Recognition and Representation of Text Characters Using Rubber Band

Jingzhi JIANG Woontae KIM Hideyoshi TOMINAGA

dept. of Electronics and Communication Engineering, Waseda University

Abstract

An important step in the communication of text including unknown characters is the representation of characters. To realize bilateral exchange of texts, representation of unknown characters, shch as foreign characters, created characters, design characters, and historical characters which are not being widely used nowadays, is an essential subject. In particular, although many Asian countries are using Chinese characters commonly, the fonts and glyphs are different.

In this paper, we propose a method to recognize the configuration of an unknown character and represent the character by composition of common components called strokes. In the proposed method, a character is firstly analyzed into small elements called parts. Then parts are merged to form a stroke. Experimental results on Chinese characters are given.

1 Introduction

Traditionally, text characters are represented by their codes. The main disadvantage of this coding transmission method is that it is difficult to transmit and display *unknown characters*'with terminals intended only for domestic use. The unknown characters include foreign characters, created characters, design characters, and historical characters which are not being widely used nowadays. For example,Chinese characters are used commonly in China, Japan, Korea, Taiwan, . Singapore, and other countries. However, their codes are different, and not only glyph but fonts are different as well.

A character font database has been proposed by Yao, Kameyama and Tominaga [1],[2], in which common components are extracted from characters and each characters is represented by composition of the common components, and a *rubber band* method has been used to code the common components. In this system, the common components are called *mechanical strokes*, and a stroke extraction method, where horizontal strokes and vertical strokes are extracted initially is used.The major problem seems to be the lack of a effective decomposition method to extract strokes.

There have been many different analytic methods used in structural analysis including transformations and decompositions of the boundary of the character, and partitions of the interior of the character into simple pieces. In particular, in the field of pattern recognition, we can find many related works, in which a structural approach is used to decompose complex scenes or objects into simple components.(For a complete survey, see Pavlidis [3].) And Bjorklund defines a low-level component as an *atom* in the global shape analysis by *k*-syntactic similarity [5]. A related term is *primitive* which has been used in syntactic pattern recognition, and *interior cluster* or *simple parts* has been used in shape description by Shapiro [4].

In this paper, we propose a method to recognize the configuration of an unknown character and represent the character by composition of common components called strokes. we propose a decomposition of character which uses *parts* and *strokes*. We firstly decompose a character into very small elements called *parts*, and parts are integrated to form a *stroke*. A stroke can coincide with a part but there may be also situations where a stroke is formed by more than parts. Strokes does not necessarily correspond with those in humans writing. In this paper, section 2 describes rubber bands, their properties and gives an algorithm to generate rubber band. Section 3 discusses several basic operations about rubber band. Section 4 describes the derivation of parts from characters and the unification of parts into strokes.

2 Rubber Band

Rubber band, *Contact line* and *Bridge line* are defined as follows (Fig.1.a).

- Rubber band : A convex polygon that includes an image block with the shortest length.
- Contact line – The lines of a rubber band, which coincide with the image block.
- Bridge line – The lines of a rubber band, that do not coincide with the image block.

Some of the properties of a rubber band are listed below:

- If the bridge lines of a rubber band exist, the number of bridge lines is equal to the number of contact lines.
- A contact line is a part of the contour of the image block.

0-8186-4960-7/93 $3.00 © 1993 IEEE

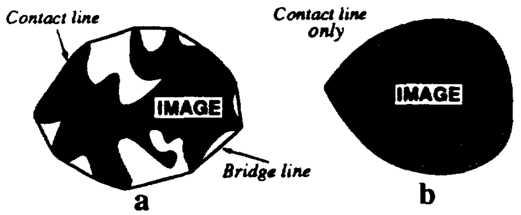

Figure 1: Rubber band, contact lines and bridge lines

- The rubber band corresponding to an image block is unique; but the image blocks corresponding to a rubber band could be determined arbitrarily.

As a general rule, the number of contact lines is always equal to the number of bridge lines, but there is also an exceptional situation when any bridge line does not exist. Fig.1.b gives an example in which an image block formed by a complete convex shape.

Now we present an algorithm to generate rubber band quickly using side distances and slopes. A, B, C, D are the dots with the shortest distances from four sides of a rectangle that surrounds an image block. The rubber band can be constructed by connecting a set of considerable protrusion points. The protrusion points are found by the following algorithm.

1. Start from A, A becomes the first protrusion point.
2. Calculate the slopes of the lines from A to another dot on the contour from A to B.
3. Set the dot that gives the maximum slope to be the second protrusion point.
4. Using new protrusion point that found in 2, reiterate 2 until B.
5. Similarly, from B to C, from C to D and from D to A, do 2 and 3.

3 Basic Operations

3.1 Unification of Rubber Band

The unification operation means to make only one rubber band against two image blocks simultaneously. If there is not any image block or a part of image block except the two image blocks in the unified rubber band, then the unification will be called as a possible unification. (Fig.2)

3.2 Modification of Rubber Band

It is necessary to modify the rubber band after the unification of two image blocks. For two image blocks *A* and *B*, the modification is accomplished by the following steps:

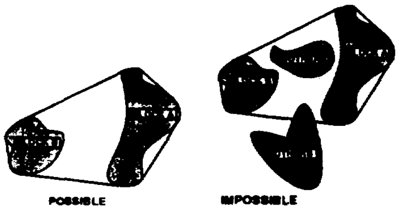

Figure 2: Possible and Impossible Unification

- We refer to the rubber band of image block *A* as *Ac* and the rubber band of image *B* as *Bc*.
- Find the corners of the rubber band *Ac* and the rubber band *Bc* with Rosenfeld-Johnston method [6], examine not only the minimum of the corners, but the connections between the corners and image block *A* and *B* as well, then make conclusions about corner points finally.
- Detect a dot *a1* from the corner points of *Ac* and the other dot *b1* from the corner points of *Bc*, which the distance from *a1* to *b1* is minimum, and detect *a2* and *b2* similarly.
- Connect *a1*, *b1* and *a2*, *b2* with two straight lines to obtain a new rubber band, or connect *a1*, *b2* and *a2*, *b1* when the two straight lines intersect.

3.3 Inverting of Rubber Band

Inverting means turning black pixels to white and white ones to black, in the interior of a rubber band. Even if we invert an image, only little information concerning intrusion and protrusion of the original image block is lost. An inverted image is formed by s set of small and simple image blocks generally. It is easy to study an inverted image instead of the original image block.(Fig.3)

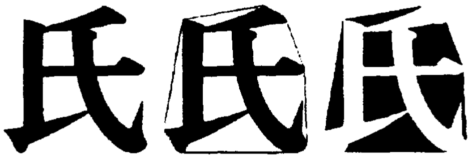

Figure 3: The inverting of a rubber band

3.4 Judgment Criteria of Stroke

In the course of the decomposition of characters, the following parameters are considered:

1. A rate of the black area to the white area in the interior of a rubber band.

2. A rate of the length of the rubber band to that of the contour of a character.

3. A rate of the length of the bridge lines of the rubber band to that of the contour which do not coincide with the rubber band.

We define L_{rubber} as the length of the rubber band, and L_{out} as the length of the contour of a character. an image block is judged whether a stroke or not a stroke as follows.

$$If \quad \frac{L_{rubber}}{L_{out}} \leq L_{+h}, \quad then \quad Not\ Stroke,$$

$$If \quad \frac{L_{rubber}}{L_{out}} > L_{+h}, \quad then\ Stroke.$$

where L_{+h} is a threshold. Empirically we use 0.9 for L_{+h}. If a image is not a stroke, the decomposition process of the image will be continued. It is certainly that we can use the other parameters to judge a image whether is a stroke or not a stroke. It was illustrated by experiment that the second rate seems more efficient.

DATA	L_{rubber}	L_{out}	Rate	Conclusion
DATA 1	466 dots	577 dots	0.808	NOT
DATA 2	413 dots	432 dots	0.956	STROKE
DATA 3	594 dots	732 dots	0.811	NOT
DATA 4	235 dots	241 dots	0.975	STROKE

Table 1: The judgment of the data in Fig.6

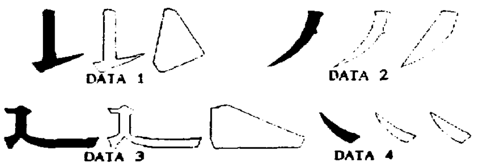

Figure 4: A set of data for judgment

3.5 Shrink of Rubber Band

This operation means to make a new rubber band after eliminating all dots of the contact lines from a rubber band. Sometimes it is necessary to extend the possibility of the unification of rubber bands. And as a major weakness, the result of decomposition of a character is very sensitive to noise and the changes of fonts. The noise sensitivity can be reduced by preprocessing and smoothing of the boundary, but the fonts sensitivity can not. Using the shrink of rubber bands, it is possible to reduce the fonts sensitivity and to extend the possibility of the unification of rubber bands.

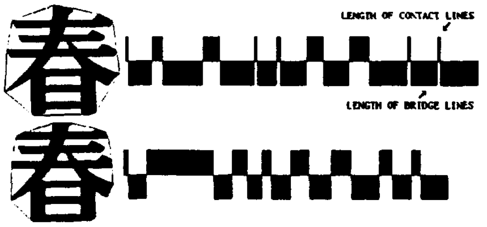

Figure 5: Shrink of rubber band

4 Extraction of Strokes

To represent unknown characters, the decomposition of characters by extracting strokes is an essential function. If one character consists of more than one image block, we begin with labeling each image block of the text character. In this section we describe the derivation of parts from one of these image blocks and integration of parts into a stroke. Fig.6 gives the flowchart.

A common process is used not only in the derivation of parts but also in the unification of parts into strokes, which derive the region pinched by two image blocks. This process executes in the following steps:

1. Make the rubber band of one image block and invert it.

2. Choose all of the pairs of blocks that their unifications of the rubber band are possible.

3. For one of the pairs of blocks, unify and modify the rubber bands.

4. Invert the rubber bands which have been unified and modified in 3.

Then, the region which adjoins two image blocks simultaneously will be considered as a ***part***.

After derivation of parts, it is necessary to unify them into strokes and eliminate them from the original character. It should be noted, however, that there may be situations where some parts have become strokes already. The region which pinched by two parts can be derived using the same procedure 2, 3, 4 as described above, that was used in derivation of parts. Thus, this region will be connected with correspondent blocks, when it coincides with the original character perfectly, and against two lines that connected in modifying process, two criterion must be satisfied: 1)the angle is less than threshold θ and 2)the length rate is greater than $1/d$ or less than d. We fixed threshold θ at $\pi/20$, and d at 2.0. If there is not any part that can be unified into strokes, the biggest part(weight) will be considered as a stroke. It is noticeable, however, that we eliminate the ***parts*** from original characters not ***unified strokes***.

Experimental results on Chinese character are given in Fig.7, Fig.8.

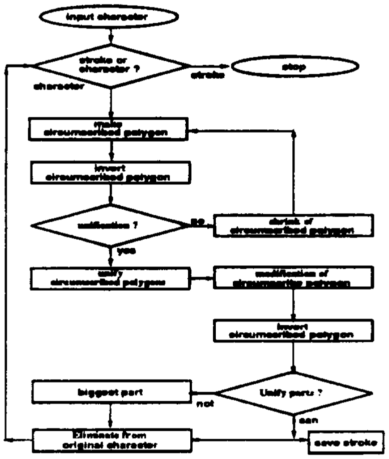
input character
stroke or character ?
stop
stroke
character
make circumscribed polygon
invert circumscribed polygon
unification ?
no
shrink of circumscribed polygon
yes
unify circumscribed polygons
modification of circumscribe polygon
invert circumscribed polygon
biggest part
Unify parts ?
not
can
Eliminate from original character
save stroke

Figure 6: Flowchart of Extraction of Strokes

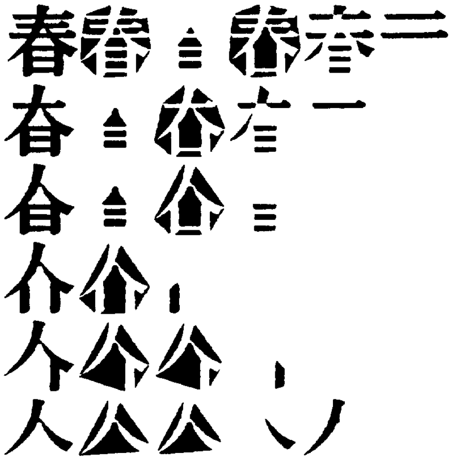

Figure 7: Decomposition of a Chinese Character

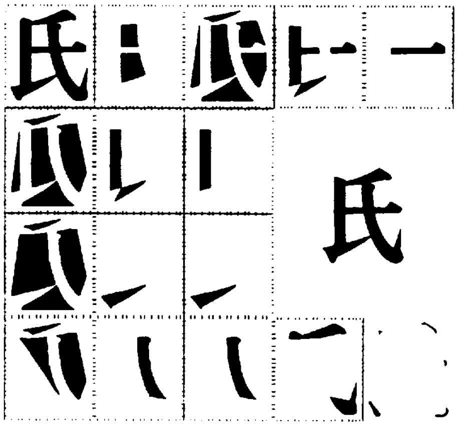

Figure 8: Decomposition of a Chinese Character

5 Discussion and Conclusions

We defined a rubber band and described its properties. Then we proposed a method to represent an unknown character by composition of common components. However, the results reported here are only initial experimental results. We expect to perform many more experiments using more extensive data and studying further the effect of different fonts. In summary, the representation of characters using rubber bands as a whole seems to be promising tool in extraction of strokes of unknown character communication system.

References

[1] Yao, Kameyama and Tominaga:"A Study Automatically Generating System of Structural Character Fond Database,"IEICE Technical Report of Japan(Sep.1990)

[2] Yao, Kameyama and Tominaga:"A Character Font Database Capable of Learning Unknown Characters and its Document Communication Protocol,"IEICE Technical Report of Japan(Mar.1991)

[3] T.Pavlidis,"A review of algorithm for shape analysis,"Comput.Graphics Image Processing, vol.7, pp.243-258, Apr.1978.

[4] L.G.Shapiro:"A structural model of shape,"IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-2, No.2 pp. 111-126, March.1980

[5] C.Bjorklund and T.Pavlidis:"Global Shape Analysis by k-Syntactic,"IEEE Trans. Pattern Anal. Machine Intell.. vol. PAMI-3, No. 2 pp. 144-155, March. 1981

[6] A.Rosenfeld and E.Johnston:"Angle detection on digital curves,"IEEE Trans.Comput., vol. c-9, 1973, pp.875-878

[7] H.Nishida and S.Mori:"Algebraic Description of Curve Structure,"IEEE Trans. Pattern Anal. Machine Intell.. vol. 14, No. 5 pp. 516-533, May. 1992