Offline Recognition of Chinese Handwriting by Multifeature and Multilevel Classification

Yuan Y. Tang, *Senior Member, IEEE Computer Society*,
Lo-Ting Tu, Jiming Liu, *Senior Member, IEEE Computer Society*,
Seong-Whan Lee, *Senior Member, IEEE Computer Society*,
Win-Win Lin, and Ing-Shyh Shyu

Abstract—One of the most challenging topics is the recognition of Chinese handwriting, especially offline recognition. In this paper, an offline recognition system based on multifeature and multilevel classification is presented for handwritten Chinese characters. Ten classes of multifeatures, such as peripheral shape features, stroke density features, and stroke direction features, are used in this system. The multilevel classification scheme consists of a group classifier and a five-level character classifier, where two new technologies, overlap clustering and Gaussian distribution selector, are developed. Experiments have been conducted to recognize 5,401 daily-used Chinese characters. The recognition rate is about 90 percent for a unique candidate, and 98 percent for multichoice with 10 candidates.

Index Terms—Offline Chinese handwriting recognition, multifeature, multilevel classification, overlap clustering, Gaussian distribution selector.

(a) (b)

Fig. 1. (a) An example of Chinese character with complex structure. (b) Examples of similar Chinese characters.

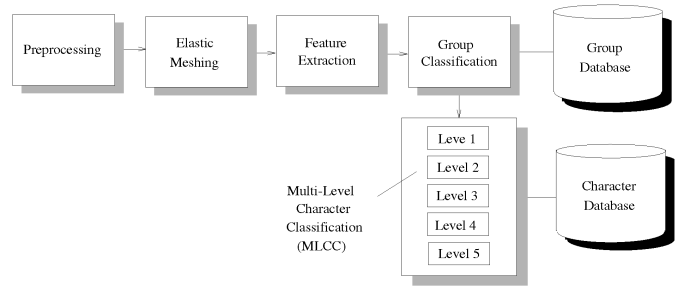

Fig. 2. Block diagram of the Chinese handwriting recognition system.

1 Introduction

Chinese handwriting recognition is one of the most challenging topics in optical character recognition. The difficulty is attributed to three factors:

1) The character set is very large: 3,000-7,000 characters are often used, 7,000-10,000 characters are collected in small dictionaries, and 70,000 Chinese characters are collected in large contemporary dictionaries.
2) In general, the structure of a Chinese character is much more complex than that of an alphabetic letter. An example can be found in Fig. 1a, which means "snuffle with a cold" or "speaking through the nose" in English. This Chinese character consists of 36 strokes.
3) Many Chinese characters have similar shapes, some examples can be found in Fig. 1b.

A survey of optical recognition of handwritten Chinese characters has recently been published [1]. Handwriting recognition is divided into online and offline categories. The online recognition has the advantage of capturing the temporal or dynamic information of the handwriting. This information consists of the number of strokes, the order of the strokes, the direction of the writing for each stroke, and the speed of the writing within each stroke. An extensive review of online character recognition can be found in [4]. As contrasted with online recognition, the offline case is much more difficult because it lacks the above knowledge. In this paper, we focus on a study of the offline recognition. An offline recognition system for handwritten Chinese characters is presented.

A block diagram of this Chinese handwriting recognition system is illustrated in Fig. 2.

The first two blocks in Fig. 2 will not be discussed in this paper, since the preprocessing operations such as segmentation, digitization, binarization, etc. are standard. The elastic meshing has been used to recognize Chinese characters printed in different fonts in our earlier work [3] and applied to nonlinear shape normalization [2]. In this paper, the rest of the components of the system, multifeature extraction and multilevel classification, will be presented in great detail.

The selection of stable and representative sets of features is at the heart of a pattern recognition system. For such complex and huge numbers of Chinese characters, a recognition system will fail if only one or two types of features are used. In our study, 10 different kinds of features are employed in the system. These features are called multifeatures, which will be presented in Section 2. The multilevel classification scheme consists of a group classifier and a five-level character classifier. In order to distribute the load of the character classification, a group classifier is designed. It breaks down all the characters into a smaller number of groups, hence the number of candidates for the process in the next step drops sharply. A new scheme called *overlap clustering* is used in the proposed group classification. In Section 3, first, the overlap clustering will be introduced, then a measure of similarity will be presented followed by an algorithm for the group classification. Section 4 deals with the multilevel character classification, which is composed of five levels. In the first level, a Gaussian distribution selector is used to select a smaller number of candidates from several groups. From the second level to the fifth one, matching approachs using different features are employed, respectively. Section 5 reports the experiments of the proposed system, which can recognize 5,401 daily-used Chinese characters with high recognition rates. The last section gives some conclusions.

- *Y.Y. Tang is with the Department of Computing Studies, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, and the Centre for Pattern Recognition and Machine Intelligence, GM-606, Concordia University, Montreal, Quebec H3G 1M8, Canada. E-mail: yytang@comp.hkbu.edu.hk.*
- *L.-T. Tu, W.-W. Lin, and I.-S. Shyu are with the Application Software Department, Computer and Communication Research Lab, Industrial Technology Research Institute, Chutung, Hsinchu, Taiwan 310, Republic of China.*
- *J. Liu is with the Department of Computing Studies, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.*
- *S.-W. Lee is with the Department of Computer Science, Korea University, Seongbuk-ku, Seoul 136-701, Korea.*

Manuscript received 29 Apr. 1996; revised 13 Mar. 1998. Recommended for acceptance by M. Mohiuddin.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 106571.

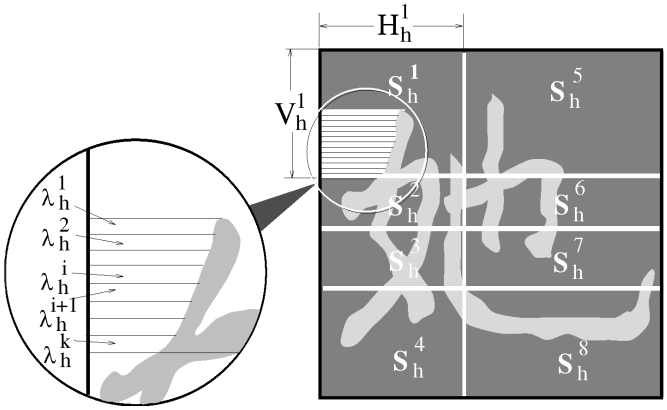

Fig. 3. Feature extraction using the HPBA and HPLD.

Fig. 4. Feature extraction using the HRSD and HRSDD.

2 MULTIFEATURES

Three types of features are used in this study, namely:

1) peripheral shape features,
2) stroke density features, and
3) stroke direction features.

The character image frame is first meshed by a 2×4 subframe, as shown in Fig. 3. A character is separated into left and right parts, each of which consists of four horizontal strips. Hence, there are a total of eight strips labeled by $S_h^1, S_h^2, \ldots, S_h^8$ in the character image frame. Strip S_h^i consists of several lines. Suppose k lines of them, namely, $\{\lambda_h^i | i = 1, 2, \ldots, k\}$, may touch the strokes of the character. Let $|\lambda_h^i|$ be the distance between the outermost stroke edge and the character image frame along line λ_h^i, and $H_h^i \times V_h^i$ be the subarea of the ith horizontal strip S_h^i. The horizontal peripheral background area (HPBA) in the ith horizontal strip S_h^i can be represented by the following formula:

$$\aleph_h^i = \left\{ \sum_{i=1}^{k} |\lambda_h^i| \right\} / \left(H_h^i \times V_h^i \right). \tag{1}$$

Similarly, horizontal peripheral line difference (HPLD) Ψ_h^i, vertical peripheral background area (VPBA) $\aleph_v^i$, and vertical peripheral line difference (VPLD) Ψ_v^i can be defined below:

$$\begin{aligned}
\aleph_v^i &= \left\{ \sum_{i=1}^{k} |\lambda_v^i| \right\} / \left(H_v^i \times V_v^i \right) \\
\Psi_v^i &= \left\{ \sum_{i=1}^{k} \left(\left\| \lambda_v^i | - | \lambda_v^{i+1} \right\| \right) \right\} / \left(H_v^i \times V_v^i \right) \\
\Psi_h^i &= \left\{ \sum_{i=1}^{k} \left(\left\| \lambda_h^i | - | \lambda_h^{i+1} \right\| \right) \right\} / \left(H_h^i \times V_h^i \right)
\end{aligned} \tag{2}$$

Stroke density is another very significant characteristic for representing the structure of characters. In this system, 16 features related to regional stroke densities are used. These features can be extracted in two different ways, depending on the directions of the elastic meshing used (see Fig. 4). Suppose there are m horizontal inspection lines in horizontal strip S_i, the crossing count on inspection line r_j is

$$\tfrac{1}{2} \Sigma_x \overline{f(x,y)} f(x+1,y).$$

The horizontal regional stroke density (HRSD) is the quotient of the accumulation of the crossing counts in S_i divided by the number of horizontal inspection lines in horizontal strip S_i. It is denoted by $\Re_h^i$ and can be computed by the following formula:

$$\Re_h^i = \left\{ \frac{1}{2} \sum_{j=1}^{m} \sum_{x} \overline{f(x,y)} f(x+1,y) \right\} / m. \tag{3}$$

Similarly, horizontal regional stroke density distribution (HRSDD) Φ_h^i, vertical regional stroke density (VRSD) $\Re_v^i$, and vertical regional stroke density distribution (VRSDD) Φ_v^i can be calculated by:

$$\begin{aligned}
\Phi_h^i &= \Re_h^i / \sum_S \Re_h^i, \\
\Re_h^i &= \left\{ \frac{1}{2} \sum_{j=1}^{m} \sum_{y} \overline{f(x,y)} f(x,y+1) \right\} / m, \quad \Phi_v^i = \Re_v^i / \sum_S \Re_v^i
\end{aligned} \tag{4}$$

Direction feature is another kind of important feature widely used in character recognition. In our Chinese handwriting recognition system, two sets of direction features are selected. They are:

- four-orientation stroke direction (4-OSD) Λ_k^j, $k = 0, 1, 2, 3$;
- eight-orientation stroke direction (8-OSD), Λ_k^j, $k = 0, 1, 2, \ldots, 7$.

3 GROUP CLASSIFICATION

In order to distribute the load of the character classification, a group classifier is designed. It breaks down all the characters into a smaller number of groups, hence the number of candidates for the process in the next step drops sharply. A new scheme called *overlap clustering* is used in the proposed group classification. In this section, first the overlap clustering will be introduced, then a measure of similarity will be presented, followed by an algorithm for the group classification.

3.1 Overlap Clustering

Overlap clustering is used in the group classification. Given m patterns, $x_1, x_2, \ldots, x_m$, contained in the pattern space **X**, the process of overlap clustering can be regarded as a way of seeking the regions $S_1, S_2, \ldots, S_k$ such that every pattern x_i, $i = 1, 2, \ldots, m$, falls into some of these regions, instead of only one region in ordinary clustering. The overlap clustering can be represented as follows:

Pattern space **X** is described by m-dimensional feature vector space, which is composed of n-dimensional feature vectors:

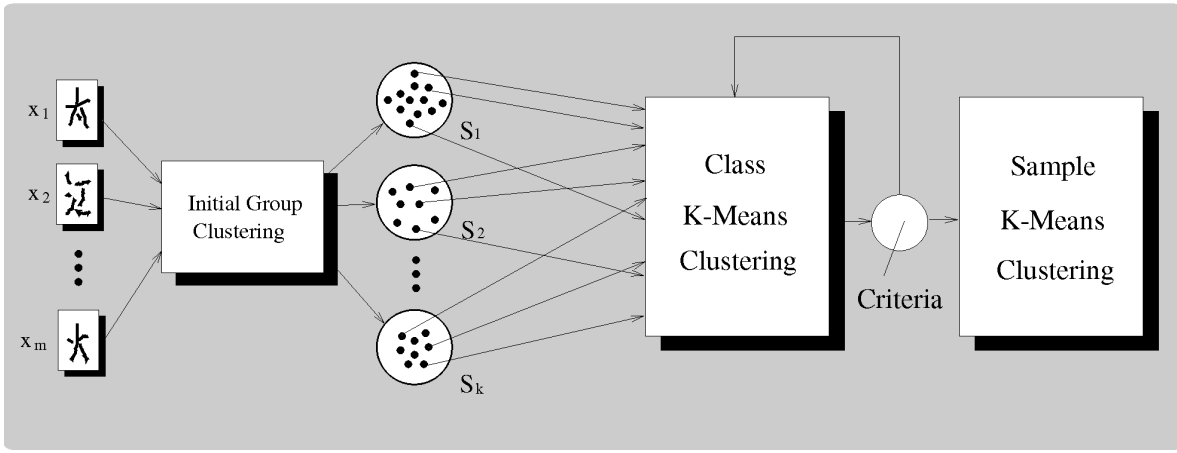

Fig. 5. Block diagram of the group classification.

$$\mathbf{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_m^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}. \tag{5}$$

DEFINITION 1. *Suppose patterns* $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{in}), i = 1, 2, \ldots, m$, *are clustered into k groups,* $S_1, S_2, \ldots, S_q, \ldots, S_k$. *It is called overlap clustering if the following conditions are satisfied*:

$$\begin{aligned} &S_1 \cup S_2 \cup \ldots \cup S_k = \mathbf{X} \\ &\exists_i \left(x_i \in S_q | q = 1, 2, \ldots, k \right) \\ &\exists_{i \neq j} \left(S_i \cap S_j \neq 0 \right) \end{aligned} \tag{6}$$

Please note that the inequality of $S_i \cap S_j \neq 0$ implies overlaps among different groups. This completely differs from the regular clustering.

3.2 Similarity Measure

To define a pattern cluster, it is necessary to first define a measure of similarity that will establish a rule for assigning patterns to the domain of a particular cluster.

Let $x_i^{(j)} = \left(x_{i1}^{(j)}, x_{i2}^{(j)}, \ldots, x_{ir}^{(j)}, \ldots, x_{in}^{(j)} \right)^T$ be the ith pattern in the group S_j, and N_j be the number of patterns in this group, i.e., $i = 1, 2, \ldots, N_j$. $x_{ir}^{(j)}$ denotes the rth feature in $x_i^{(j)}$. The expected value of $x_{ir}^{(j)}$ in group S_j is defined by $M_r^{(j)} = \frac{1}{N_j} \sum_{i=1}^{N} x_{ir}^{(j)}$, which is the center of group S_j for the rth feature. All expected values $M_r^{(j)}$ in group S_j constitute an expectation vector: $M^{(j)} = \left(M_1^{(j)}, M_2^{(j)}, \ldots, M_r^{(j)}, M_n^{(j)} \right)^T$, where n stands for the number of dimensions for each pattern vector, $M^{(j)}$ is the center of group S_j for all features.

The expected value of centers for the rth feature in the pattern space $\mathbf{X}$ is formulated by $M_r^G = \frac{1}{k} \sum_{j=1}^{k} M_r^{(j)}$. The variance of centers for the rth feature can be calculated by $\sigma_r^G = \frac{1}{k} \sum_{j=1}^{k} \left(M_r^{(j)} - M_r^G \right)^2$, where k denotes the number of groups.

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir}, \ldots, x_{in})^T$ be the ith pattern in the pattern space $\mathbf{X}$, and x_{ir} denotes the rth feature in x_i. The expected value of x_{ir} in the entire character space is defined by $M_r = \frac{1}{m} \sum_{i=1}^{m} x_{ir}$. The variance of x_{ir} can be calculated by $\sigma_r = \frac{1}{m} \sum_{i=1}^{m} \left(x_{ir} - M_r \right)^2$, where m denotes the number of characters in the pattern space. The measure of similarity used in this group cluster is a modified Euclidean distance called *normalized Euclidean distance (NED)*, which is described below.

DEFINITION 2. *Let* $M^{(j)}$ *be the expectation vector in group* S_j *and* $M_r^{(j)}$ *be the expected value of the* rth *feature in it.* σ_r *denotes the variance of* x_{ir}, *and* σ_r^G *the variance of centers for the* rth *feature. NED* $\mathcal{D}$ *is defined as*

$$\begin{aligned} D &= \left\{ \left(x_i - M^{(j)} \right)^T \left(x_i - M^{(j)} \right) \right\} / \left\{ \frac{1}{n} \sum_{r=1}^{n} \left(\sigma_r^G + \sigma_r \right) \right\} \\ &= \left\{ \begin{bmatrix} x_{i1} - M_1^{(j)} & x_{i2} - M_2^{(j)} & \ldots & x_{in} - M_n^{(j)} \end{bmatrix} \begin{bmatrix} x_{i1} - M_1^{(j)} \\ x_{i2} - M_2^{(j)} \\ \cdots \\ x_{in} - M_n^{(j)} \end{bmatrix} \right\} / \\ &\quad \left\{ \frac{1}{n} \sum_{r=1}^{n} \sqrt{\sigma_r^G \times \sigma_r} \right\} \end{aligned} \tag{7}$$

where n stands for the number of dimensions for each pattern.

3.3 Algorithm for Group Classification

In our study, two kinds of K-means methods are developed, namely, class-K-means and sample-K-means. The algorithm for the group classification proposed in this paper can be described by a block diagram, as shown in Fig. 5. The operation of this algorithm is divided into three phases:

- Initial group clustering—Choose K initial groups. These are arbitrary and selected using a randomizer;
- Class-K-means clustering—Use the expected values for each class in the ordinary K-means algorithm;
- Sample-K-means clustering—To implement overlaps clustering, this phase uses the individual value of each sample in the ordinary K-means algorithm.

Algorithm 1: Group Classification

Input: m classes of characters, $x_1, x_2, \ldots, x_i, \ldots, x_m$, each class consists of n samples, i.e., $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$;

Output: K clusters, $S_1, S_2, \ldots, S_k$;

Step 1: Arbitrarily choose K classes as the initial cluster centers:

$$M^{(1)}(0), M^{(2)}(0), \ldots, M^{(j)}(0), \ldots, M^{(K)}(0);$$

Step 2: Distribute all of m characters (classes $x_1, x_2, \ldots, x_m$) among

the chosen cluster domains according to (7);

Step 3: Recompute the cluster centers, i.e., take the average of the classes in their domains as their new cluster centers with the following formula:

$$M^{(j)}(t+1) = \frac{1}{N_j} \sum_{x \in S_j(t)} x, \quad j = 1, 2, \ldots, K;$$

Step 4: If $M^{(j)}(t+1) = M^{(j)}(t)$ then go to Step 2, else go to Step 5;

Step 5: Distribute all of $m \times n$ samples (x_{i1}, x_{i2}, ..., x_{in}, $i = 1, 2, ..., m$) among the chosen cluster domains according to (7);

Step 6: Update the cluster centers with the following formula:

$$M^{(j)}(t+1) = \frac{1}{N_j} \sum_{x_{rs} \in S_j(t)} x_{rs}, \quad j = 1, 2, \ldots, K;$$

Step 7: If $M^{(j)}(t+1) = M^{(j)}(t)$ then go to Step 5, else go to Step 8;

Step 8: End.

After the group classification, characters x_1, x_2, ..., x_m have been clustered into K groups S_1, S_2, ..., S_p, S_q, ..., S_K with overlap. $M^{(1)}$, $M^{(2)}$, ..., $M^{(p)}$, $M^{(q)}$, ..., $M^{(K)}$ are their centers. η_1, η_2, ..., η_p, η_q, ..., η_K denote the number of candidates in groups S_1–S_K, respectively. Thus, the group classifier $\Re_{gc}$ can be represented as

$$\Re_{gc} = \begin{bmatrix} S \\ M \\ \aleph \end{bmatrix} = \begin{bmatrix} S_1 & S_2 & \cdots & S_p & S_q & \cdots & S_K \\ M^{(1)} & M^{(2)} & \cdots & M^{(p)} & M^{(q)} & \cdots & M^{(K)} \\ \eta_1 & \eta_2 & \cdots & \eta_p & \eta_q & \cdots & \eta_K \end{bmatrix}. \tag{8}$$

4 Multilevel Character Classification

There are five levels in the multilevel character classification (MLCC). In the first level, a Gaussian distribution selector is used to select a smaller number of candidates from several groups. From the second level to the fifth one, matching approachs using different features are employed, respectively. Because the matching approach used is a standard technique, it will be omitted here in order to save space. In this section, we concentrate on the first level, where Gaussian distribution selector will be presented in detail.

From the group classifier, m classes of characters have been clustered into K groups with overlaps. In the first level of the MLCC, for each input character, the similarities between it and different groups are computed. The order of the similarities can be arranged in ascending, or low-to-high, order by the *Gaussian Distribution Selector (GDS)*. Then the first h_k nearest clusters are selected.

Let x be an input character. The density function of $P(x \mid S_i)$ can be expressed by Gaussian distribution function $N(\mu^{(i)}, \Sigma_i)$. The result of the group classification can be represented by (8). The similarities can be calculated using the normalized Euclidean distance represented by (7) and are listed in ascending order:

$$\begin{bmatrix} S_1 & S_2 & \cdots & S_p & S_q & \cdots & S_K \\ M^{(1)} & M^{(2)} & \cdots & M^{(p)} & M^{(q)} & \cdots & M^{(K)} \\ \eta_1 & \eta_2 & \cdots & \eta_p & \eta_q & \cdots & \eta_K \end{bmatrix}$$

$$\forall_{p<q}(\mathcal{D}(x, M^{(p)}) < \mathcal{D}(x, M^{(q)})).$$

The density function of $P(x \mid S_i)$ can be expressed by Gaussian distribution function $N(\mu^{(i)}, \Sigma_i)$.

A candidate selector based on Gaussian distribution function has been designed in this system, GDS. From GDS, h_k clusters S_1, S_2, ..., S_{hk} have been selected, such that

$$\begin{gathered} \overbrace{S_1, S_2, \ldots, S_{h_k}}^{i \le h_k}, \overbrace{S_{h_k+1}, S_{h_k+2}, \ldots, S_K}^{j > h_k} \\ \left(\mathcal{D}\left(x, M^{(i)}\right)\right)_{i \le h_k} < \left(\mathcal{D}\left(x, M^{(j)}\right)\right)_{j > h_k} \end{gathered}. \tag{9}$$

Suppose each class consists of β samples. In the process of the multilevel character classification, the reductions of the number of candidates can be presented as follows:

- Level 1: The number of candidates is reduced from $\beta \times m$ to $\sum_{i=1}^{h_k}(S_i \times \eta_i)$ by a rate of $\frac{1}{R_1}$;
- Level 2: The number of candidates is reduced from $\sum_{i=1}^{h_k}(S_i \times \eta_i)$ to $\frac{1}{R_2}\sum_{i=1}^{h_k}(S_i \times \eta_i)$;
- Level 3: The number of candidates is reduced from $\frac{1}{R_2}\sum_{i=1}^{h_k}(S_i \times \eta_i)$ to $\frac{1}{R_2R_3}\sum_{i=1}^{h_k}(S_i \times \eta_i)$;
- Level 4: The number of candidates is reduced from $\frac{1}{R_2R_3}\sum_{i=1}^{h_k}(S_i \times \eta_i)$ to $\frac{1}{R_2R_3R_4}\sum_{i=1}^{h_k}(S_i \times \eta_i)$;
- Level 5: The candidates become unit from $\frac{1}{R_2R_3R_4}\sum_{i=1}^{h_k}(S_i \times \eta_i)$ by $\frac{1}{R_5}$;

The above reductions can be illustrated in the form of

$$\begin{aligned} &\beta \times m \overset{Level\,1}{\Rightarrow} \frac{1}{R_1}(\beta \times m) = \sum_{i=1}^{h_k}(S_i \times \eta_i) \overset{Level\,2}{\Rightarrow} \\ &\frac{1}{R_2}\sum_{i=1}^{h_k}(S_i \times \eta_i) \overset{Level\,3}{\Rightarrow} \frac{1}{R_2R_3}\sum_{i=1}^{h_k}(S_i \times \eta_i) \overset{Level\,4}{\Rightarrow} \\ &\frac{1}{R_2R_3R_4}\sum_{i=1}^{h_k}(S_i \times \eta_i) \overset{Level\,5}{\Rightarrow} \frac{1}{R_2R_3R_4R_5}\sum_{i=1}^{h_k}(S_i \times \eta_i) = 1 \end{aligned} \tag{10}$$

The reduction rates R_1, R_2, R_3, R_4, and R_5 depend on the number of pattern classes and the quality of the samples. In our work, they can be selected as

R_1	R_2	R_3	R_4	R_5
$3 \sim 5$	$3 \sim 5$	$4 \sim 6$	$6 \sim 8$	$10 \sim 15$

5 Experimental Results

Experiments have been conducted on two types of handwritten data: one type with average quality and another type with above average quality. The database used in our study for training and testing consists of 5,401 daily-used Chinese characters. Each character consists of 200 samples. Thus, there are a total of 1,180,000 samples. They were selected from 1.4 million samples written by 3,000 people in Taiwan. More details about the database can be found in [5]. For each class, 200 samples have been separated equally into two groups:

1) training group has 100 samples odd numbered and
2) testing group has 100 samples even numbered.

In this experiment, four feature vectors, V_1, V_2, V_3, V_4, have been used:

1) Sixty-four features have been employed in the stage of group classification. They are peripheral shape features and regional stroke density ones. They constitute a 64-dimensional vector:

$$V_1 = \left(\aleph_h^i \aleph_v^i \Psi_h^i \Psi_v^i \Re_h^i \Re_v^i \Phi_h^i \Phi_v^i\right)^T, \; i = 1, 2, \ldots, 8$$

2) For the multilevel character classification, HPLD, VPLD, HRSD, VRSD, HRSDD, and VRSDD produce a 48-dimensional vector, which has been used in its first level.

TABLE 1
GROUP CLASSIFICATION

Number of Classes	5,401
Number of Samples	540,100
Recognition Rate	99.76% (538,712/540,100)
Overlaps	3.8

TABLE 2
REDUCTIONS OF NUMBER OF CLASSES IN THE MULTILEVEL CLASSIFICATION

Number of Classes	5,401
Number of Classes in the First Level	1,000-1,200
Number of Classes in the Second Level	200-400
Number of Classes in the Third Level	50-150
Number of Classes in the Fourth Level	10-15
Number of Classes in the Fifth Level	1

$$V_2 = \left(\Psi_h^i \Psi_v^i \Re_h^i \Re_v^i \Phi_h^i \Phi_v^i\right)^T,\ i = 1, 2, \ldots, 8$$

3) HPBA, VPBA, and four-orientation stroke directions (4-OSD) $\Lambda_0^j, \Lambda_1^j, \Lambda_3^j, \Lambda_4^j$ $(j = 1, 2, \ldots, 6)$ form a 40-dimensional feature vector V_3, which has worked in the second and third levels.

$$V_3 = \left(\aleph_h^i \aleph_v^i \Lambda_k^j\right)^T,\ i = 1, 2, \ldots, 8;\ j = 1, 2, \ldots, 6;\ k = 1, 2, \ldots, 4$$

4) The eight-orientation stroke directions (8-OSD) $\Lambda_0^j, \Lambda_1^j, \ldots, \Lambda_8^j$ $(j = 1, 2, \ldots, 6)$ become 48-dimensional feature vector V_4, which is used in the last two levels:

$$V_4 = \left(\Lambda_k^j\right)^T\ \ j = 1, 2, \ldots, 6;\ k = 1, 2, \ldots, 8.$$

In the group classification, the number of clusters affects the final recognition rate of the whole recognition system. If we choose K too small, then the load of the next levels of classification (character classifier) will be heavy, which will result in a lower recognition rate. On the other side, if we choose K too large, then the load of the group classification will be massive, making the hit recognition rate go down and, finally, making the total recognition error rate go up. In our study, according to the experiments, the number of clusters in the group classifier has been chosen to be 72, namely, $K = 72$. The number of overlaps among these groups is 3.8 in our experiment. The hit recognition rate is over 99 percent. The experimental result of the group classification is listed in Table 1.

Experimental results for the multilevel classification can be found in Table 2, Table 3, and Table 4, respectively. In the first level, seven groups of candidates have been selected from 72 clusters by the GDS. This level outputs about 1,000 classes. Although the number of classes has been reduced by the first level, it is still large; it is necessary to decrease them gradually by the rest of the multilevel classification scheme. The reductions through each level are shown in Table 2.

From the experimental results shown in Table 3 and Table 4, it is obvious that this system possesses a high recognition rate for offline handwriting with a large set of Chinese characters. After processing through three levels of the multilevel character classification, the recognition rate is near 80 percent for unique candidates and 96.60 percent for 10 multichoice candidates. The final results form the fifth level of the multilevel character classifier, the accuracy rate is 89.03 percent for a unique candidate and 98.13 percent for 10 multichoice candidates.

By analyzing the recognition errors in our experiments, it can be found that the major reason of misrecognitions is the ambiguities of shapes in some Chinese characters. Several examples of the ambiguous pattern pairs are illustrated in Fig. 6. In some of the pairs, for instance, (a) and (b), (c) and (d), (e) and (f), and (i) and (j) in Fig. 6, the similar characters differ only in the short strokes. Some ambiguous pattern pairs differ only in their radical parts, such as in (g) and (h) and in (k) and (l) in Fig. 6.

This system has been implemented on a 486/33 MHz IBM personal computer. The average recognition speed is 3.5 characters per second.

TABLE 3
SELECTION OF CANDIDATES IN THE FIRST THREE LEVELS

Selection of Candidates in the First Level							
Number of Candidates	1	2	3	4	5	10	20
Recognition Rate	55.67%	67.77%	73.57%	77.19%	79.71%	86.30%	91.27%
Number of Candidates	30	40	50	60	70	80	90
Recognition Rate	93.51%	94.80%	95.69%	96.31%	96.79%	97.18%	97.48%
Selection of Candidates in the First Level							
Number of Candidates	1	2	3	4	5	10	20
Recognition Rate	75.51%	84.98%	88.64%	90.72%	92.05%	95.11%	97.08%
Number of Candidates	30	40	50	60	70	80	90
Recognition Rate	97.86%	98.30%	98.67%	98.76%	98.89%	98.99%	99.08%
Selection of Candidates in the Fourth Level							
Number of Candidates	1	2	3	4	5	10	20
Recognition Rate	79.61%	88.31%	91.47%	93.15%	94.21%	96.60%	98.04%

TABLE 4
SELECTION OF CANDIDATES IN THE LAST TWO LEVELS

Selection of Candidates in the Fourth Level						
Number of Candidates	1	2	3	4	5	10
Recognition Rate	88.4761%	94.00%	95.67%	96.51%	97.02%	98.12%
Selection of Candidates in the Fifth Level						
Number of Candidates	1	2	3	4	5	10
Recognition Rate	89.03%	94.34%	95.91%	96.69%	97.15%	98.13%

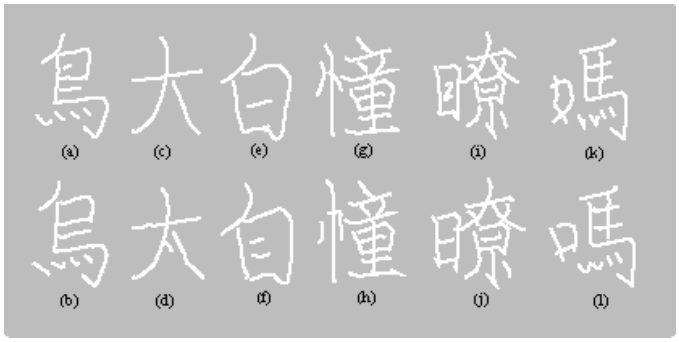

Fig. 6. Examples of ambiguous pairs of Chinese characters.

6 Conclusions

In this paper, a new offline system has been presented to recognize Chinese handwritten characters. Recognizing Chinese handwriting is a challenging topic in the area of character recognition. In contrast with online recognition, offline recognition is much more difficult. The offline recognition system developed here is based on a multimodel that includes a multifeature scheme and a multilevel classifier.

The selection of stable and representative sets of features is the heart of the design of a pattern recognition system. In this system, four feature vectors, V_1, V_2, V_3, V_4, have been used. V_1 consists of 64 features, V_2: 48, V_3: 40, and V_4: 48. These are called multiple features extracted from peripheral shapes, stroke densities, and stroke directions.

Chinese characters have three characteristics:

1) The vocabulary is typically large, involving more than 3,000 classes.
2) The structures of Chinese characters are much more complex than those of alphanumeric letters.
3) Many Chinese characters have similar shapes.

The design of classifier is an important task for such a huge number of complex Chinese characters. In our system, the classifier has been designed elaborately, where a multilevel classifier has been developed. It contains a group classifier and a five-level character classifier, where two new technologies, overlap clustering and Gaussian distribution selector, have been developed.

This system has been implemented on a 486/33 MHz IBM personal computer. Experiments have been conducted on both quality and general quality handwritten data. The database used in our study for training and testing consists of 5,401 daily-used Chinese characters, with 200 samples. Thus, there are a total of 1,180,000 samples that have been selected from 1.4 million handwritten by 3,000 people in Taiwan. From the experimental results, it is obvious that this system possesses a high recognition rate for offline handwriting with a large set of Chinese characters. The final recognition rate from the five-level character classifier is 89.03 percent for unique candidates and 98.13 percent for 10 multichoice candidates. The average recognition speed is 3.5 characters per second.

To improve this system, the following should be done in our further research:

- We have not fully investigated the effect of a multilevel classifier on the performance and the error rate, although we think that we can approach this by analyzing the effects at the individual level or pairs of consecutive levels. This effect needs a lot of in-depth investigation that is beyond the scope of this study, and we will do it as one of our further tasks.
- Some classes of multifeatures that are used in this system could be correlated each other and could greatly reduce their efficiency for handwritten Chinese character recognition, for example, HPBA with HPLD, VPBA with VPLD, HRSD with HRSDD, VRSD with VRSDD, and 4-OSD with 8-OSD. We will remove the redundancy and find other effective features.
- The samples used in this study are traditional Chinese characters from a database in Taiwan. In the next study, the simplified form from Beijing will be applied as the samples.

Acknowledgments

This work was supported by research grants received from Research Grant Council (RGC) of Hong Kong and Faculty Research Grant (FRG) of Hong Kong Baptist University. This work was also a partial result of the project no. 35N1300 conducted by the Industrial Technology Research Institute under the sponsorship of the Minister of Economic Affairs, Republic of China, and the Hallym Academy of Science, Hallym University, Korea.

References

[1] T.H. Hildebrandt and W. Liu, "Optical Recognition of Handwritten Chinese Characters: Advances Since 1980," *Pattern Recognition*, vol. 26, no. 2, pp. 205-225, 1993.

[2] S.-W. Lee and J.-S. Park, "Nonlinear Shape Normalization Methods for the Recognition of Large-Set Handwritten Characters," *Pattern Recognition*, vol. 27, no. 7, pp. 895-902, July 1994.

[3] C.Y. Suen, Y.Y. Tang, and Q.R. Wang, "Feature Extraction in the Recognition of Chinese Characters Printed in Different Fonts," *Proc. 1986 Int'l Conf. Chinese Computing*, pp. 136-143, Singapore, Aug. 1986.

[4] C.C. Tapperet, C.Y. Suen, and T. Wakahara, "The State of the Art in Online Handwriting Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 8, pp. 787-808, Aug. 1990.

[5] L.-T. Tu, Y.S. Lin, C.P. Yeh, I.S. Shyu, J.L. Wang, K.H. Joe, and W.W. Lin, "Recognition of Handprinted Chinese Characters by Feature Matching," *Proc. First Nat'l Workshop Character Recognition*, pp. 166-175, 1991.