each set consists of 3755 categories of GB2312-80 level 1 Chinese characters) were used to simulate FWCCs with strokes fully connected for our experiments, we call those characters simulative finger-written Chinese characters (SFWCCs). There characters were written by 300 different individuals. And all the characters are written

Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)
0-7695-2822-8/07 $25.00 © 2007 IEEE

naturally with no constraint in stroke order, stroke number and writing style. Fig. 7 shows some samples.

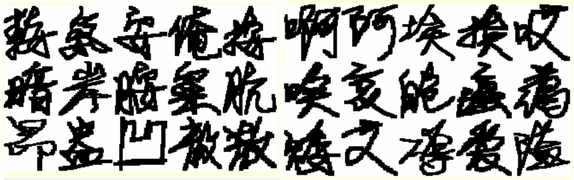


Figure 7. Some simulative finger-written Chinese characters.

200 sets of SFWCCs are used to train the classifier. The remained 100 sets and 11265 FWCCs cover 3755 categories collected by our camera-based system are used for testing.

Table 1 shows the recognition rates of SFWCCs with different amounts (n in the section 4) of the SCSs using 8-directional features and the proposed methods.

Table .1 The recognition rates of SFWCCs with different n using eight directional features and the proposed methods.

The number of SCSs(n)	100	400	700	1000
Recognition rate(%)	95.04	95.52	95.59	95.69
Storage space(M)	24.8	34.5	37.1	40.2

From table 1, we can see that both the recognition rate and the storage space of the proposed method increase with the increase of n (the number of *similar category sets*). In some sense, this means LDA contributes more to discriminating similar characters when it is applied in more SCSs.

Table 2 compares the performances of our method (with 400 SCSs) and Modified Quadratic Discriminant Function (with 30 eignvectors for each category). MQDF proposed by Kimura *et al.* [5] aims to improve the computation efficiency and classification performance of QDF via eigenvalue smoothing, which have been used successfully in handwriting recognition, which can reach high recognition rate[6] with large storage space.

In this paper, the eight-directional feature of a FWCC is a 512-dimensional vector. To reduce the storage space of MQDF classifier, we use standard LDA algorithm to reduce the dimensions of the vectors (to 256). Therefore, the parameters of MQDF classifier consist of two parts. The first part is LDA transformation matrix and the transformed means for each category, whose size is 256*3755+512*256. The second part is the principle eigenvectors and eigenvalues, whose size is 256*3755*30 +31*3755. If every datum occupies 4 bytes space, the total storage space of MQDF is 114.66M.

Table 2. Performance comparison of MQDF and our method using eight directional features (*n =400*)

	Recognition rate		Storage space
	SFWCCs	FWCCs	
Previous method[3]	---	90.8%	--
MQDF	96.05%	93.37%	114.66M
Our methods	95.52%	93.06%	34.52M

From the table 2, we can see that the proposed method is almost as good as MQDF in classification ability and is better than MQDF in storage space. Our method almost reaches the same recognition rate as MQDF with 34.52M storage space.. Under ideal conditions, there should be no difference between the recognition rates of FWCCs and SFWCCs. However, it is hard for a camera based system to collect Chinese characters with high quality as the touch screen of a PDA. Therefore, the recognition rate of FWCCs is lower than that of SFWCCs as table 2 shows.

Figure 8 shows some similar FWCCs labeled with different categories which can be distinguished by the proposed method.

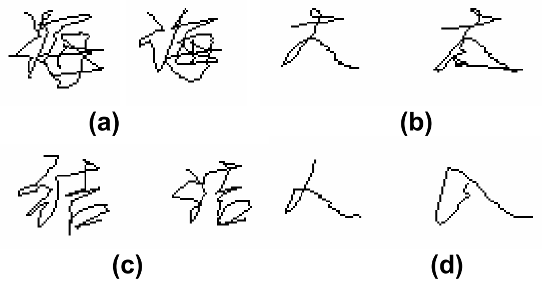


Fig. 8. Some similar FWCCs distinguished by our method. (a) “海”and “诲” ; (b) “大” and “太”; (c) “结”and “洁” ; (d) “人”and “入”.

6. Conclusion

This paper researches FWCCs recognition. We use eight-directional features to represent the patterns and propose a new method to discriminate similar characters. Using the method, some similar characters of different categories can be distinguished correctly and the recognition rate can be improved from 90.8% to 93.3%. Comparing with MQDF, it can almost reach the same classification performance with less storage space. In the future, we are going to research the new algorithm to discriminate similar characters based on the *similar category sets* obtained by this paper.

7. Acknowledgments

This paper is sponsored by the following research foundations: New Century Excellent Talent Program of MOE of China (No.NCET-05-0736) and Motorola Human Interface Lab Research Foundation (No.D84110).

8. References

[1]R.Casey and G.Nagy, ”Recognition of printed Chinese characters ”, *IEEE transaction on Computer*, vol.1, no.15, pp.91-101,1966.

[2] Previous work, In of the 18th International Conference on Pattern Recognition, HongGong, 2006.

[3] Previous work, In of Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition , Seoul, Korea, August.2005.

[4] Zhen-Long BAI and Qiang HUO,” A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition”, In of Proceedings of the Eight International Conference on Document Analysis and Recognition, Seoul, Korea, August. Vol. 1, pp.232-236, 2005.

[5] F.Kimura, K.Takashina, S.Tsuruoka, and Y.Miyake, “Modified quadratic discriminant functions and the application to Chinese character recognition ” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.9, No.1,pp.149-153,1987.

[6] Hailong Liu, xiaoqing ding. “Handwritten Character Recognition Using Gradient Feature and Quadratic Classifier with Multiple Discrimination Schemes”, In of Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, Seoul, Korea, August.Vol.1 19-23, 2005.

[7] F.Kimura, K.Takashina, S.Tsuruoka, and Y.Miyake, “Modified quadratic discriminant functions and the application to Chinese character recognition,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, No.1, pp. 149-153, 1987.