A Structural Approacn To On-Line Chinese Character Recognition

Y. J. Liu J. W. Tai

Institute of Automation, Academia Sinica

Beijing , China

1. Introduction

Chinese characters are pictorial patterns consisting of curve Segments as well as straightline segments. A Chinese character Which consists of one part with several strokes is called a single component character .In contarast ,a Chinese character which consists of a few parts , and each part has several strokes is called a compound character. Two strokes of a character may be unconnected, or one stroke of a character may contain some connecting points which join the stroke to other strokes.

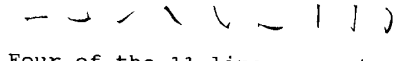
The structural properties of Chinese characters are important , for on-line Chinese character recognition .The composing structure of Chinese characters can be divided into four levels . 1. Whole character level , 2. Parts of character level,3. Stroke level, and 4. Line segment level. We can find there are 7000-10000 whole characters,200-500 parts of character, 40-80 strokes and 4-8 straightline segments . Level 1 is the hightest level with complex structure,and level 4 is the lowest level with simple structure.

On the basis of previous works[2-5] by second author of the article ,a structural (or syntactic) approach for on-line Chinese character recognition is proposed in this article.Some strokes or straightline segments can be considered as a set of primitives , and production rules can also be found according to the properties of characters . In this article , A fuzzy attributed finite-state grammar and corresponding automaton are introduced for recognizing line segments . And according to intrinsic structural properties of Chinese character ,an order arrangement is provided. As a result, an on-line Chinese character recognition system has implemented by the techniques mentioned above. The system consists of a small input tablet,an interface board with a IBM-PC/XT/AT. Which can recognize about seven thousand on-line handwriting Chinese characters with some constraints.

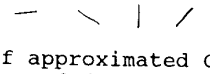
2. Description of distorted versions of stroke

The stroke extraction is one of the basic problems for on line Chinese character recognition. Various methods have suggested. It is known, a stroke is composed of a straight line sequence or can be approximated by a straight line sequence. According to statistical data, there are probably 38 kinds of regular strokes, but lots of people can not write in regular stroke form , When they write Chinese characters . Even though hand writing is constrained , the kinds of stroke are greater than 38. In fact,it has to extract 70 kinds of stroke at least.

The stroke samples can be decomposed into a sequence of straight line segments. For regular strokes of constrained hand writing Chinese character , which are composed of 11 kinds of generalized line segments as follows :



Four of the 11 line segments are more basic , and others can be approximated by these four basic line segments



Most of approximated Chinese characters consisting of four basic line segments can classify by computer .

Make a comparsion between strokes and regular strokes , we can find most of strokes include line segments appearing in the corresponding regular strokes. The difference is that some line segments may be added into those strokes , especially, at starting position and inflexional position of the strokes , and the proportion of length of segments in main directions may be different . We consider that the hand writing strokes are distorted versions of regular strokes. In order for stroke recognition , both conventional templet matching technique and attributed finite state grammar approach as well as fuzzy set [8] are used .

If a regular stroke is considered as a

CH2614-6/88/0000/0808$01.00 © 1988 IEEE

templet ,which consists of n line segment $a_1, a_2, \ldots a_n$, the length of a_i is $A(a_i) = \bar{l}_i$. i = 1,2, ... n . The noisy strokes and distorted versions are derived from two kinds of variations .One is there may be some new short line segments added into starting position and concatenation position of two consecutive line segments of regular stroke. Another one is lengths of segments of regular stroke may be varied . Such a set of distorted versions of regular stroke is suitably described by an attributed finite state grammar G[7].

From the recognition point of view, we can define an attributed finite state automaton T as follows :

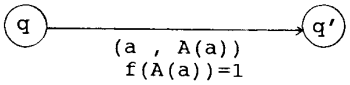
$$T = (\sum , Q , \delta , q_o , H , q_f)$$

where $\sum$ is a finite set of input symbols with attributes represented by (a_i , l_i) , Q is a finite set of states, each state is associated with a reference value , δ is a mapping of $Q*\delta$ into Q, $q_o \in Q$ is the initial state , H is a finite set of functions , $q_f \in Q$ is the final state .

The difference between an arrtibuted finite state automaton and a conventional finite state automaton is transition condition . The interpretation of

$$\delta(q,(a,A(a))=q' \text{ if } f(A(a))=1, q,q' \in Q , a \in \sum$$

is that the automation T , in state q and scanning the symbol a with attribute A(a), goes to state q'if A(a) satisfies condition f(A(a)) = 1 . The state transition digram corresponding to δ (q, a, A(a))= q', f(A(a)) = 1 is shown as follows :



An attributed finite state automaton is simple , and its function is better than a finite state automaton.Actually,attributed finite state automata are not sufficient for recognizing hand writing strokes. The reason is that state transition has only two possibilities, stays at the state or goes to next state, no context information can be used .

3. Fuzzy attributed finite-state automaton

From above discussion, an attributed finite state automaton is not so powerful for recognizing hand writing strokes. It is reasonable to realize a set of similar hand writing strokes should be a fuzzy set, and by means of fuzzy information processing technique, a fuzzy attributed finite state grammar can be defined[7].

Corresponding to a fuzzy grammar G , a fuzzy attributed finite state automaton is defined in the following.

A fuzzy attributed finite state automaton is an seven-tuple $T_f = (\sum , Q, F, \delta, H, L, q_o)$, where $\sum$, Q and q_o are the same as T, H is a finite set of given functions , and $L = \{l_i\}$ is a finite set of positive real attributes corresponding to states .

F is a mapping of $\sum *Q*L$ to L , the interpretation of

F [q_i , $\tilde{l}_i$, (a,u)] = l_i is that $\tilde{l}_i$ represents attributes accumulated by transitions at state q_i before current transition , (a,u) is one symbol a with attribute u as input , the transition rule is

$$l_i \xrightarrow[(a,u)]{} \begin{cases} \tilde{l}_i + 0 & \text{if } a \bar{\in} \sum_i , \sum_i \subset \sum \\ \min[(\tilde{l}_i + u), \bar{l}_i] & \text{if } a \in \sum_i , \sum_i \subset \sum \end{cases}$$

where $\bar{l}_i$ is attribute of line segment of templet .

And δ is a mapping of $\sum *Q*L$ to Q .

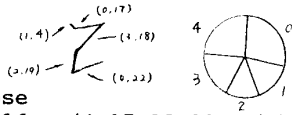
The membership of a string x is

$$\mu(x) = \frac{\sum_i^n l_i}{\sum_j^m u_j} \text{ , and } \sum_j^m u_j = k \sum_i^n \bar{l}_i$$

where the numerator represents summation of attributes corresponding to the mapping of F, and denominator represents summation of attributes of the input string . And all $q \in Q$ can be the final state.

Example 3.1 An example is given to illustrate a hand writing stroke is recognized by T. The regular stroke is considered as templet , L={4,8,4}, $Q=\{q_o, q_1, q_2\}$, $\sum$={ 0,1,2,3,4 },the positive constant K is introduced for normalization purpose. The transition rules of attribute and transition rules of state has been given in [7] :

For an input string
x = (1,4)(0,17)(3,18)(2,19)(0,22), the stroke corresponding to x is shown as fellows:



We chose
$k=(\sum_j^m u_j)/16 = (4+17+18+19+22)/16 \doteq 5$.
So $k\bar{l}_0/2 = 10$, $k\bar{l}_1/2 = 20$, $k\bar{l}_2/2 = 10$, ($\sum_{i=0}^{2} \bar{l}_i = 16$) .
The $\mu(x)$ is cbtained as follows:

$$\mu(x) = \frac{17+37+20}{4+17+18+19+22} = 0.925$$

A positive real number r is chosen as threshold , $r < 1$,so long as $\mu(x)$ is greater than r for an input string x,then x is recognized as the referential regular stroke of automaton .

4.Order arrangement of line segments.

From previous discussion,straight line

segments of a Chinese character can be extracted by fuzzy attributed finite state automaton.Another interesting point is how to get an order of line segment in order for recognizing Chinese characters. We know when someone writes a Chinese character on an input tablet, an order of stroke is obtained, but different writers may have different order of stroke. A method for order arrangement independent of writers is proposed in [6] . For two line segments of a Chinese character, according to the intrinsic structural properties of the character , some criteria are given to verify which line segment takes precedence over another. If the appearance of two line segments satisfies the structural criteria,an index 1 is assigned to the relation between two line segments , otherwise an index 0 is assigned. For instance, consider a set of four basic line segments as primitives , two segments are connected and relation between them is (m,m) , by structure properties of Chinese characters, a precedence matrix is given in Fig. 4.1

A \ B	—	\	|	/
—	0	1	1	1
\	0	0	0	0
|	0	1	0	0
/	0	1	1	0

Fig. 4.1 A precedence matrix of four basic line segments .

From order of appearance of line segments , two samples of Chinese character " 禾 " written by different persons are given in Fig. 4.2 (a) and (b). The order of line segments s_1 is different from order of line segments s_2 , where

s_1 = ／ , — , | , ／ , ＼

s_2 = ／ , | , — , ／ , ＼

According to instrinsic structural properties of order of line segments is

s = ／ , — , ／ , | , ＼

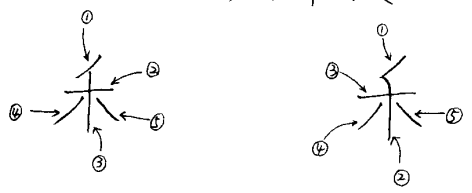


Fig. 4.2 (a) (b)

A sample of 禾 A sample of 禾

We can transform different order of line segments to s, the procedure of transformation do not discuss here , which can be found in [6] .

An on line Chinese character recognition system has implemented by the techniques mentioned above . The system consists of a small input tablet , an interface board and an IBM PC-XT . Some constraints of writing characters are required of writers .During a short period of training, the error rate of recognition is less than 5% for recognizing 7000 Chinese characters .

5.Concluding remark

The work of fuzzy attributed finie-state grammar and corresponding automaton is an extension of the previous work on attributed grammar [2-6] by the authors. The point view of a fuzzy attributed grammar is embedding a regular stroke into a set of distorted strokes with different structure , then the mumbership function is considered as a similarity measure for character classification . Which is better than a distance measure , sush as levenshtein distance . The reason is that the structure of distorted strokes may be different from structure of regular stroke, The high recognition rate is hardly to obtain by error correcting technique[1].

As a matter of fact , this approach can be applied not only for Chinese character classification but also for line drawn pattern recognition .

References

1. K. S. Fu, Syntactic Methods in Pattern Recognition,Academic Press, 1974.
2 J. W. Tai and K. S. Fu, Semantic Syntax-Directed Translation for Pictorial Pattern Recognition , Proc. 6th Int. Conf. Pattern Recognition , Munich, 1982.
3. J. W. Tai, A Kind of Attributed Grammar for Pattern Recognition,Acta Automatica Sinica, Vol. 9, No.2 , 1983.
4. J. W. Tai, A Line Drawing Pattern Recognition Method , Acta Automatica Sinica, Vol.11, No.3 , 1984.
5. J. W. Tai,A Syntactic-Semantic Approach for Chinese Character Description , Computer Processing of Chinese and Oriental Languages, Vol.1, No.3 , 1984.
6. Y. J. Liu and J. W. Tai, A Method of stroke order Arrangement for On Line Chinese Character Recognition, to be published on Automatica Sinica, Vol.14, No.3, 1988.
7. Y. J. Liu and J. W. Tai , A Fuzzy Attributed Finite State Automaton for On Line Chinese Character Recognition, to be published on Automatica sinica , Vol.14,No.2,1988.
8. A. Kandel, Fuzzy Techniques in Pattern Recognition, John Wiley and Sons, 1982.