Open Journal of Applied Sciences, 2012, 2, 184-187
doi:10.4236/ojapps.2012.23027 Published Online September 2012 (http://www.SciRP.org/journal/ojapps)

A New Method for Chinese Character Strokes Recognition

Yan Xu, Xiangnian Huang, Huan Chen, Huizhu Jiang
Xihua University, Chengdu, China
Email: 501625379@qq.com

Received July 6, 2012; revised August 4, 2012; accepted August 16, 2012

ABSTRACT

In this paper, the problem of stroke recognition has been studied, and the strategies and the algorithms related to the problem are proposed or developed. Based on studying some current methods for Chinese characters strokes recognition, a new method called combining trial is presented. The analysis and results of experiments showed that the method has the advantage of high degree of steadiness.

Keywords: Chinese Character Recognition; Stroke Recognition; Stroke Combining

1. Introduction

Unconstrained handwritten Chinese character recognition has been a difficult research area of character recognition, and its stroke recognition is an important part of structural analysis of Chinese characters. The stroke is constituted by a number of direction strokes(referred to as strokes). Studies have shown that despite the everchanging of unconstrained handwritten Chinese character, stroke is very stable. Strokes contained in stroke belong to the same section of a natural stroke (a written), and generally, they are not separated in two or more natural strokes, and this paper uses this feature to complete stroke recognition. Although the definition of stroke varied, a common feature of stroke is that it can be divided into two categories of the basic stroke and composite stroke. Basic stroke refers to the direction strokes, this segment of space partition is shown in **Figure 1**. Strokes are easy to determine and extract algorithm, and this article discusses the complex stroke, posed by multiple strokes [1]. For the Chinese character recognition system based on structural analysis, effects of stroke recognition have a major impact on the whole word recognition. Those more typical stroke recognition methods have direct synthesis method, dynamic programming method and so on. Based on the analysis of these methods, this paper presents a new method of Chinese characters called combining trial [2].

2. Direct Synthesis

Direct synthesis method of stroke, which uses the extracted HSPN four direction strokes, directly synthesize strokes according to the definition of given stroke. Algorithm idea: set some determine conditions in the stroke identification module ,combine the extracted strokes serial number string in the same natural stroke, then compose them to different types of strokes according to preconditions to complete the stroke recognition. The biggest feature of this algorithm is simple and quick. Considering a greater deformation of unconstrained handwritten Chinese character, the algorithm is generally suitable for the occasion to write a more standardized, the high limit handwritten of more restrictive writings and printed texts.

3. Dynamic Programming

This algorithm is proposed to overcome adverse effects of the deformation of the stroke to stroke judgment. It uses a similar matching strategy: first, define a standard strokes code string (standard stroke) and the distance calculation rules between strings; then, take matching operations between the unknown code string and all of the defined standard code strings one by one, calculate the distance between them ;finally, take the smallest as matching result.

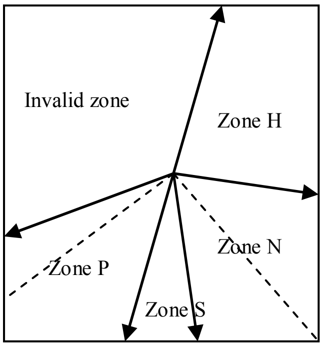


Figure 1. Space division of strokes types.

Algorithm for real: Through the gradual increasing or reducing the length of a code string such as X, search the code string in the standard code string set which has the minimum distance with the unknown code string as the results, and transform the stroke matching problem into solving the best matching problem between two symbol strings.

There are a lot of attached strokes through the free writing, and strokes string to be identified is not known in advance, then we need to take stroke splitting, and each splitting produces a series of strokes to be knowledge, so this method has multi-step iterative and volume computation. We recognize that: 1) Long stroke is more steady than short stroke; 2) Different types of strokes have different possibility of confusion. Focusing on these two differences, this paper has proposed to take "fuzzy numbers of length" and "fuzzy numbers of direction" treatment strategy. See the following section.

4. Combining Trial

Because there are a large number of connected strokes that is multiple strokes connected with unconstrained handwritten Chinese characters (belong to the same natural strokes), and we can not ask the writer to separate a painting to each stroke, stroke recognition includes two processes which are stroke splitting and stroke judgment, requiring to make a split decision from the entire word which comes from an unknown mode and then make a decision. These two processes complement each other, alternately. So, we propose the combining trial to recognize stroke.

The recognition process of combining trial is a tentative process of splitting and judgment constantly:

Split→Judge→Split again→Judge again …

Input: the direction strokes code sequence of a natural stroke (string);

Output: the stroke code sequence of the natural stroke (string);

Rule number one: Principle of giving priority to take large;

Rule number two: Folded strokes determine the segments.

4.1. Strategy of Strokes Splitting-Increasing Test

For strokes codes sequence of a natural stroke (which may contain multiple strokes), from the first strokes codes, take a strokes code each time increasing to form a test stroke code, take the dictionary matching and save the matching results temporarily; then take the next strokes code to form a new test stroke code for matching operations, and so on. If you have extracted a predetermined one of a few more strokes off, that is, the extraction of strokes contained in all segments is completed, the stroke has been identified. If the strokes code of natural stroke code does not end, then clear the data structures which placed test stroke code (delete the test stroke code),then take the next strokes code to re-formed a test stroke code, and to determine the next stroke (the natural stroke with multiple strokes), Until all of the strokes code of natural stroke have been taken into the match, then take the stroke which contains the largest number of strokes as the match result and return it (whichever is greater priority), to give priority to extract folded stroke.

4.2. Strategy of Stroke Judgment-Similar Matching

Precise matching technology requires that signature code string to be identified must be equal to the stored signature code string in the feature dictionary. So, the reference template of the pattern in the dictionary must have equivalent coverage, *i.e.*, comprehensive, can cover the most common deformation of the pattern. Obviously, the high matching accuracy and the fast determination speed are obvious advantages of the technology, and recognition performance mainly depends on the completeness of the reference template. Because mode deformation range can not be limited, it may have rejection (the code is not in the library). In view of this, the establishment of a more complete feature library is an important task [3,4].

Similarity matching, it is raised by the unpredictable issues of deformation, it is usually take "distance" or "similar degree" of model to be knowledge and reference model as model criterion, and the definition of these criteria varied. The system uses the strategy of accurate identification first and then similar identification, that is, when exact match Produces rejection, it is transferred to produce similar matching module, to add similar identification of strokes of rejection. The string similarity, matching techniques are varied, more typical of them are dynamic programming method, fuzzy property law, the error correction method and various weighted matching method. When the stroke code uses non-equal length strokes code string, in order to reflect these differences of strokes in length and in type, this paper proposes fuzzy numbers of length and fuzzy numbers of direction to deal with them.

Take several written of folded stroke "乙"as an example, **Figure 2(a)** as the standard wording, standard stroke code is $G = aca$ and others are stroke variant. The strokes code set $\{a, b, c, d\}$ is the 4 yards direction code of the strokes. These graphics are similar but clearly different. The difference is that the String to be identified has produced a distortion, and we called those strokes code that out of standard strokes or produced a distortion as deformed strokes. If you take exact matching and stroke variants are not covered by the dictionary, the recognition algorithm will refuse to identify.

When taking similar code string matching, we consider the differences of ranging symbol in length and direction, and we take fuzzy number of the length and direction to define the symbol distance (the distance of strokes code):

$$d\left(x_k,g_i\right)=\begin{cases}0, x_k=g_i\\ f^D\left(x_k\right)f^L\left(x_k\right), x_k\neq g_i\end{cases} \tag{1}$$

where: $x_k\in X$ denotes the first K symbol of strokes code string to be identified. $g_i\in G$ denotes the first i s-ymbol of standard strokes code string associated with x_k. $f^D\left(x_k\right)$ denotes the fuzzy number of direction associated with x_k symbol. $f^L\left(x_k\right)$ denotes the fuzzy number of le-ngth associated with x_k symbol. the code string distance (stroke distance) is defined as the sum of distance of each symbol associated with X string:

$$D\left(X,G\right)=\sum_{k=1}^{m}d\left(x_k,g_i\right) \tag{2}$$

where, X denotes the stroke to be knowledge; G denotes standard stroke; m denotes the total number of strokes has been matched in X (some deformed strokes may be deleted);

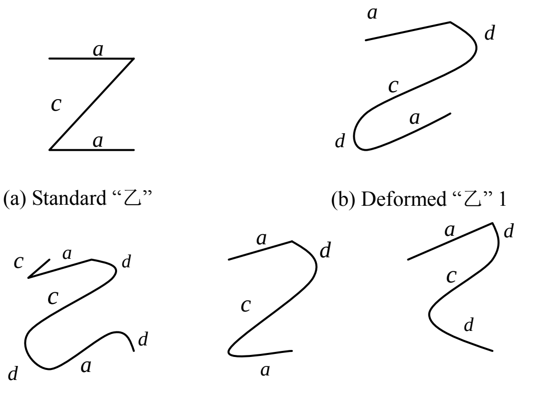
For stroke matching, first, computing the matching distance of stroke to be matched and the standard stroke according to (1), and then take the standard stroke with minimum distance as the judgment of stroke X to be knowledge. Can be seen, the main task is to calculate the fuzzy number of direction $f^D\left(x_k\right)$ and the fuzzy number of length $f^L\left(x_k\right)$, as follows.

1) Calculate $f^L\left(x_k\right)$:

Observation and analysis indicate that excess strokes usually shorter than the previous one and (or) the after one. So the fuzzy number calculation of length of the f-irst K strokes x_k in X can be defined:

$$f^L\left(x_k\right)=1-\frac{Len\left(x_k\right)}{\min\left\{Len\left(x_{k-1}\right),Len\left(x_{k+1}\right)\right\}} \tag{3}$$

where, $f^L\left(x_k\right)\in\left[0,1\right]$; If $f^L\left(x_k\right)<0$, take $f^L\left(x_k\right)=0$; $Len\left(x_k\right)$ denotes point length of strokes x_k.



(a) Standard "乙" (b) Deformed "乙" 1

(c) Deformed "乙" 2 (d) Deformed "乙" 3 (e) Deformed "乙" 4

Figure 2. Standard and deformed writings of folded strokes "乙".

2) Calculate $f^D\left(x_k\right)$:

The distance of two different strokes is determined by the degree of similarity of them, so the fuzzy number calculation of direction of the first K strokes x_k in X can be defined:

$$f^D\left(x_k\right)=1-sim\left(x_k,g_i\right) \tag{4}$$

where, $sim\left(x_k,g_i\right)$ denotes the similarity of strokes x_k and standard strokes g_i. It can be determined by querying the strokes similar table (see strokes space subdivision plans):

$$Sim=\begin{array}{c|cccccc|}
H & 1.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0\\
S & 0.0 & 1.0 & 0.7 & 0.7 & 0.9 & 0.9\\
P & 0.0 & 0.7 & 1.0 & 0.0 & 0.9 & 0.0\\
N & 0.2 & 0.7 & 0.0 & 1.0 & 0.0 & 0.9\\
F_P & 0.0 & 0.9 & 0.9 & 0.0 & \otimes & \otimes\\
F_N & 0.0 & 0.9 & 0.0 & 0.9 & \otimes & \otimes\\
 & H & S & P & N & F_P & F_N
\end{array} \tag{5}$$

Here are two strokes (stroke code string) matching algorithms and matching rules.

When taking similar match for code string, we need first select matching starting-point of string, *i.e.*, for code string to be knowledge, we from which symbol to start to match. Based on analysis of the discipline of shape variation, to consider the amount of computational algorithm as small as possible, we design two algorithms of starting-point selection method: methods of fixed starting point and floating starting-point. Fixed starting-point method, that is, no matter the first symbol code string to be identified and the first standard symbol code string are the same, we take it as the starting point to match. The method has no problem in match **Figures 2(a)** and **(b)** "*aca*" (standard code string) and "*adcda*" (code string to be identified) for the first symbol of this two code string are the same symbol "a". However, when matching code string (c) "*cadcdad*" and standard code string (a) "*aca*", there are dislocation of code string, which leads to matching errors. so ,we need to design floating-point fast matching method: First, compare the first symbol, if they are equal, we chose it as a starting point, or we need calculate the matching distances of the first symbol, second symbol with the standard string, taking as a starting point from the smaller one(to ensure the string long enough). Starting point is chosen, then we need take string symbol to be knowledge one by one to match with the standard symbol. If the two symbols are the same, continue to the next symbol match, or calculate the matching distances of the symbol, the next symbol with the standard string, taking as a matching result from the smaller one. Finally, we need calculate the matching distance of the two strokes, and a similar matching of stroke is completed [5].

5. Conclusion

For changes of unconstrained handwritten Chinese character, this paper presents a new Chinese character strokes recognition method: combining trial which is based on the structural analysis technology. The algorithm uses increase test technology to splitting and combining strokes, and the algorithm combines the proposed similarity matching technique based on the fuzzy number of length and the direction fuzzy number to complete stroke identification. The algorithm contains two processes of stroke splitting and stroke judgment, and this two processes are conducted alternately, *i.e*., we take split and composition as well as judgment, and in accordance with the principle of priority to take a large to extract the maximum possible compound stroke. Finally, we collect the first Chinese characters of nation standard written by 500 people and take tests. The results prove the effectiveness of the algorithm and the accuracy of stroke extraction, which lay a good foundation for the development of the follow-up whole word recognition algorithm.

6. Acknowledgements

This work was supported by the research fund of Sichu-an key laboratory of intelligent network information processing (SGXZD1002-10) and the key laboratory of the radio signals intelligent processing (Xihua University) (XZD0818-09).

REFERENCES

[1] B Jia, X. D. Tian and F. Yang, “Off-Line Handwritten Chinese Character Recognition Based on Double Contour Feature,” 2009 *International Symposium on Intelligent Information Systems and Applications*, Qingdao, 28-30 October 2009, pp. 399-402,

[2] T. H. Su, T. W. Zhang and D. J. Guan, “Off-Line Recognition of Realistic Chinese Handwriting Using Segmentation-Free Strategy,” *Pattern Recognition*, Vol. 42, No. 1, 2009, pp. 167-182. doi:10.1016/j.patcog.2008.05.012

[3] X. N. Huang, “An Multiple Classifiers Integrated System of On-Line Natural Handwritten Chinese Characters Recognition,” *Journal of Chinese Information Processing*, Vol. 14, No. 5, 2000, pp. 37-41.

[4] X. N. Huang, “Extraction of Natural Handwritten Chinese Character Strokes and Roots,” *Journal of Chongqing University* (*Natural Science*), Vol. 23, No. 5, 2000, pp. 104-107.

[5] Y. F. Sun, Y. Chen and Y. Z. Zhang, “Symmetry-Based Recognition Method for Similar Chinese Characters,” *Journal of Chinese Information Processing*, Vol. 18, No. 2, 2004, pp. 51-57.