Fuzzy Stroke Type Identification for On-Line Chinese Character Recognition

Jyh-Yeong Chang and Min-Hwa WAN
Department of Electrical and Control Engineering
National Chiao Tung University, Taiwan, R.O.C.

Abstract– This paper presents an on-line Chinese character stroke type recognition system based on fuzzy set theory. According to the writing stroke sequence, each character is described by an 1-D stroke string model. The input written characters can be loosely constraints, which are quite flexible on size and allow various variations. The stroke segments of input strokes are extracted firstly and the strokes of input character are then identified as a sequence of primitive stroke types by a fuzzy methodology. A character recognition system, based on this stroke type identification scheme and Modified Dynamic Programming Forward (MDPF) matching, Modified Dynamic Programming Backward (MPDB) matching, and A^* matching algorithms, has been built and tested to be very successful.

I. INTRODUCTION

On-line handwritten Chinese character recognition (OLCCR) is one of most natural and efficient ways of high speed text input to the machine. Stroke-based analysis is one of the most traditional approaches in on-line Chinese character recognition which utilizes line segments of the strokes as the feature to describe the construction of Chinese characters. A set of basic strokes is usually selected as the primitives and there are about 29 stable primitive stroke types defined for the construction of the Chinese character on-line model [1], [2]. For stroke-based approach, the character is decomposed into strokes and the types of the strokes have to be recognized. A string of stroke type sequence according to the writing order can therefore be obtained. Then we can use the recognized stroke string to look up the reference dictionary for finding the proper class of the input character. Therefore, stroke and stroke type identification modules play the most important role in a stroke-based character recognition system.

In recent years, fuzzy set has provided a powerful scheme of knowledge representation with its clear-cut logical properties [3]. In particular, fuzzy set theory has been an appropriate framework to address many of the problems encountered in handwriting recognition, because various kinds of variation found in handwritten characters can be resolved by fuzzy logic [4], [5] and consequently the constraints of careful writing can be greatly released. Because the primitive strokes constituting the Chinese character are fuzzy in nature, fuzzy membership function will be constructed and then is used as a similarity measure basis of primitive strokes. This paper presents a fuzzy set approach for identifying the stroke type of handwritten Chinese characters. Instead of using direction codes to represent a stroke segment and then identify the input stroke by performing the feature matching between the given segment string and section string of the input stroke, we extract the angle of a stroke segment directly. For each stroke type, we define the angle membership functions as features to measure the degree of similarity for input strokes. An input character then can be recognized as a sequence of primitive strokes, called stroke string. Based on this stroke type identification scheme, Modified Dynamic Programming Forward (MDPF) matching, Modified Dynamic Programming Backward (MDPB) matching, and A^* matching algorithms, are used to construct a character recognition system to best match the input script to the reference stroke database. Finally, the results of these three matching algorithms are fused with fuzzy integral technique. The character with the maximum confidence value is then recognized as the target character.

II. PRIMITIVE STROKE TYPE IDENTIFICATION

From the analysis of [1], [2], 29 types of primitive strokes defined by primitive codes as shown in Fig. 1 are enough to construct a Chinese character. Because each primitive stroke type is composed of several stroke segments, the stroke segments of an input stroke will be extracted first to identify the primitive stroke type.

A. STROKE SEGMENT EXTRACTION

The directions of line segments for most Chinese characters are composed of horizontal, vertical, and diagonal lines. Thus we define four basic types of segments, which are called primitive segments in this paper, to extract the stroke segments from an input stroke. These four types of primitive segments, (1) horizontal segment (→), (2) ver-

0-7803-4778-1/98 $10.00 © 1998 IEEE

tical segment (↓), (3) right-slanting segment (↙), and (4) left-slanting segment (↘), are shown in Fig. 2(a).

A primitive segment is a line segment with roughly the same gradient, i.e., the gradient of a primitive segment is permitted to vary, but within a prespecified bound. By allowing a range of tolerance, this classification can tolerate angle differences of the line segments in handwriting a character. As shown in Fig. 2(b), a line segment with slanting angle lying between $-\alpha$ and α is defined to be the horizontal segment. A line segment with slanting angle lying between $\frac{\pi}{2}+\beta$ and $\frac{\pi}{2}-\beta$ is defined to be vertical segment. In the same way, line segments with slanting angles lying between α and $\frac{\pi}{2}-\beta$ and between $-\frac{\pi}{2}+\beta$ and $-\alpha$ are defined to be the right-slanting segment and the left-slanting segment, respectively. Here the α and β are both chosen to be 10°. To this end, let l be the slope of the line segment and tp be the type of the primitive segment, the slope bounds of these four primitive segment types are given as follows:

If	$-0.176 \leq l \leq 0.176$	then	tp is 1,
else if	$l \leq -5.67$ *or* $l \geq 5.67$	then	tp is 2,
else if	$0.176 \leq l \leq 5.67$	then	tp is 3,
else			tp is 4.

A stroke, in accordance with the stylus pen output of a script may consist of more than one primitive segment. In our stroke segment extraction method, we extract a sequence of primitive segments from the input stroke. Every time when the distance of the input script of a character reaches a length of 40 under the sampling rate of 9600 bps, this short line segment is assigned by a primitive segment code as defined above and thus a stroke segment is represented by a sequence of four primitive segments as described above. For example, the first stroke segment "⌐" of the stroke "7" consists of five consecutive primitive code "1" as shown in Fig. 3. A stroke may contain another line segments with different gradients. The second stroke segment "/" of the stroke "7" consists of six consecutive primitive code "3" as shown in Fig. 3. Since a stroke segment is represented by a sequence of same certain primitive code, these primitive segments are somewhat colinear. Thus the stroke of Fig. 3, with primitive segments with same primitive segment code being combined, is represented by two stroke segments as "1 3" only, as shown in Fig. 3.

After extracting the stroke segments for the input stroke, we can compute the angle θ_i, which represents the arguments of the stroke segment vector, corresponding to the sequence of stroke segment vectors. The range of the angle θ_i is defined in $[0°, 360°]$. Fig. 4. shows an example for computing the angles of a stroke "7".

Note that the extracted stroke segments are sometimes not necessarily accurate owing to the input stroke may be wavy and/or hooked. Thus to obtain a reliable stroke segments, a preprocess that can delete the noisy or redundant ones from the original segments is presented. It is evident from Fig. 1 that the angle difference between any two consecutive stroke segments is at least 45°. Thus after computing the corresponding angle for each stroke segment of the input stroke, if the angle difference of any two consecutive stroke segments is less than a preset value, 30° in this paper, the one with the smaller length of stroke segment is deleted. For example, as shown in Fig. 5, stroke "7," which leads to a stroke segment "4 1 3," would conform to stroke"7." Due to the angle between stroke segments "4" and "1" is just 20°, and the length of stroke segment "4" is small, so we delete the redundant stroke segment "4." After deleting redundant segments, the reliable stroke segments are extracted from the input stroke of a script. The angles of all the stroke segments will be used as the feature vector for primitive stroke type identification.

B. Fuzzy Stroke Type Identification

Two strokes are of the same type if the angles of stroke segments they form are similar enough. That is, the possibility of two strokes to be of the same type is inversely proportional to the angle difference of stroke segments. To perform the stroke identification for the input stroke, the angles of the stroke segments of the primitive strokes are extracted first as shown in the previous subsection. The feature angles of training scripts of characters will be used to construct the fuzzy membership function for each stroke segment of a primitive stroke type.

On observing these 29 primitive stroke types from Fig. 1, it is easily seen that the maximum stroke segments in a primitive stroke type is five. Thus, for each primitive stroke, its feature vector is denoted as $\theta_f = (\theta_{f1}, \theta_{f2}, \theta_{f3}, \theta_{f4}, \theta_{f5})$, consists of direction angles of each corresponding stroke segment and the feature angle θ_{fi} is the angles of the i-th stroke line segment for a stroke type input in the training set. The don't care components for each primitive stroke type are assigned with stroke segments less than five and filled with the pseudo number $-180°$ in this paper.

Now we will fuzzify the input stroke type feature vector by a membership function to accommodate the ambiguity caused from the various variations among the angles of the stroke segments of a script. According to the fuzzy set definition, the membership function describes the possibility over the universe of discourse. We fuzzify each component of the feature vectors of the 29 stroke types, one for each type by a trapezoidal fuzzy membership function which has degree unity around θ_{fi} and a decreasing degree apart from θ_{fi}. As shown in Fig. 6, for a stroke segment of a primitive stroke F, the membership function

could be described by its corresponding feature angle θ_{fi}. The term d represents the length which the membership grades are equal to 1. In this paper, the term d is specified by the variation range of the angles of corresponding stroke segment for all the same stroke type inputs in the training scripts. The variable λ represents the fuzzification coefficient which, in fact, determines the slope of the membership function. In this paper, the variable λ is chosen as $\frac{1}{180-d/2}$.

Membership grades $\mu_F^i(\theta_i)$ express the possibility of crossing at the feature angle θ_{fi} for an input stroke segment with angle θ_i. In this manner, each stroke line segment of the input stroke now has membership grades, which represents the crossing fuzziness at the direction θ_{fi}, by a value in the interval $[0, 1]$. Therefore, a five membership grade vector can then be generated for each input stroke type. These five fuzzy sets form the reference vector for each primitive stroke type to classify the type of a given hand-written stroke type.

For a given input stroke, determine the $\boldsymbol{\theta}$=(θ_1, $\theta_2, \theta_3, \theta_4, \theta_5$) for this stroke. Then the following average operator is used to determine the similarity degree of the given stroke to each primitive stroke: the degree to which a given $\boldsymbol{\theta}$ belongs to primitive stroke type F equals the average of the degrees of all the five components according to the corresponding membership functions constructed above, i.e.,

$$\mu_F(\boldsymbol{\theta}) = \frac{1}{5}\sum_{i=1}^{5} \mu_F^i(\theta_i). \tag{1}$$

For the convenience of computing the similarity degree and noting that the membership function is symmetrical at θ_{fi}, Eq. (1) can be rewritten as

$$\mu_F(\boldsymbol{\theta}) = \frac{1}{5}\sum_{I=1}^{5} \mu_F^i(\theta_I), \tag{2}$$

where

$$\theta_I = min(|\theta_{fi} - \theta_i|, 360° - |\theta_{fi} - \theta_i|).$$

The membership functions of θ_i and θ_I are shown in Figs. 6 and 7, respectively.

We use this average operator because it is the aggregated property of all the five components that specifies the type of the input stroke. Fig. 8 shows an example of determining the similarity degree between the input stroke "  " and the primitive stroke "  ." The stroke segments of the input stroke "" are extracted first shown in Fig. 8(c). The corresponding angles for each stroke segments, denoted as θ_i, are then computed and shown in Fig. 9. Finally, from Eq. (2), the $\boldsymbol{\theta}$ leads to a similarity degree of 1 between the stroke "" and "."

Each input stroke will be compared to each primitive stroke types and obtain 29 membership grades, which correspond to 29 primitive strokes types. The maximal value of these similarity degrees corresponds to the best match. Maximum rule is proposed here to classify the given hand-written stroke type, i.e.,

$$m_z = \max_F[\mu_F(\boldsymbol{\theta})] \qquad F = 0, \cdots, 28. \tag{3}$$

Then, the input stroke defined by θ_i is classified as stroke type F corresponding to m_z.

After performing the fuzzy stroke type identification, the character is represented by a sequence of primitive stroke types, which will be called stroke string. Fig. 10 shows an example of the character "" which is recognized as a sequence of primitive strokes string "2, 14, 2, 14, 15" by its writing stroke sequence "", "", "", "", "". Some more examples of test characters which are recognized as a sequence of primitive stroke string are shown in Fig. 11.

III. Character Recognition by Stoke-string Matching

A Chinese character is composed of a collection of primitive stroke types. Hence the stroke type sequence can be used as an on-line model for recognizing the Chinese characters. Since the Chinese character database is large (about 5401 characters in daily use), preliminary classification is required in order to improve the performance and reduce the time consuming. The pre-classification process reduces the number of candidates by taking the number of strokes and top/bottom level radical of the input character into account. From the analysis of our character database, stroke-number variation of input pattern comparing to corresponding reference patterns are almost in the range $[-2, +1]$. Hence, those reference patterns with stroke numbers from $n-2$ to $n+1$ are selected to match input pattern with stroke number n. Furthermore, the writing sequence of a character may vary from person to person but the initial and last writing sequences of the radicals in each Chinese character are rather stable features that can be utilized for classification. The radicals of input character are extracted by traversing the initial writing sequence of input character forward and last writing sequence backward to compare top/bottom radical database. If the candidate characters satisfy these three conditions, they can be sent into the next stage for further matching.

The effective matching schemes must be chosen carefully to develop a good recognition system. For stroke-number and stroke-order free system, it is advised to use nonlinear matching approach. Here, in this paper, we used three modified searching schemes as the sub-recognizers. They are Modified Dynamic Programming Forward matching (MDPF), Modified Dynamic Programming Backward matching (MDPB) [1], [6], and A^* [7], [8]

searching approaches. We use the fuzzy integral to combine the results of three sources of matching algorithms [9]. The fuzzy integral can be thought as searching for the maximal grade of agreement between the objective evidence and subjective expectation of prediction. The objective evidence is obtained from the recognition results of the three matching algorithms. As to the subjective prediction, fuzzy density, are assigned based on how these algorithms performed on validation data. The fuzzy density of each matching algorithm, the importance of each classifier, can not be fixed to one value and used through the huge number of character classes. For this reason, we develop the system in which the densities are character and matching algorithm dependent. The fuzzy densities are calculated beforehand and stored in database. After the input character is processed by the three matching algorithms individually, the outputs of three matching algorithms are then merged by the fuzzy integral fusion process [9].

IV. The Simulation Results

To investigate the feasibility of the recognition system proposed in this paper, experiments are performed for the recognition of 224 characters written by 4 persons in our laboratory and 3630 characters selected randomly from the database HCCR of ITRI, Taiwan [9]. The database, which contains 5401 classes of common Chinese characters that are written by more than 300 persons. There are about 21 variations per class.

The recognition system is implemented on PC (PENTIUM-166) in Turbo C 3.0++ and the input device is OmniPen KD1000 digitizer. Experiments have been done to examine the performance of the approach. Performance of the proposed system as well as the three individual classifiers were evaluated and compared. The results listed in Table I shows that the proposed system incorporating evidence fusion technique is better than the traditional character recognizer. The recognition rate, which is the results of recognizing the 3854 characters by the proposed system, is 95.02%. The cumulative classification rate of choosing top three most similar characters is up to 97.2%.

V. Conclusion

In this paper, we have proposed an effective stroke type extraction method for Chinese characters using fuzzy stroke concept. The stroke-based recognition system has been considered as a promising direction for Chinese character recognition, and stroke identification plays the most important part of the recognition system. Our fuzzy stroke type identification scheme has been implemented and then used for the matching stage of the Chinese recognition system. We have obtained a very high recognition rate of 97.2%.

References

[1] C. K. Lin *et al.*, "A knowledge model based on-line recognition system," *IEEE Int. Conf. on ASSP*, vol. 3, pp. 157-160, 1992.

[2] K. S. Chou *et al.*, "An on-line Chinese character recognition system," *Int. Conf. on CPCOL*, pp. 149-153, 1991.

[3] H. -J. Zimmermann, *Fuzzy Set Theory and Its Application.* Boston: Kluwer Academic, 1991.

[4] F. H. Cheng, W. H. Hsu, and C. A. Chen, "Fuzzy approach to solve the recognition problem of handwritten Chinese characters," *Patt. Recog.*, vol. 22, no. 2, pp. 133-141, 1989.

[5] H. M. Lee and C. W. Huang, "Fuzzy feature extraction on handwritten Chinese characters," *Proc. FUZZ-IEEE*, vol. 3, pp. 1809-1814, 1994.

[6] C. K. Lin, K. C. Fan, and F. T. P. Lee, "On-line recognition by deviation-expansion model and dynamic programming matching," *Patt. Recog.*, vol. 26, no. 2, pp. 259-268, 1993.

[7] K. S. Chou *et al.*, "Knowledge model based approach in recognition of on-line Chinese charactersine," *IEEE Journal on Selected Area in Communications*, vol. 12, no. 9, pp. 1566-1575, 1994.

[8] Q. Z. Wu *et al.*, "A new stroke string matching algorithm for stroke-based on-line character recognition," *IEEE Int. Conf. on ASSP*, vol. 1, pp. 645-648, 1993.

[9] R. F. Liao, "On-line character recognition system using fuzzy integral," *Master Thesis, Inst. of Control Eng., Chiao Tung University*, Taiwan, 1994.

Stroke Type	Stroke Code	Stroke Type	Stroke Code
→	0		15
↓	1		16
↙	2		17
↘	3		18
↗	4		19
	5		20
	6		21
	7		22
	8		23
	9		24
	10		25
	11		26
	12		27
	13		28
	14		

Fig. 1. The primitive stroke types and their corresponding stroke codes.

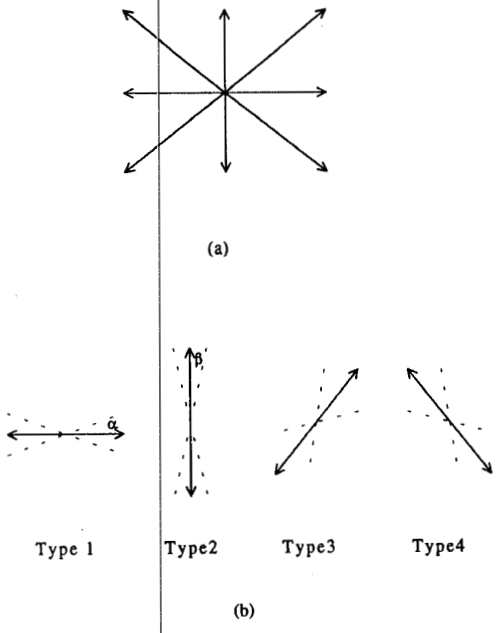


Fig. 2. (a) The directions selected for most Chinese characters. (b) Four primitive segments and their tolerance ranges.

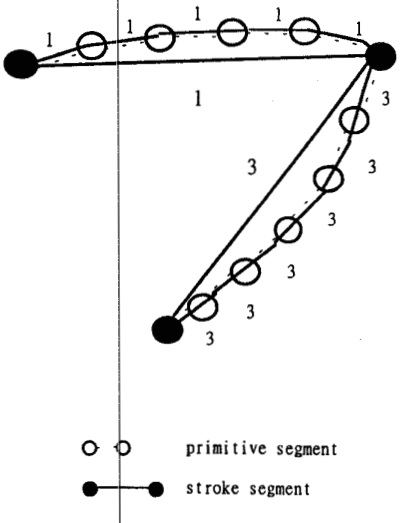


Fig. 3. The stroke input and its primitive segment "1 1 1 1 1 3 3 3 3 3 3." After line approximation, the input stroke is composed of two stroke segments, "1" and "3."

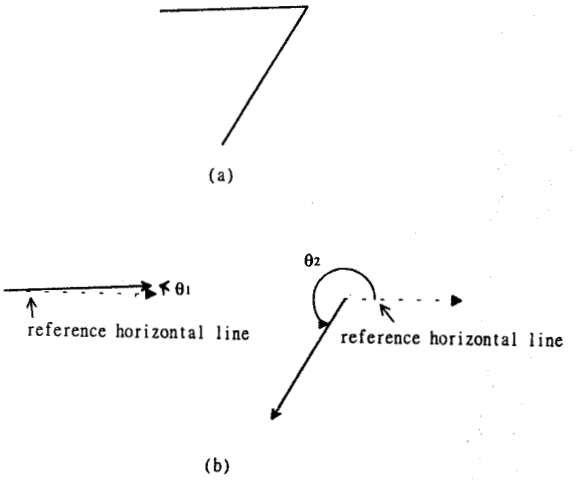


Fig. 4. (a) The stroke segments of the primitive stroke type "7." (b) The angles of the stroke segments.

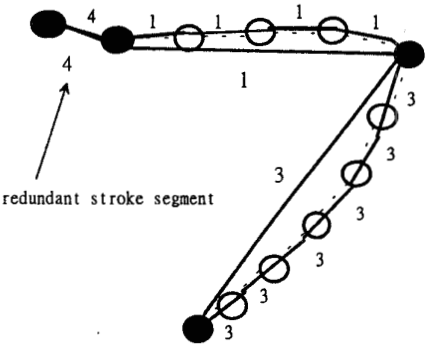


Fig. 5. The stroke "7" would be conformed to stroke "7" after a preprocessing.

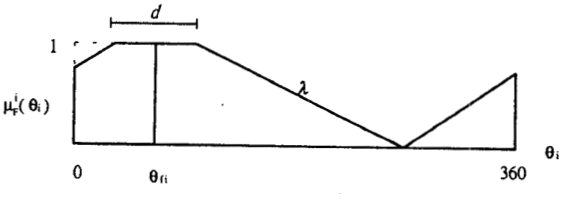


Fig. 6. The angle membership function of a stroke segment with the feature angle.

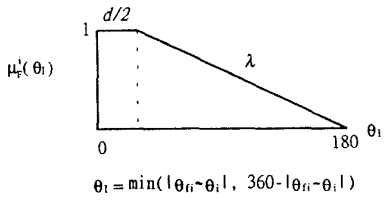


Fig. 7. Simpler form of the membership function of Fig. 6.

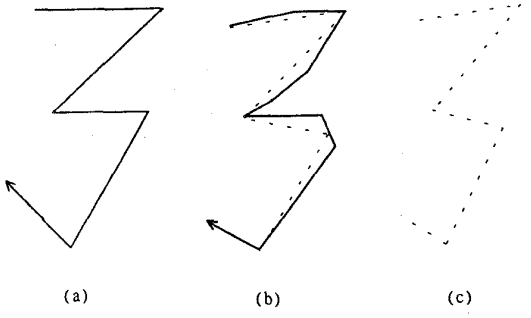
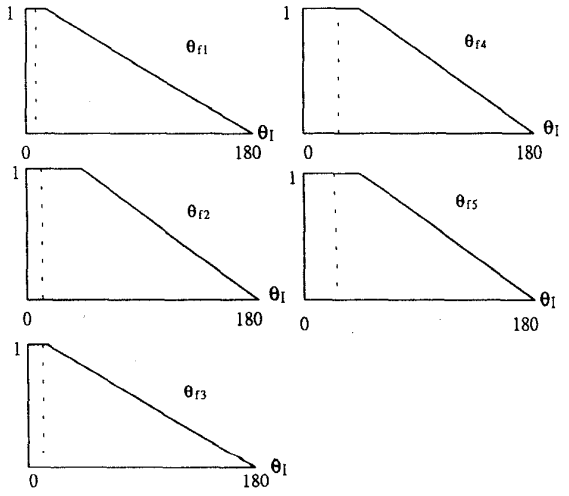


Fig. 8. (a) The primitive stroke type 28. (b) A test stroke. (c) The extracted stroke segments.



$\theta_f = (0, 255, 0, 225, 135)$

$\theta = (7.3, 237.2, 16.8, 253.3, 163.7)$

$\theta_I = (7.3, 12.2, 16.8, 28.3, 28.7)$

$\mu_{28}(\theta_I) = 1$

Fig. 9. The five angle membership functions of the feature angle. The similarity degree between the Fig. 7(a) and (c) is 1.

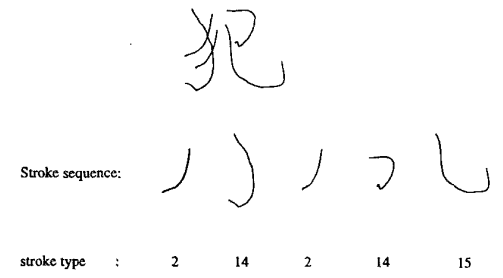


Fig. 10. The character "犯" which is recognized as a sequence of primitive stroke string "2, 14, 2, 14, 15."

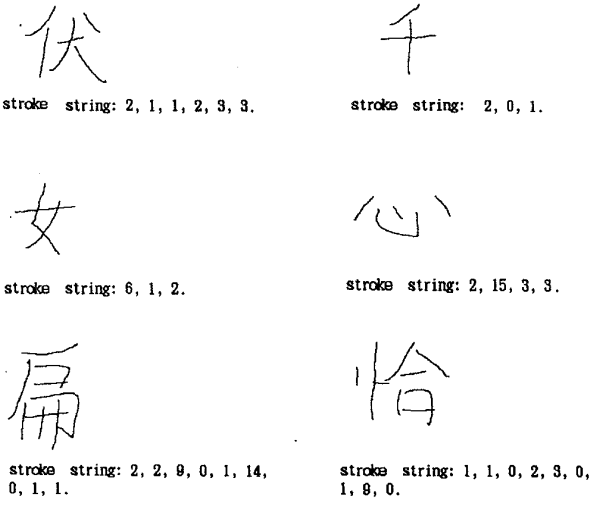


Fig. 11. Some examples of test characters which are recognized as sequence of primitive stroke string.

Table I. The results of recognizing the 3854 characters by the proposed system

Classifier	Test sample	Error	Error rate(%)	Recognition rate(%)
MDPF	3854	285	7.39	92.61
MDPB	3854	252	6.54	93.56
A^*	3854	302	7.84	92.16
Fusion of MDPF, MDPB, and A^*	3854	192	4.98	95.02