Segmentation and Recognition of Continuous Handwriting Chinese Text

Chen Hong, Gareth Loudon, Yimin Wu and Ruslana Zitserman.
Apple-ISS Research Center, National University of Singapore,
Heng Mui Keng Terrace, Singapore 119597
E-mail: chenhong@apple-iss.iss.nus.sg

ABSTRACT

This article introduces the basic segmentation problems in Chinese handwriting and several prior work to solve the problems. A new segmentation method is proposed, which is applicable to both on-line and off-line system for free-format handwritten Chinese character sentences. This method performs basic segmentation and fine segmentation based on varying spacing thresholds and minimum variance criteria. The five most probable ways of segmentation are derived from this stage. Each way of segmentation is fed through character recognizer individually and the lattice search engine is applied along the linguistic information to determine which way of segmentation is most likely. The sentence level accuracy, with an average of eight handwritten characters in each sentence, is 49% for top one choice and 56% for top five choices. The average character recognition accuracy is 85%. Finally, Further work is briefly described.

1. INTRODUCTION

A typical handwriting recognition system, for either on-line or off-line, for free-format and continuous text consists of the following processing stages:
text input -> segmentation -> recognition -> (language model) -> result
Segmentation is a process of separating characters from one another. It is a very important step to prepare character data for the recognizer because most recognizers can only deal with isolated characters. A great deal of research work has been done to solve the Chinese character recognition problem [1][2][3][4][5][6]. The overall accuracy of a system is determined by both segmentation and recognition.

Several methods [7][8] have been reported for the segmentation of Roman alphabets writing. Most of these methods make use of time and spatial information or character shape information provided by the character recognition process. However, Chinese characters are quite different from alphabets. Their structure is more complicated. In addition, our goal is to find a generic algorithm that is applicable to both on-line and off-line system. Therefore the time-related information such as stroke sequence cannot be used.

1.1 Characteristics of Chinese Handwriting

The following are the characteristics of Chinese handwriting and their related segmentation and recognition problems.

(1) Most of the Chinese characters consist of more than 2 radicals and any of these radicals can be an individual character in itself. This makes segmentation difficult because from a recognition point of view, the result is a valid even if a character is mis-segmented into several radicals. Therefore linguistic knowledge is needed to distinguish correct segmentation just as human beings usually group radicals into a character by the meaning and context.
(2) In free-format handwriting, space between characters and space between radicals vary considerably. Though it is common for people to leave a larger gap between characters than that between radicals, the spatial information is not reliable.
(3) Writing style varies from person to person; however, for a particular person, the writing style is somewhat consistent. This conclusion is drawn from the observation that the correct segmentation of a sentence is frequently the one where the width of each segment is evenly distributed.
(4) Chinese characters are written either horizontally or vertically. Without losing its

generality, the following study focuses on the horizontally written sentences and it can be easily applied to vertically written sentences with only small modifications.

(1) and (2) tell us that spatial information may be of some help for the segmentation, but should not be the only criteria. (3) gives us a hint that the width of character is a important factor to consider.

1.2 Prior Work

Some segmentation methods have been proposed to separate Japanese characters which are also ideographic in nature. Though Japanese characters are very similar to Chinese characters, Japanese usually write more neatly in handwriting.

In [9], Shunji introduced following stages to split the text string into isolated character: preprocessing -> segmentation of non-touching characters -> segmentation of compound characters -> segmentation of touching Japanese characters -> segmentation of English characters. In each stage, some hypotheses for character segmentation are generated on the basis of information obtained in the earlier stages and these hypotheses are verified by the character recognition results. The method is quite effective for the printed document.

Murase [10] proposed another important approach in which character recognition results and grammatical constraints are used for character segmentation. First, all possible character candidates are extracted and recognized. Then, the most plausible sequences of characters are sought by dynamic programming and they are verified under grammatical constraints. His method is devoted to on-line handwritten character recognition and cannot solve the touching character problem.

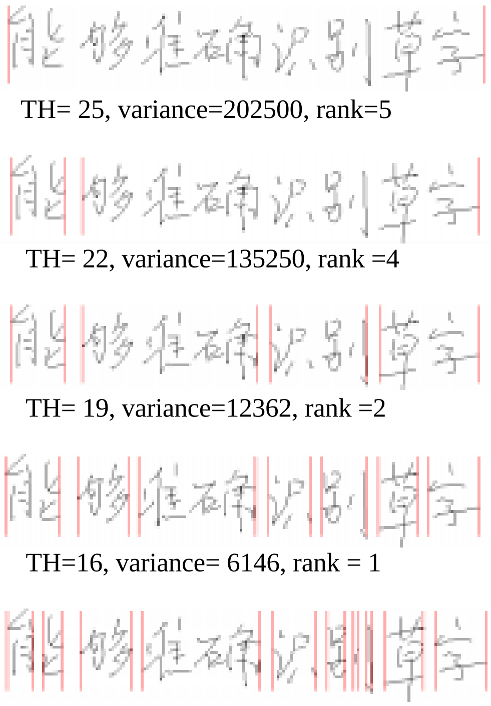
2.SEGMENTATION ALGORITHM

The segmentation process is broken down into several steps: first, basic segmentation is done by varying the space threshold. Five possible ways of basic segmentation are obtained from this stage. Then a minimum variance criteria is used to fine tune the possible ways of segmentation recursively. Touching characters are separated and the wrongly separated parts are merged during this stage. The possible ways of segmentation increase as the procedure goes on and are sorted based on the variance criteria. The top five ways of segmentation are selected and combined to form the lattice. Each possible segment in the lattice is fed into a character recognizer. The recognition results and their probabilities are integrated into the lattice. Finally the lattice search engine is utilized to find the overall result for the sentence.

2.1 Basic Segmentation

The basic segmentation is very simple. Five ways of segmentation are obtained by using five different space thresholds. First, all the data points are projected onto X-axis and Y-axis. Scan the projection from left to right. If a gap is larger than the threshold, mark the gap as a boundary. Vary the threshold value and do the above process again. Once a particular segmentation is decided, the width median of each segment and the width variance are calculated and stored for the next stage process.

Figure 1 shows an example of this basic segmentation with different threshold values.



TH= 25, variance=202500, rank=5

TH= 22, variance=135250, rank =4

TH= 19, variance=12362, rank =2

TH=16, variance= 6146, rank = 1

TH= 13, variance=13622, rank =3

Figure 1. Basic segmentation

2.2 Fine Segmentation

The basic segmentation is not effective in the case where more than one radical are joined together or the distance between radicals is too large. In the fine segmentation stage, these problems are solved by splitting joined segments and merging over-segmented parts. Fine segmentation is performed based on the result from the basic segmentation. Five possible ways of basic segmentation are sorted according to the variance. The top two ways with minimum variance are selected to do the fine segmentation.

The following process is repeated until no more new segmentation is generated. Each segment is scanned from left to right. If the width of the segment is too small compared with the median, try to combine the segment with the next few segments. Calculate the median and variance of this new segmentation. If the variance decreases, regard the new segmentation as a valid way and store it. Otherwise, ignore the new segmentation. If the width of segment is too large, estimate how many parts the segment should be split into and split the segment. Again, calculate the variance of the new segmentation and decide whether to count it as a new way of segmentation or to discard it.

Figure 2 gives one fine segmentation result which is obtained by splitting one of the segment.

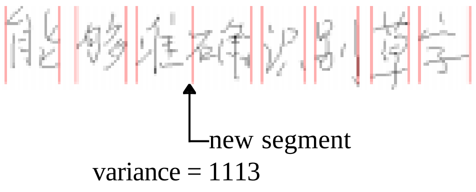


Figure 2. Fine segmentation

2.3 Lattice Formation and Search

Segmentations are re-sorted according to the variance criteria and chosen the top five ways to form the lattice. The next stage is to apply recognizer and language model to get the overall result for the sentence.

Each possible segment is fed into the character recognizer to get 50 candidates together with their probabilities. The character recognizer we used here is an on-line recognizer based on discrete HMMs. Theoretically, the recognizer can be of any type, on-line or off-line [4], according to the application.

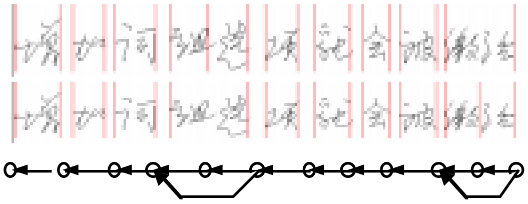


Figure 3. Lattice Generation

The result of segmentation and candidate character recognition is expressed in the form of lattice which is a two-dimensional weighted directional graph. A node represents a possible boundary in the basic segmentations; An arc is associated with the candidate characters and the corresponding probabilities from the recognizer. A sample lattice is shown in Figure 3. In this simple example, neither basic or fine segmentations are completely correct. However, by using the lattice engine, the correct segmentation is included in the search paths. So it is possible to be recovered.

Next, the language model is applied to search through all the possible paths in the lattice and the results are listed according to the recognition score and language model score.

3. EXPERIMENTAL RESULTS

One hundred sentences were collected from seven different users for the test with eight characters on average in each sentence. No rules were imposed on the way and style of writing. Most of the sentences were written in a very cursive style. Table 1 gives the experimental results.

	accuracy
Basic and fine segmentation	79% (top1) 96% (top5)
Sentence (with lattice)	49% (top1) 56% (top5)
Character (without lattice)	70.3%
Character (with lattice)	85%

Table 1. Experiment result

In the sentence accuracy statistics, the sentence is reguarded as correct when both the segmentation and recognition result are correct. And the character accuracy without lattice is obtained by estimating the character accuracy of top 1 choice for each segment in the most likely segmentation.

4. DISCUSSION

Although out test is performed on on-line data, the method proposed here is equally applicable to both on-line and off-line systems. The system can be easily customized by plugging in different recognizers since the algorithm doesn't make use of any time-related information.

From the experimental results, it can be seen that the language model helps a great deal in the recognition accuracy. In addition, it was observed that the number of candidates for each segment is critical to the overall speed and accuracy. A smaller number pulls down the accuracy because the correct character may be missed out. In this case, the language model cannot do anything if the correct character is not included in the candidate list. On the other hand, a larger number of candidates requires more memory and slows down the search speed. Fifty was found to be the optimal number for this application through the experiments.

5. CONCLUSION

Continuous Chinese handwriting recognition has always been a difficult problem because of both segmentation and recognition issues. In this paper, a generic segmentation method based on minimum variance criteria is described. Combined with a lattice search engine, the algorithm provides large flexibility to handle free-format Chinese handwriting. The experimental results show that the method provides a promising solution for a natural and fast Chinese pen input system.

However, this method only performs segmentation, recognition and linguistic analysis after a whole sentence has been completely written. Therefore, the respond time is still not satisfactory for an on-line recognition. Our future work will focus on the improvement of the segmentation algorithm and the recognizer so that the system can still perform segmentation and recognition while users are still writing.

REFERENCES

[1] R. Nag, et al., Script Recognition using hidden Markov models, Proc. ICASSP, pp. 2071-2074 (1986)

[2] M.Y. Chen et al., Variable Duration Hidden Markov Model and Morphological Segmentation for Handwritten Word Recognition, IEEE Transactions on Image Processing, Vol.4, No.12, pp. 1675-1688 (1995)

[3] J.C. Anigbogu and A. Belaid, Hidden Markov Models in Text Recognition, IJPRAI, Vol.9, No.6, pp. 925-958 (1995)

[4] Gareth Loudon, Chen Hong, Yimin Wu and Ruslana Zitserman, The Recognition of Handwritten Chinese Character From Paper Records, Proc. IEEE-TECON, to be published, (1996)

[5] M.Y Chen et al, Off-line Handwritten Word Recognition Using Hidden Markov Model Type Stochastic Network, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No.5, pp. 481-496 (1994)

[6] S.B. Jeng, C.H. Shih et al., On the Use of Discrete-state Markov Process for Chinese Character Recognition, Visual Communications and Image Processing, Vol. 1360, pp. 1663-1670 (1990)

[7] G.F.Groner, Real-time recognition of handprinted text, Proc.FJCC, pp. 591-601 (1966)

[8] C.C.Tappert, Cursive script recognition by elastic matching, IBM J. Research and Development, 26, pp. 765-771 (1982)

[9] S. Ariyoshi, A Character Segmentation Method for Japanese Printed Documents Coping with Touching Character Problems, Proc. ICPR, pp. 313-316 (1992)

[10] H. Murase, Online recognition of free-format Japanese handwritings, Proc. ICPR , pp. 1143-1147 (1988)

[11] Y. Maeda, F. Yoda, K. Matsuura and H. Nambu, Character Segmentation in Japanese hand-written document images, Proc. ICPR, pp. 769-772 (1988)

[12] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.