Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 6, October 2017, pp. 985-997
DOI: 10.23940/ijpe.17.06.p19.985997

A Method for Segmenting Uyghur Symbols

Xiangwei Qi, Yong Yang, Weimin Pan*

School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China

Abstract

In consideration that Uyghur symbols are poorly recognized and recognition algorithms are impertinent for connecting characters, five kinds of features are extracted from Uyghur symbols after feature analysis and pre-processing, including 8-directional features, fuzzy features and primitive features. According to characteristics of Uyghur, features of Uyghur symbols such as aspect ratio and dynamic speed are extracted. Features are selected by LDA transformation, and the methods for judging relationships among Uyghur primitives are determined. Minimum distance classifier and MQDF classifier are suggested to be used. By recognizing primitives, effectiveness of algorithms is experimentally verified with online Uyghur recognition algorithms of decision fusion strategies by classifiers.

Keywords: Uyghur language; feature extraction; symbol segmentation; feature classification

(Submitted on May 29, 2017; Revised on July 12, 2017; Accepted on September 17, 2017)

1. Introduction

This paper aims at strokes, which, in a strict sense, refers to an integration of all Uyghur letters or connected strokes that are entered online by users. After a range of processing, the Uyghur letters or connected segments corresponding to this group of strokes are identified. The key of this survey consists in conversion of strokes into letters or connected segments. To this end, two major problems are concerned. One is recognition of Uyghur letters or connected segments. The degree of matching between an input set of strokes and some Uyghur letters or connected segments shall be identified. The other is segmentation of Uyghur symbols. For a set of strokes that can't be generally recognized, it is necessary to confirm which strokes of the set can be recognized as a whole. In other words, that set of strokes is supposed to be grouped, in order that each group of strokes can be recognized as a Uyghur letter or connected segment. To solve these two problems, favorable support is required from basic databases of handwriting samples.

To recognize Uyghur symbols, it is essential to find out the characteristic differences between a Uyghur symbol and other symbols. In this paper, letters, connected segments or words input by hands are considered as a whole. These differences are just boundaries among different Uyghur symbols.

At present, domestic and foreign studies about recognition of handwritten Uyghur words have been mostly performed based on recognition of Chinese and English characters. The algorithms for recognizing Uyghur letters mainly extract relationships among directional element features, structural letter features and strokes. The classification methods mostly include Euclidean distance classifier for Chinese classification, SVM, HMM and neural networks. Owing to limitations upon inherent features of agglutinative language, the final recognition effects are limited to certain extent no matter in theories or practices [3,4].

2. Feature Extraction

To acquire features of Uyghur symbols and connected segments more completely, this paper creates a sample database after pre-processing and improves recognition through classifier fusion.

* Corresponding author.
E-mail address: *47266861*@qq.com.

2.1. Direction Features

To recognize characteristics based on strokes, there is a need to take direction of strokes into account. Therefore, characters are usually denoted by directional statistical vectors in statistical classification. For the time being, four directional features have been widely utilized for recognizing Chinese and English characters, having achieved relatively ideal effects for recognizing characters. In recognizing handwritten Chinese characters, the aforementioned four directions correspond to four Chinese strokes, namely horizontal stroke, vertical stroke, left-falling stroke and right-falling stroke. In HCL2004, four directional features are expanded to 8 features and successfully applied in identifying Chinese handwritten characters.

In consideration of handwriting features of Uyghur, their strokes can't be simply summarized as horizontal stroke, vertical stroke, left-falling stroke and right-falling stroke. In view that very few features can be extracted from Uyghur, directional features of strokes can be hardly elaborated in detail according to 4-directional features. Since it is highly effective for recognizing handwritten Chinese characters based on 8-directional features, this paper extracts 8-directional features of Uyghur symbols.

To extract these features, stroke directions are decomposed into several fixed angled. Subsequently, the number of strokes at each angle is reckoned as eigenvalue. The set of number of directional stroke segments composes a histogram, which is referred to as directional histogram. In each sampling area of Uyghur symbols, 8-directional features are extracted and determined. Assuming that the direction vector of the point $P_j = (x_j, y_j)$ is $\underset{v_j}{\rightarrow}$,is defined as $\underset{v_j}{\rightarrow}$:

$$\underset{v_j}{\rightarrow} = \begin{cases} \overrightarrow{P_j P_{j+1}}, P_j \ is \ start \ point \\ \overrightarrow{P_{j-1} P_{j+1}}, P_j \ is \ intermediate \ point \\ \overrightarrow{P_{j-1} P_j}, P_j \ is \ finish \ point \end{cases} \tag{1}$$

$\underset{v_j}{\rightarrow}$ mmodulus is normalized and mapped upon two directions: $d_j^1 \ and \ d_j^2$, where the 8-directional features of d_j^1 is a_j^1, and 8-directional features of d_j^1 and a_j^2. Provided that P_j is not the starting point or finish point of sampling, a_j^1 and a_j^1 shall be calculated as follows [2]:

$$a_j^1 = \frac{|d_x - d_y|}{s} \tag{2}$$

$$a_j^2 = \frac{\sqrt{2} min(d_x, d_y)}{s} \tag{3}$$

Where, d_x=$|x_{j+1}$-$x_{j-1}|$, d_y=$|y_{j+1}$-$y_{j-1}|$, $s = \sqrt{d_x^2 + d_y^2}$. Provided that P_j is a starting or finish point, the conditions shall be altered. Assuming that P_j is a starting point, d_x=$|x_{j+1}$-$y_{j-1}|$ and d_y=$|x_{j+1}$-$x_j|$, If it is a finish point, d_x=$|x_j$-$x_{j-1}|$ and d_y=$|y_j$-$y_{j-1}|$.As shown in the following Figure 1.

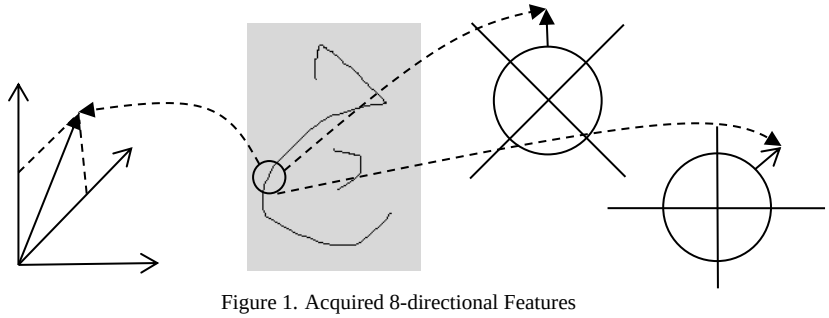

Figure 1. Acquired 8-directional Features

After 8-directional features of all points (*P*) are calculated by above methods, the overall 8-directional feature model may be conveyed as {B_d=[f_d(x,y)], x,y=1,2,…,64; d=D1,D2,…,D8}. Assuming that $f_{d_j^1}(x, y) = a_j^1$ and $f_{d_j^2}(x, y) = a_j^2$(j is the number of *P*). Besides, the value of all other points other than d_j^1 and d_j^2 is given to be zero.

Subsequently, the ultimate 8-directional features are determined. First of all, a max filter is used for setting the value within any 3*3 neighbourhood in each model. Next, the graphs in each direction are divided into 8*8 grids, the center of which is reckoned as sampling center of the area. Then, Gauss filter is employed, so 8-directional features may be conveyed as follows [5]:

$$F_d(x_i, y_i) = \sum_{x=-N}^{y=N} \sum_{y=-N}^{y=N} f_d(x_i + x, y_i + y) G(x, y) \quad (4)$$

Where, (x_i, y_i) is the coordinate of the sampling point, N is a filtered area. Here, N is set to be a constant and determined as 2λ. $G(x, y) = \frac{\kappa^2}{\delta^2} e^{-\frac{\kappa^2(x^2+y^2)}{2\delta^2}}$ (λ is wavelength (determined to be 8 in this paper), κ=2π/λ, δ=π) is a Gaussian model and f_d is a model generated after maximum filtering. All 8-directional features are combined to determine 8-direction vector. Each handwritten Uyghur symbol has 8 models, so there are 8*8 sampling points of directional features in each model, and each Uyghur symbol may be expressed by 8*8*8=512-dimension vector.

2.2. Fuzzy Characteristics

After Uyghur symbols are pre-processed through normalization, a bounding box of Uyghur symbols made up of 12*8 small squares can be defined, in order to extract the fuzzy characteristics. Without considering direction of strokes, the small squares may be numbered $A_i (i \in \{1,2,\cdots,96\})$ and ordered from the left to the right, or in a top-down manner. Figure 2 shows the bounding box of a connected segment composed of three Uyghur letters.

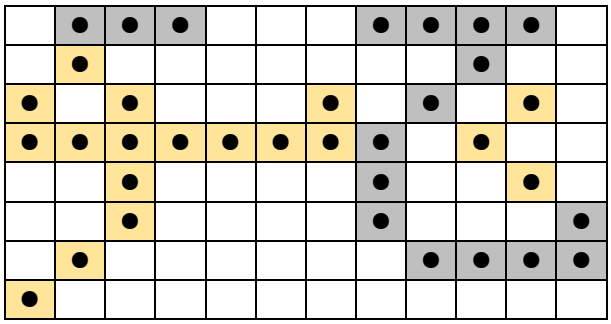

Figure 2. Acquired 8-directional Features

Assuming that A_i has N_i sampling points and $N_i > 0$.

$$\bar{x}_{A_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} v_j^x \quad (5)$$

$$\bar{y}_{A_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} v_j^y \quad (6)$$

Where, v_j^x and v_j^y are A_i x and y coordinates inside the squares, while the coordinate of each average point is denoted as $T_{A_I} = (\bar{x}_{A_i}, \bar{y}_{A_i})$.

As shown in Figure 2, the average points of Uyghur symbols are seen inside the small squares. Each square only has one point. Even if these points can't exactly convey features of Uyghur symbols, they can record the basic skeleton of a Uyghur symbol and are favorable for identifying symbols.

Here, the width and height of the bounding box of Uyghur symbols are conveyed by *w* and *h* respectively. The coordinate of the point in the upper left corner is (x_{TL}, y_{TL}), the average point-to-point (x_{TL}, y_{TL}) distance of the bounding box is as follows:

$$D_{A_i} = \left\{ \left[\frac{64}{w} \left(\bar{x}_{A_i} - x_{TL} \right) \right]^2 + \left[\frac{84}{h} \left(\bar{y}_{A_i} - y_{TL} \right) \right]^2 \right\}^{\frac{1}{2}} \tag{7}$$

By multiplying the above formula by $\frac{64}{w}$ and $\frac{84}{h}$, the size is normalized as 84×64, and D_{A_i} is a feature for symbol recognition. In this case, each Uyghur symbol has 96 features.

Supposing that each Uyghur symbol has n features and collects m samples, the symbol will form a m*n matrix of feature sets. The D_{A_i} of 96 small squares is separately calculated. Next, the mean and variance of each feature are calculated. Thus, 96 means and 96 variances are determined for each Uyghur symbol. They are calculated as follows:

$$d_{A_i} = \frac{1}{N_{A_i}} \sum_{j=1}^{N_i} D_{A_i}^j \tag{8}$$

$$\delta_{A_i}^2 = \frac{1}{N_{A_i}} \sum_{j=1}^{N_i} \left(D_{A_i}^j - d_{A_i} \right)^2 \tag{9}$$

Where, N_{A_i} is A_i the number of samples inside squares and $D_{A_i}^j$ is the feature of A_i in the j[th] sample. Then, d_{A_i} and $\delta_{A_i}^2$ are stored in the sample library as features of Uyghur symbols.

2.3. Elementary Features of Uyghur Symbols

Uyghur symbols are arbitrarily written, from which very few features can be extracted, and many local features will be wasted even when only overall features of strokes are extracted. Geometric division of strokes and extraction does not only increase efficiency of recognition, but are also effective for identifying more features.

In writing Uyghur by hands, strokes of symbols are mostly composed of straight lines and arcs. Even if they are points, they are usually extended into short straight lines, and multiple points are often extended to form an arc. According to ideas of Mohamed et al. for feature classification, this paper particularly classifies Uyghur strokes into three categories, namely straight lines, arcs and points.

Among straight lines and arcs of Uyghur symbols, arc changes are generally significant. In this paper, strokes are segmented and classified by extracting sudden turns of strokes along the directions where the strokes are generated. The segmented parts are known as primitives, which may convey some geometric features of strokes. In this paper, primitives of Uyghur symbols are classified as follows, including straight lines (horizontal, perpendicular, leftward-sloping and rightward-sloping lines), arcs (O-shaped, ∪-shaped, ∩-shaped, ⊂-shaped, ⊃-shaped, S-shaped and ∼-shaped arcs) and points, as shown in Figure 3:

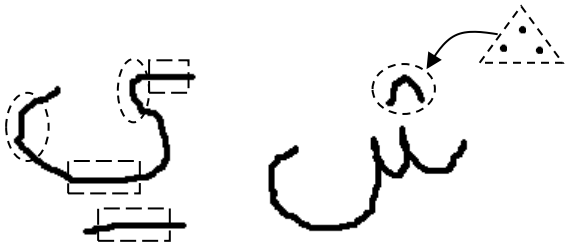

Figure 3. The framework of cell image classification
(As shown above, □, ○ and △ represent straight lines, arcs and points respectively.)

It is necessary to judge which category the strokes of an input Uyghur symbol belong to, or whether there are sudden turning points. Provided that there are n sudden turning points (n>0) in a stroke, the stroke will be segmented into n+1 primitives. In case of no sudden turning point (n=0), the whole stroke will be reckoned as a primitive to record its features.

Supposing that the point sequence of strokes of a Uyghur symbol is p_1, p_2, ..., p_n, the angle between $\overline{p_i p_{i-3}}$ and $\overline{p_i p_{i+3}}$ is $\theta_i < \delta$, there will be sudden turns in the strokes. In this paper, the value of the threshold δ is taken as 90°, mainly because there are many arcs in Uyghur symbols. When abrupt changes happen to strokes, the angle θ_i will be generally smaller, as shown in Figure 4.

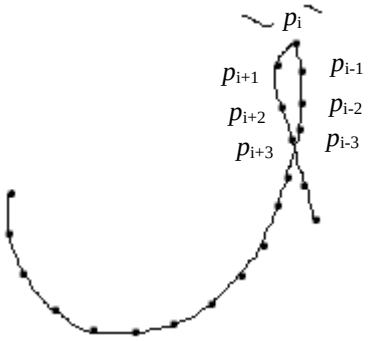

Figure 4. Sudden Turns

In many cases, there will be several consecutive points when the angle between vectors (a) is smaller than 90°. For instance, for the angle between $\overline{p_{i-1}p_{i-4}}$ and $\overline{p_{i-1}p_{i+2}}$ in Figure 4, only a turning point is taken as a point of division. Assuming that the angle between the point p_n, p_{n+1}, ... , p_{n+t} and the third point in front of or behind it is smaller than 90°, and p_k is a sudden turning point, $k = n + \text{int}(\frac{t+1}{2})$. It is difficult to confirm whether a stroke is a straight line or an arc, so this paper determines categories of primitives of Uyghur symbols according to fuzzy logics.

(1) Straight lines

Supposing that the point sequence on primitives is p_1, p_2, ..., p_r, the value of μ will be determined by the following formula to reflect the curvature of strokes:

$$\mu = {d(p_1, p_r)}\Big/{\sum_{k=1}^{r-1} d(p_k, p_{k+1})} \tag{10}$$

Where, $d(p_i,p_j)$ is the distance between p_i,p_j . Through an experiment, it is clarified in this paper that when $\mu > 0.9$, the primitives will be straight lines, or else they will be arcs. If they are straight lines, following membership functions will be defined to identify categories of straight lines:

Functions of perpendicular lines:

$$\mu_V = max\{f(x, 90,90), f(x, 90,270)\}$$

Functions of horizontal lines:

$$\mu_H = max\{f(x, 90,0), f(x, 90,180), f(x, 90,360)\}$$

Functions of rightward-sloping lines:

$$\mu_R = max\{f(x, 90,45), f(x, 90,225)\}$$

Functions of leftward-sloping lines:

$$\mu_L = max\{f(x, 90,135), f(x, 90,315)\}$$

Where, $f(x, b, c) = \begin{cases} 1 - 2\times\left|\frac{x-c}{2}\right| & (c - {b}/{2}) \le x \le (c + {b}/{2}) \\ 0 \end{cases}$

x is direction angle of a right angle, b=90°, and *c* is the maximum, dependent upon category of straight lines.

(2) Arcs

It is critically important to acquire effective features by classifying arcs of Uyghur symbols. Once strokes are classified based on features of sudden turning points, the arc primitives shall be classified. In order to not make the classification of primitives too detailed, a primitive is suggested to contain six points at least in this paper. In the following formula, when $\mu < 0.1$, the primitive will be a ring. Many strokes of Uyghur symbols are written clockwise and counterclockwise. In case that certain stroke can be written in these two ways, the starting and finishing points are nearly on a horizontal line, the stroke of the arc will be "∼-shaped". The arc-shaped stroke will be "S-shaped" if the starting point is nearly perpendicular to the finish point. For other arcs, the category of primitives is judged in line with fuzzy theories, and their membership functions are respectively listed as follows:

$$\mu_{\cap} = min\left[1, \frac{\sum_{i=0}^{n} a_y}{n}\right], a_y = \begin{cases} 1, y_i > \dfrac{y_s + y_E}{2} \\ 0 \end{cases} \tag{11}$$

$$\mu_{\cup} = min\left[1, \frac{\sum_{i=0}^{n} b_y}{n}\right], b_y = \begin{cases} 1, y_i > \dfrac{y_s + y_E}{2} \\ 0 \end{cases} \tag{12}$$

$$\mu_{\subset} = min\left[1, \frac{\sum_{i=0}^{n} l_x}{n}\right], l_x = \begin{cases} 1, x_i > \dfrac{y_s + y_E}{2} \\ 0 \end{cases} \tag{13}$$

$$\mu_{\supset} = min\left[1, \frac{\sum_{i=0}^{n} r_x}{n}\right], r_x = \begin{cases} 1, x_i > \dfrac{y_s + y_E}{2} \\ 0 \end{cases} \tag{14}$$

2.4. Relationships between Primitives and Their Codes

A Uyghur symbol is made up of one or several strokes. To better recognize features of strokes with primitives, spatial positions, including top bottom, left and right, of the strokes are extracted and utilized. In view of practical requirements for the segmentation process, primitives shall be identified at first prior to judging their relationships. In case of any new primitives, it will be necessary to judge their spatial relationships with previous primitives and whether the stroke is a symbol or part of the symbol. If not, previous strokes will exactly constitute a symbol and the new primitives will be categorized as other symbols. On the contrary, the new primitives and previous strokes would make up a symbol. By repeating this process, strokes of Uyghur symbols are processed with codes of primitives, and membership of primitives is determined based on fuzzy theories. MPRF algorithms (namely improved PRF algorithms) are utilized. Likewise, it is stipulated in MPRF algorithms that any Uyghur symbol to be recognized shall not have more than 10 strokes. This limitation is in accordance with the practical situation. In general, a Uyghur letter only has less than 4 strokes. For common connected segments and words, there are no more than 10 strokes.

The MDRF algorithm is generally as follows. Firstly, primitives are extracted, and strokes are numbered according to primitives. Then, spatial relationships of strokes are judged to unify strokes written in different orders by identifying relationships between primitives and codes. Consult the dictionary of primitive relations in the sample library to confirm whether any Uyghur symbol reveals such encoding relationship. If so, perhaps the symbol would be a substroke of another symbol. Instead, it is necessary to continue extracting the next stroke. This process shall be repeated in order to identify relationships of all strokes in order. Provided that this group of input codes is finally not found in the dictionary, fuzzy methods shall be used for determining the membership of the character that is the closest to the inputs. A threshold variable is defined. In case the maximum degree of membership is higher than the threshold, the input recognition result is the symbol with the highest degree of membership. In contrast, it means that the primitives are not successfully recognized. Next, remove the last stroke for re-coding, and further consult the dictionary of codes. This process shall be repeated until only a stroke is left.

(1) Primitive Coding

To extract "joint features" from strokes, primitives are coded for strokes of all input Uyghur symbols.

Line: "—" is a horizontal line； "|" is a perpendicular line, "/" is a rightward-sloping line and "\" is a leftward-sloping line.

Arc: "∪" is a concave arc; "∩" is a convex arc; "⊂" is a leftward convex arc; "⊃" is a rightward convex arc; "S" is a perpendicular continuous arc; "~" is a horizontal continuous arc and "O" is a closed arc.

Dot.

The codes of primitives are shown as follows:

Table 1. Codes of Primitives

Primitives	O	—	|	/	\	∪	∩	⊂	⊃	S	~	dot
Codes	0	1	2	3	4	5	6	7	8	9	a	b

(2) Relationships between Primitives

For convenience, 9 kinds of relationships are defined between primitives of Uyghur symbols, including left horizontal, left lower, left upper, right under, right above, right horizontal, upper right, lower right and crossing.

In light of these relationships, the codes are defined as follows:

Table 2. Codes of Primitives

Relations	Right Horizontal	Right Upper	Above	Left Upper	Left Horizontal	Lower Left	Below	Lower Right	Cross
Codes	0	1	2	3	4	5	6	7	8

2.5. Aspect Ratio

The aspect ratio of a character refers to the ratio of the width to the height of the bounding box of a character. Unlike Chinese and English, each Uyghur character has similar aspect ratio. This ratio is an evident feature for classifying letters of Uyghur symbols to be recognized. It can be considered as a feature for effectively distinguishing certain type of characteristics. In this paper, it is specified that the aspect ratio won't be judged for Uyghur symbols with more than 4 strokes. On the contrary, it will be identified. Figure 5 shows several instances of aspect ratio. It has been experimentally demonstrated that this feature is quite effective for broad classification.

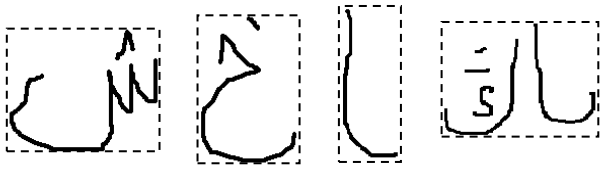

Figure 5. Several Instances of Aspect Ratio
(Figure5(1-3) show the letters, while common connected segments are indicated in Figure4(4))

2.6. Features of Dynamic Speed

The writing process of Uyghur symbols differs from those of Chinese and English characters. In some nodes of strokes, the changes to writing speed are evident. It is helpful for recognizing characters and classifying primitives by statistically analyzing and recording these changes. Figure 6 shows that the deeper the color, the slower the writing speed. In contrast, the lighter the color, the faster the writing speed. According to experimental results, this research finding is fairly effective for broad classification.

Figure 6. Several Instances of Aspect Ratio

3. Feature Classification

In the research field on recognition of handwritten characters, strengths and weaknesses of classification algorithms are critical for determining the value of recognition rate. In these classifications, it is necessary to calculate the similarities between extracted input feature vectors and feature vectors of databases. Based on the similarities, a degree of matching is determined. The symbols with the highest degree of matching will be recognized. At present, promising results have been achieved in recognition by classifiers based on Bayesian theories, which are also fairly effective for recognizing handwritings with simpler letter structures, including English, Arabic and Japanese. More representative classifiers include minimum distance classifiers and MQDF (corrected quadratic discriminant functions) classifiers. When minimum distance classifiers are utilized, the covariance matrix of each category is reckoned as a unit matrix and problems are simplified into mean value estimations. MQDF classifiers are improved QDF classifiers, improving classification efficiency and performances of QDF based on smooth eigenvalues. In this section, these two categories of classifiers will be analyzed at first. Then, their performances for recognizing Uyghur symbols are improved with decision fusion techniques in combination with primitive processing.

3.1. Minimum Distance Classifier and MQDF

Assuming that the input sample to be classified is X, $\omega_1, \omega_2, \cdots, \omega_n$ is n types, prior probability of X that belongs to ω_i is denoted to be $P(\omega_i)$, P(X) is the probability of X, $P(X/\omega_i)$ is a given category, ω_i is the conditional probability of sample X, $P(X/\omega_i)$ is posterior probability for sample X, the Bayes' formula can be conveyed as follows:

$$P\left(\frac{\omega_i}{X}\right) = \frac{P(\omega_i, X)}{P(X)} = P\left(\frac{X}{\omega_i}\right) * P(\omega_i)/P(X) \tag{15}$$

When all posterior probability $P(X/\omega_i)$ is determined, the category of X will be dependent upon the maximum $P(X/\omega_i)$ in case that various risks of misjudgements are the same. In other words, a $P(\omega_i/X_i)$-related decision function $g_i(X)$ can be created. According to the decision rules, when for all $i=1,2,\ldots,n$, $i \neq k$, $g_k(X) \geq g_i(X)$, X will belong to ω_k. $g_i(X)$ can be taken to be $P(\omega_i/X)$, or considered as a function of $P(\omega_i/X)$. When $g_i(X)$ is taken to be $P(\omega_i/X)$, the Bayesian decision will be the minimum probability of error. If $P(X/\omega_i)$ is in line with the following formula (multivariate normal distribution) [1]:

$$P\left(\frac{X}{\omega_i}\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \|\sum_i\|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(X-\mu_i)^T \sum_i^{-1} (X-\mu_i)\right\} \tag{16}$$

Where, μ_i is the mean vector of ω_i and $\sum_i$ is the covariance of ω_i. P(X) is treated as a constant, so the decision function $g_i(X)$ is determined as follows:

$$P\left(\frac{\omega_i}{X}\right) = \frac{P(\omega_i)}{(2\pi)^{\frac{n}{2}} \|\sum_i\|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(X-\mu_i)^T \sum_i^{-1} (X-\mu_i)\right\} \tag{17}$$

(1) Minimum Distance Classifier

Assuming that prior probability $P(\omega_i)$ is the same in $g_i(X)$ and the covariance matrix is the same as a diagonal unit matrix, namely $P(\omega_i) = P(\omega)$ for $i=1,2,\ldots,n$, the decision function will be determined as follows:

$$g_i(X) = -\|X-\mu_i\| \tag{18}$$

In this case, $j=1,2,\ldots,n$, $j \neq i$, when Bayesian recognition methods are adopted. Provided that $\|X-\mu_i\| = min_j \|X-\mu_j\|$, X will be categorized as ω_i.

The minimum distance is determined based on the Euclidean distance between samples and other categories of samples within the characteristic space, so as to judge the exact category of samples. In spite of its general classification capacity, minimum distance classifiers only need to store template vectors of each category. It is possible to complete the

classification just by calculating the Euclidean distance between all vectors, so minimum distance classifier requires very little space for storage and achieves extremely high classification efficiency. To recognize Uyghur symbols, this paper gives priority to these simple and feasible classifiers.

(2) MQDF

For the decision function $g_i(X)$, provided that the prior probability of all categories $P(\omega_i)$ is the same, then:

$$g_i(X) = \frac{1}{(2\pi)^{\frac{n}{2}} \| \sum i \|^{\frac{1}{2}}} exp\left\{ -\frac{1}{2}(X-\mu_i)^T \sum\nolimits_i^{-1} (X-\mu_i) \right\} \tag{19}$$

After transformation, the quadratic discriminant function can be calculated as follows:

$$g(X, \omega_i) = (X-\mu_i)^T \sum\nolimits_i^1 (X-\mu_i) + log \mid \sum_i \mid \tag{20}$$

In fact, QDF, as a measure of distance, is the distance between input vectors and templates. The eigendecomposition of the covariance matrix may be expressed as follows:

$$\sum\nolimits_i = B_i \Lambda_i B_i^T \tag{21}$$

Where, $\sum_i$ is the covariance matrix, Λi is the diagonal matrix, $\sum_i$ composes diagonal elements, and the vectors corresponding to eigenvalues make up the matrix B_i. Then, QDF is corrected based on eigenvalues and feature vectors, and smaller eigenvalues are replaced by constants, so MQDF is obtained [6,12]:

$$\begin{aligned} g_2(X, \omega_i) &= \sum_{j=1}^{k} \left(\frac{1}{\lambda_{ij}} \right) \left[\beta_{ij}^T (X-\mu_i) \right]^2 + \sum_{j=1}^{k} log(\lambda_{ij}) + \sum_{j=k+1}^{d} \left(\frac{1}{\delta_i} \right) \left[\beta_{ij}^T (X-\mu_i) \right]^2 + (d \\ &\quad - k) log(\delta_i) \\ &= \sum_{j=1}^{k} \left(\frac{1}{\lambda_{ij}} \right) \left[\beta_{ij}^T (X-\mu_i) \right]^2 + \sum_{j=1}^{k} log(\lambda_{ij}) + \left(\frac{1}{\delta_i} \right) r_i(X) + (d-k) \end{aligned} \tag{22}$$

Where, k is the number of the main axis, X is the mapping of the secondary axis, which is conveyed as $r_i(X)$. In view that Euclidean distance is unchanged, it is solved that:

$$r_i(X) = \|X-\mu_i\|^2 - \sum_{j=1}^{k} \left[\beta_{ij}^T (X-\mu_i) \right]^2 \tag{23}$$

δ_i, as a constant of MQDF, may be calculated by the following formula:

$$\delta_i = \frac{tr(\sum_i) - \sum_{j=1}^{k} \lambda_{ij}}{d-k} = \sum\nolimits_{j=k+1}^{d} \lambda_{ij} / (d-k) \tag{24}$$

Where, $tr(\sum_i)$ is the value of covariance determinant, δ_i is the mean of eigenvalues, the eigenvalue is determined on average, excluding the first k eigenvalues.

MQDF classifier can eliminate the errors for estimating relatively small eigenvalues. The storage and computation load can be reduced simply by storing the first k feature vectors, eigenvalues and the mapping of X on the main axis.

3.2. Feature and Decision Fusions

Although there are many effective recognition and classification methods, all feature selection and classification methods have their respective strengths and weaknesses. It is extremely difficult to realize ideal classification just by a single classifier. To improve ultimate recognition effects, multi-feature and multi-classifier fusions have become a critically important research interest of present pattern recognition. Various features and classification methods are usually complementary to each other to certain extent, so recognition efficiency can be improved by certain fusion technologies. Currently, feature fusion and decision fusion are two common fusion technologies.

In general, a feature extraction technology can extract some features of samples, but can't manifest sample characteristics and highlight some discriminant features. Meanwhile, certain noise and redundancy are caused. Hence, multiple kinds of features are fused, adjusted and optimized to highlight effective features, reduce or eliminate noise and redundancy, in order to effectively improve recognition performances. Figure 7 shows a recognition framework based on feature fusion. Its key consists in optimizing identified information via feature fusion.

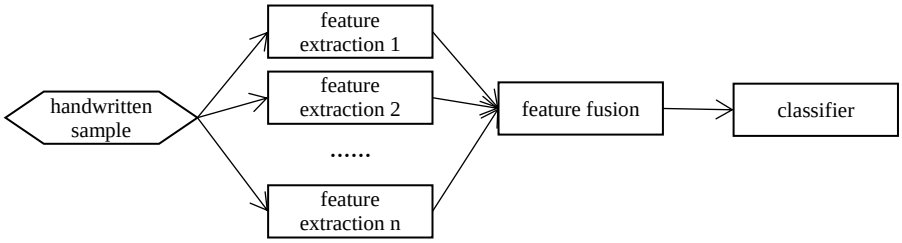

Figure 7. Feature Fusion Procedures

There are mainly two feature fusion techniques, namely serial and parallel fusion. Serial fusion means serial connection of several kinds of features for generating a higher-dimension feature vector, recognizing and classifying the features in that high-dimension vector space. By parallel classification methods, two or more kinds of features are integrated to make up a complex vector, on which features are recognized and classified. To a certain extent, both feature fusion technologies can improve classification and recognition performances [10].

Classifier integration is a decision fusion technique. Like sample characteristics, classifiers are somewhat independent and complementary. Classifier integration means integrating different classifiers to make up a higher-level recognition system for the final purpose of improving overall performances. In classifier fusion technologies, feature selection and classification methods vary among different parallel layers. To integrate results of different layers, decision fusion is necessary. Then, the results can be rearranged in order to improve final recognition performances. The recognition framework based on classifier fusion is shown in Figure 8. In this framework, it is essential to optimize results of different layers through decision fusion.

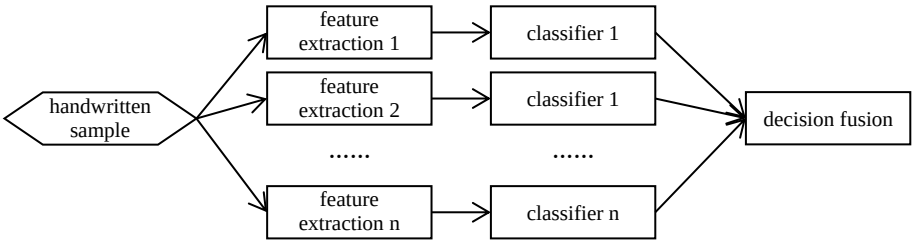

Figure 8. Feature Fusion Procedures

Compared with single classifiers, classifier integration technologies have evident strengths. All recognizers may make different contributions to the final conclusions. In other words, the feature vectors that can't be identified by some recognizers would be easily recognized by some other recognizers. Although some recognizers are unhelpful for judging the category of samples, they are likely to easily eliminate some types, so as to greatly reduce wrong classifications while indirectly enhancing recognition and classification.

This paper proposes a confidence-based decision fusion method. Through parallel connection of classifiers, the recognition and classification conclusions on each layer are converted into values of confidence, based on which the candidate results can be rearranged in order. The confidence is calculated as follows:

Assuming that the probability of categories $P(\omega_i)$ is uniformly distributed, the posteriori probability $P(^{\omega_i}/_X)$ may be conveyed as follows [7,11]:

$$P(\omega_i|X) = \frac{p(X|\omega_i)}{\sum_j P(X|\omega_i)} = \frac{\frac{1}{|\sum|_i^{1/2}} exp\left\{\frac{(X-\mu_i)^T \sum_i^{-1}(X-\mu_i)}{-2}\right\}}{\sum_j \frac{1}{|\sum|_j^{1/2}} exp\left\{\frac{(X-\mu_j)^T \sum_j^{-1}(X-\mu_j)}{-2}\right\}} \tag{25}$$

The recognized distance calculated by MQDF (d_i) with MQDF is determined as follows:

$$d_i = (X-\mu_i)^T \sum\nolimits_i^{-1} (X-\mu_i) + log\,|\sum_i| \tag{26}$$

Besides, $p(X|\omega_i)$ is proportional to $\exp\{-d_i/2\}$,

$$p(X|\omega_i) \propto exp\{-d_i/2\} \tag{27}$$

so posteriori probability $P(\omega_i|X)$ is presented as follows:

$$P(\omega_i|X) = \frac{p(X|\omega_i)}{\sum_j P(X|\omega_j)} = \frac{exp\{-d_i/2\}}{\sum_j exp\{-d_j/2\}} \tag{28}$$

Where, d_i, i=1, …,k, which denotes the value of distance among the first k candidate classifications.

In consideration that the numerical value of distance (d_i) is much higher in some cases, it will be zero if it is taken to be negative, and the operations will be inconvenient, so it will be adjusted at certain scale. That is:

$$d_i' = d_i/D_0 \tag{29}$$

By substituting d_i into the above formula, the confidence will be as follows:

$$P(\omega_i|X) = \frac{exp\left(\frac{d_i/D_0}{-2}\right)}{\sum_j exp\left(\frac{d_j/D_0}{-2}\right)} \tag{30}$$

Where, D_0 is a scale factor, and its value shall be estimated based on the distance between recognized samples. d_{min} and d_{max} are statistically analyzed for a fixed amount of samples. They respectively represent the minimum and maximum distance between the first and the final candidate characters. Empirically, the following formula is tenable [8,9]:

$$\begin{cases} p(Final\ Candidate|X) \approx 0 \\ p(First\ Candidate|X) \geq 0.5 \end{cases} \tag{31}$$

To make it convenient for calculating and adjusting parameters, D_0 is taken to be 2^N. Then, this scale factor (D_0) may be adjusted as follows:

$$\begin{cases} exp\{\frac{d_{max}/2^N}{-2}\} \leq 10^{-10} \\ \frac{exp\{\frac{d_{min}/2^N}{-2}\}}{\sum_{i=0}^{N-1} exp\{\frac{[d_{min}+i*(d_{max}-d_{min})/(N-1)]/2^N}{-2}\}} \geq 0.5 \end{cases} \tag{32}$$

Where, N is the number of candidate types (In other words, the first N candidates are selected), d_{min} and d_{max} represent the maximum and minimum distance between the first and the last candidate words respectively.

By above methods, the confidence of decision fusion can be solved. It is suppose to range between 0 and 1. By repeating this process, the confidence of results can be determined on each layer. Subsequently, results can be rearranged in order to reach final conclusions.

4. Analysis of Performances for Recognizing Uyghur Symbols

In this experiment, performances of recognition algorithms for Uyghur symbols are tested. They are experimentally tested on subsets in corresponding sample libraries of letters and connected segments. Two test sets, namely Test_c1 and Test_str1, are selected. To be specific, Test_c1 is a subset in the sample library of Uyghur words, while Test_ str 1 is a subset in the sample library of connected segments in Uyghur. 10% samples are randomly selected to compose the test set, whereas the remained samples are integrated into a training set. In this paper, the supervision set is not taken into account owing to limited research progress).

In comparative experiments, minimum distance classifiers, MDQF classifiers and HMM (DTW)-based primitive recognition are adopted as classifiers, which are connected in parallel. The experimental results are shown in Figure 9 as follows.

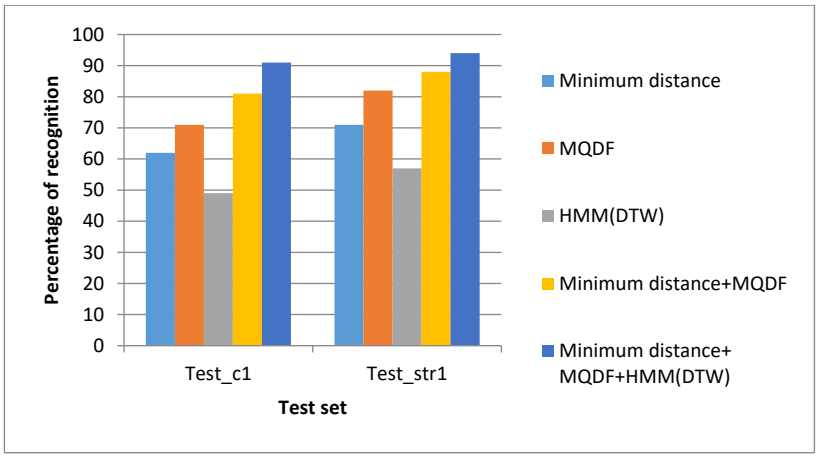

Figure 9. Recognition Effects of Classifiers

According to the experimental results,

(1) MQDF, as classifier, has better performances than the minimum distance classifier. However, its weaknesses lie in its occupation of relatively huge storage space and complicated calculations.

(2) The recognition rate of MQDF will be higher if LDA dimensions are increased.

(3) The recognition and classification effects are extremely poor when premiums are separately used. Nevertheless, it is effective to correct mistakes of other classifiers.

(4) Although classifier fusion can effectively increase recognition accuracy, the systems will become more complicated in terms of time and space.

(5) Notwithstanding high feature extraction rate and recognition rate, whole-word recognition is disadvantageous in a large vocabulary. In practical writing practices, connected segments are usually input as basic units. It is hard to cope with this situation. Despite that, many improvements remain to be further made in the final results, no better solution has been found except for integrating segmentation with whole-word recognition.

5. Conclusions

In consideration that Uyghur symbols are poorly recognized and recognition algorithms are impertinent for connecting characters, five kinds of features are extracted from Uyghur symbols after feature analysis and pre-processing, including 8-directional features, fuzzy features and primitive features. According to characteristics of Uyghur, features of Uyghur symbols such as aspect ratio and dynamic speed are extracted. Features are selected by LDA transformation, and the methods for judging relationships among Uyghur primitives are determined. Minimum distance classifier and MQDF classifier are suggested to be used. By recognizing primitives, effectiveness of algorithms is experimentally verified with online Uyghur recognition algorithms of decision fusion strategies by classifiers.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.61462088), Scientific Research Program of the Higher Education Institution of XinJiang (FSRPHEXJ), Initial Scientific Research Fund of Doctors in Xinjiang Normal University, The Key Discipline of Computer Application of Xinjiang Normal University and the Key Laboratory of Data Security of Xinjiang Normal University.

References

1. K. Anil, P. Jain, W. Robert, M. Jianchang, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 4-37, 2000
2. Y. Assabie, J. Biqun, "Offline Handwritten Amharic Word Recognition", *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1089-1099, 2012
3. H. E. Abed, V. Margner, "Arabic handwriting recognition competition", *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, pp. 3-13, 2011
4. M. Kherallah, "On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm", *Engineering Applications of Artifical Intelligence*, vol. 22, no.1, pp. 226-236, 2013.
5. S. Mozaffari, K. Faez, V. Maergner, "Two-stage lexicon reduction for offline Arabic handwritten word recognition", *International Journal of Pattern Recognition andArtificial Intelligence*, vol. 22, no.7, pp. 1323-1341, 2008
6. A. Rehman, T. Saba, "Off-line cursive script recognition: current advances, comparisons and remaining problems," *Artificial Intelligence Review*, vol. 37, no.4, pp.261-88, 2012
7. M. I. Razzak, F Anwar, S. A. Husain, "HMM and fuzzy Logic: A Hybrid Approach for Online Urdu Script-based Languages' Character Recognition", *Knowledge-Based Systems*, vol. 23, no.8, pp. 914-923, 2010
8. F. Slimane, S. Kanoun, J. Hennebert, "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution", *Pattern Recognition Letters*, vol. 34, no. 2, pp. 209-218, 2013
9. T. Saba, A. Rehman, M Elarbi-Boudihir, "Methods and strategies on off-line cursive touched characters segmentation: a directional review", *Artificial Intelligence Review*, vol. 42, no. 4, pp. 1-20, 2014
10. T. Salimi, P. Hamid, S. Giveki, A. Davar, "Arabic handwritten digit recognition based on ensemble of SVD classifiers and reliable multi-phase PSO combination rule", *International Journal on Document Analysis and Recognition*, vol. 16, no. 4, pp. 371-386, 2013
11. M. T. Parvez, A Sabri. P Mahmoud, "Offline arabic handwritten text recognition", *ACM Computing Surveys*, vol. 45, no. 2, pp. 1-35, 2013
12. N. Tagougui, M Kherallah, A. M. Alimi, "Online Arabic handwriting recognition: a survey", *International Journal on Document Analysis and Recognition*, vol. 16, no. 3, pp. 209-226, 2013

Xiangwei Qi received his Master's degree in Xinjiang Normal University. His current research interests include pattern recognition, machine learning, and intelligent information processing.

Yong Yang received his Master and doctor degrees in Xinjiang Institute of physics and chemistry, Chinese Academy of Sciences. His research interests include machine learning and intelligent information processing.

Weimin Pan received his Master's degree in Xinjiang University. Her research interests include computer networks and security, machine learning, and intelligent information processing.