

# A two-stage handwritten character segmentation approach in mail address recognition

Zhi Han<sup>1,2</sup>, Chang-Ping Liu<sup>1</sup>, Xu-Cheng Yin<sup>1,2</sup>

<sup>1</sup> Character Recognition Center, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> Graduate School of Chinese Academy of Sciences

zhi.han@mail.ia.ac.cn

## Abstract

Character segmentation has become a crucial step for mail address recognition in the automatic post mail sorting system. In this paper, a two-stage character segmentation algorithm according to the characteristics of handwritten mail address characters is proposed. In the simple segmentation stage, the block sequence is extracted from the mail address image using the structure-based methods, including projection profile analysis, connected components analysis and stroke cross number analysis. In the precise segmentation stage, all candidate segmentation paths are created by combining the neighboring blocks and represented with a candidate segmentation graph first. Then several optimal candidate paths are selected from the graph by dynamic programming searching based on recognition confidence. Finally the best segmentation path is determined by matching these paths with the known post address database. In the experiment on more than 500 real envelop images with the this approach, the correct sorting rate of address recognition is up to 79.46% and that of address-postcode integrated recognition is up to 96.26%.

## 1. Introduction

With a continual increase of the postal business, the research on automatic postal mail sorting and its application has become a prospective field in China. Up to now, postal mails are mainly sorted by means of postcode recognition because mail address recognition is highly difficult and its accuracy cannot be assured. But the performance of this postcode-based mail sorting technology is not satisfactory since the using and writing of postcode on envelopes in China are not normative. With the improvement in handwritten Chinese character recognition owing to further

research, mail address recognition has become possible and can be applied to automatic mail sorting system.

In the mail sorting system with mail address recognition as an integral part, character segmentation is a crucial step in the process of address recognition. At present the accuracy of single handwritten Chinese character recognition has reached a high level whereas the segmentation of handwritten Chinese character becomes an obstacle to the application of this technology. Handwritten Chinese character segmentation in mail sorting system is a difficult step for the following reasons:

- (1) The structure of Chinese characters varies heavily, including the number of components (radicals), each of which may be a real Chinese character itself like “唱” and “胡”, the position relationship of components like left-right structure, top-bottom structure and in-out structure.
- (2) Handwritten characters in real mail address are unrestrained and in the different writing style, so the font size, the internal (between-radical) gap and between-character gap vary greatly.

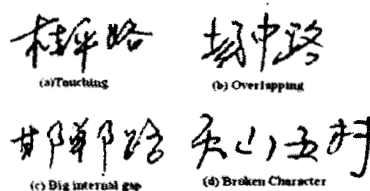


Figure 1. Some samples of tough character segmentation cases: (a) Touching characters (b) Overlapping characters (c) Big internal (between-radical) gap inside a single character (d) broken character

- (3) Overlapping, touching, separating characters in handwritten mail address are major factors that

can result in segmentation errors. Figure 1 shows some difficult cases in the character segmentation process.

At present there are mainly two categories of method of handwritten Chinese character segmentation [3,7]. One is based on topology structure and shape information such as the width and height of a character, the gap between adjacent characters, the stroke structure, etc [1,4,9,10]. Common methods in this category include vertical projection profile analysis, connected component analysis, stroke analysis, contour shape analysis, etc. The other is recognition-based segmentation in which all the possible segmentation results are recognized and the optimal one is determined according to the confidence of recognition result [2,5,6,8]. Algorithms as lexicon-driven algorithm [6], recursive segmentation algorithm [2], hidden Markov Model based algorithm [8], etc., fall in this category.

This paper proposes a two-stage character segmentation algorithm according to the characteristics of handwritten mail address characters. In the simple segmentation stage, the block sequence is extracted from the mail address image using the structure-based methods, including projection profile analysis, connected components extraction and stroke cross number analysis. In the second stage, all candidate segmentation paths are created by combining the neighboring blocks and represented with a candidate segmentation graph first. Then several optimal candidate paths are selected from the graph by dynamic programming searching based on recognition confidence and the best segmentation path is finally determined by matching these paths with the known post address database.

The rest of this paper is organized as follows: Section 2 describes the details of the algorithm. Section 3 gives experimental results and Section 4 makes concluding remarks.

## 2. Two-stage handwritten character segmentation in mail address recognition

### 2.1 Extraction of basic block sequence

The algorithm first expects to extract a basic block sequence from the mail address by means of several structure-based segmentation methods. Each block here may be a single character or a part of a Chinese character. Considering that there is a block-combining process in the following step, as many basic blocks as possible should be segmented in this step.

#### 2.1.1 Extraction of blocks based on vertical projection profile

First a basic block sequence is extracted from the binary mail address image with the vertical projection profile analysis. The projection profile can be obtained by counting the number of black pixels in a column or row. The average height of the mail address is computed according to the horizontal projection profile and denoted as  $LH$ . Then the mail address image is cut at white gaps of vertical projection profile. The basic blocks sequence thus obtained is denoted as  $\{b_1, b_2, \dots, b_N\}$  and the width of the block  $b_i$  is denoted as  $w_i$ . With the help of projection profile method, the non-overlapping and non-touching part of the mail address image can be segmented easily and quickly.

#### 2.1.2 Segmentation of overlapping characters based on connected component analysis

After the basic block segmentation with projection profile analysis, the wider blocks in the sequence are analyzed with the connected component analysis method to solve the problem of overlapping characters segmentation. The process can be divided into the following steps:

- (1) Extraction of connected components from the wider block image: Check the width of each block in the sequence. If the width of the block  $b_i$  meets  $w_i > T_w \cdot LH$ , where  $T_w$  is a predefined Width/Height threshold, then the connected components from this block image are extracted. The bounding box coordinate of each connected component  $CC_i$  is denoted as  $(l_i, r_i, t_i, b_i)$ .
- (2) Combining of connected components: In a handwritten mail address line, the two connected components overlapping in the horizontal direction are likely to compose one Chinese character. To decide whether or not the two neighboring components should be combined, the overlapping degree of these two neighboring connected components is relied on. Denote the bounding box of two neighboring connected components as  $cc_1(l_1, r_1, t_1, b_1)$  and  $cc_2(l_2, r_2, t_2, b_2)$ , and assume  $l_1 < l_2$  (ordered from left to right). Then the overlapping degree  $p_{overlap}$  is computed as

$$p_{overlap} = \left( \frac{r_1 - l_2}{r_1 - l_1} + \frac{r_1 - l_2}{r_2 - l_2} \right) / 2.$$

If  $p_{overlap} > T_{overlap}$ , where  $T_{overlap}$  is a predefined threshold, the two connected components are combined.

- (3) Update of the block sequence: If the number of connected components in one block is more than 2, delete this block from the sequence and then insert each connected component into the sequence as a new block. The updated block sequence is still denoted as  $\{b_1, b_2, \dots, b_N\}$

### 2.1.3 Segmentation of touching characters based on stroke cross number analysis

Since the touching neighboring characters cannot be correctly segmented with the methods of projection and connected components analysis, further analysis of the block sequence by means of projection and stroke cross number analysis is applied in this step to the case of touching characters.

- (1) Recognition of the wider block: Check all the blocks in the sequence. If the width of the block  $b_i$  meets  $w_i > T_{MaxW} \cdot LH$ , where  $T_{MaxW}$  is a predefined Width/Height threshold, then use the handwritten character recognition model to recognize this block. The recognition confidence of block  $b_i$  is denoted as  $c_i$
- (2) Projection and stroke cross number analysis: If the recognition confidence of block  $b_i$  meets  $c_i < T_{CF}$ , where  $T_{CF}$  is a predefined recognition confidence threshold, compute the vertical stroke cross number  $N_c(x)$  in this block image, which is the stroke number penetrated by a vertical line at point  $x$  in the block image. Then select the proper segmentation points from the set  $\{x | N_c(x) < 2 \text{ and } VP(x) < T_{vp}\}$ , after the insertion of which the width of each block meets  $T_{MinW} \cdot LH < w_i < T_{MaxW} \cdot LH$ .  $VP(x)$  is the vertical projection profile at point  $x$ .  $T_{vp}$  is a predefined project profile threshold,  $T_{MaxW}$  and  $T_{MinW}$  are predefined Width/Height thresholds.
- (3) Update of the block sequence: If block  $b_i$  is segmented into several new segments at the segmentation points, delete  $b_i$  from the sequence

and insert the new segments into the block sequence as new blocks.

## 2.2 Best segmentation path selection from the basic block sequence

### 2.2.1 Representation of candidate segmentation paths in a candidate segmentation graph

After the above block segmentation processes, the mail address image is represented in a sequence of image blocks  $\{b_1, b_2, \dots, b_N\}$ , which are ordered from left to right. The set of segmentation points that separate the block sequence is denoted as  $\{s_0, s_1, \dots, s_N\}$ . Since one or more consecutive blocks may compose one Chinese character, a candidate character pattern composed of  $k$  consecutive blocks between the segmentation points  $s_{i-1}$  and  $s_{i+k-1}$  is denoted as  $\langle s_{i-1}, \dots, s_{i+k-1} \rangle$ . All the consecutive candidate character patterns compose a candidate character segmentation path and all the possible segmentation paths are represented in a candidate segmentation graph, in which each segmentation point can be seen as a node and each edge as a candidate character pattern. Then all the possible segmentation paths in the mail address are represented as the paths from the starting node  $s_0$  to the ending node  $s_N$  in the graph. Figure 2 shows an example of candidate segmentation graph.

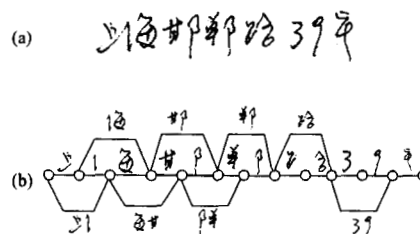


Figure 2. An example of candidate segmentation graph (a) original address image (b) candidate segmentation graph including all the possible candidate segmentation paths, each node in the graph representing a segmentation point and each edge representing a candidate character pattern.

To reduce the number of candidate character pattern combinations and improve the efficiency of algorithm, the following factors are considered to prohibit some neighboring blocks from combining if they meet some limitation requirement:

- (1) The number of consecutive blocks: In view of the complexity of Chinese character structure and the small noise factors, one to at most four consecutive blocks can be combined into a candidate character pattern.
- (2) The width of candidate character patterns: If the bounding box width of several consecutive blocks  $w_p$  meets  $w_p > T_w \cdot LH$ , where  $T_w$  is a predefined Width/Height threshold, these neighboring blocks are prohibited to combine
- (3) The gap between two neighboring blocks: Define  $g_M$  as the maximum of the gaps between every two neighboring blocks among all the blocks that may compose a candidate character pattern. If  $g_M > T_{gap} \cdot LH$ , where  $T_{gap}$  is a predefined Width/Height threshold, these blocks are prohibited to combine.
- (4) Recognition confidence of candidate patterns: If the recognition confidence of a candidate character composed of several consecutive blocks is lower than a predefined threshold, viz.  $p_{CF} < T_{CF}$ , where  $p_{CF}$  is the recognition confidence of the candidate patterns and  $T_{CF}$  is a predefined threshold, these blocks are prohibited to combine.

### 2.2.2 Selection of optimal candidate paths based on recognition and dynamic programming

The set of candidate segmentation paths in the candidate segmentation graph can be represented as  $CSP = \{sp_1, sp_2, \dots, sp_M\}$ , where  $sp_i$  is a candidate path and  $M$  is the number of the segmentation paths. Let  $w_{ij}$  be the weight from node  $i$  to node  $j$  and  $E(sp_k)$  be the set of all the edges on the path  $sp_k$ . The total weight of a candidate path  $sp_k$  in the graph can be determined as follows:

$$W(sp_k) = \sum_{\langle i, j \rangle \in E(sp_k)} w_{ij}$$

The recognition confidence of each candidate pattern, after recognized with single handwritten character recognizer, becomes the weight of the corresponding edge of this candidate pattern in the candidate segmentation graph. In this paper, 20 optimal candidate segmentation paths are selected by searching the 20 longest paths in the candidate segmentation graph with dynamic programming method.

### 2.2.3 Selection of the best segmentation path

Since the accuracy rate and the confidence of handwritten character recognition are not so high as those of printed character recognition, it may not be extremely reliable and efficient to retrieve the optimal segmentation paths by character recognition confidence alone. More knowledge is needed to further confirm the best segmentation path from these 20 optimal paths. As this algorithm is applied to the mail sorting system, the city postal address database, which stores the names of all the districts and streets of the city, is adopted as the knowledge database in the process of searching for the best segmentation path. The matching degree between the recognition result of the optimal candidate path and the postal address database is used to evaluate the 20 candidate optimal paths and the path with the highest matching degree is selected as the best segmentation path.

## 3. Experiments

Up to now the character segmentation algorithm cannot be evaluated and tested very accurately because there is no standard automatic method to test and evaluate the performance of the character segmentation algorithm. For the small sample test, the character segmentation algorithm is usually evaluated with the help of manual statistics; for the large sample image test, the final recognition accuracy of the system applied is often used to evaluate the performance of the segmentation algorithm. In this paper, the second method is adopted.



Figure 3. The segmentation results of some mail address images with proposed algorithm

The above segmentation algorithm of handwritten Chinese characters in mail addresses is applied to the mail sorting system based on postcode and mail address recognition. The whole system captures the

gray mail images from the mail-sorting machine, and extracts the mail address images with the binary process and layout analysis. Then the algorithm is used to segment the mail address images into characters and then output the recognized address information. The recognized mail address and postcode are used to sort mails and the correct sorting rate is used to evaluate the algorithm. The test sample set includes 589 envelope images captured from mail sorting machine. Table 1 shows the test result. Figure 3 shows the segmentation result of some mail address samples.

The number of test envelope images	589
The number of correctly sorted mails with address recognition alone	468
The correct sorting accuracy with address recognition alone	79.46%
The number of correctly sorted mails with both address recognition and postcode recognition	567
The correct sorting accuracy with both address recognition and postcode recognition	96.26%

**Table 1. The test result of the mail sorting application using the character segmentation algorithm proposed in this paper**

#### 4. Conclusion

In this paper, a two-stage handwritten Chinese character segmentation approach in mail address recognition is proposed. This approach is applied to the automatic mail sorting system. First a block segmentation process is fulfilled using the structure-based methods to extract the block sequence from the mail address image. Next all candidate segmentation paths are created by combining the neighboring blocks and represented in a candidate segmentation graph. Then several candidate paths are selected from the graph by dynamic programming searching based on recognition confidence and the best segmentation path is determined with the help of the knowledge analysis of the known post address database. An experiment with this algorithm was carried out on more than 500 real envelop images, with the correct sorting rate of address recognition up to 79.46% and that of address and postcode integrated recognition up to 96.26%.

In the test of this algorithm, the segmentation result of touching digits in the mail address is not satisfactory. Sometimes the touching digits are wrongly recognized as a Chinese character, such as the case in Figure 3(d). In the future, a special module to solve the problem of touching digits recognition may be added to further improve the performance of the mail sorting system.

#### References

- [1] S. Ariyoshi, "A Character Segmentation method for Japanese Printed Documents Coping with Touching Character Problems", Proceedings of 11th International Conference on Pattern Recognition, Ten Haag, The Netherlands, 1992, Vol.2, pp.313-316
- [2] R. Casey, G. Nagy, "Recursive Segmentation and Classification of Composite Patterns", Proceedings of the 6th International Conference on Pattern Recognition, Munich, Germany, 1982, pp.1023-1026.
- [3] R. Casey, E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", IEEE Trans. on Pattern Recognition and Machine Intelligence, 1996, Vol.18, No.7, pp.690-706.
- [4] Y. Kobayashi, K. Yamada, J. Tsukumo, "A Segmentation Method for Handwritten Japanese Lines Based on Transitional Information", Proceedings of 11th International Conference on Pattern Recognition, Ten Haag, The Netherlands, 1992, pp. 487-491.
- [5] M. Koga, T. Kagehiro, H. Sako, et al., "Segmentation of Japanese Handwritten Characters Using Peripheral Feature Analysis", Proceeding of 14th International Conference on Pattern Recognition, Brisbane, Australia, 1998, Vol.2, pp.1137-1141
- [6] C. L. Liu, M. Koga, H. Fujisawa, "Lexicon-Driven Segmentation and Recognition of Handwritten Character Strings for Japanese Address Reading", IEEE Trans. on Pattern Recognition and Machine Intelligence, 2002, Vol. 24, No.11, pp.1425-1437
- [7] Y. Lu, "Machine printed character segmentation - an overview", Pattern Recognition, 1995, Vol.28, No.1, pp.67-80.
- [8] Y. H. Tseng, H. J. Lee, "Recognition-based Handwritten Chinese Character Segmentation Using a Probabilistic Viterbi Algorithm", Pattern Recognition Letters, 1999, Vol. 20, No.8, pp. 791-806.
- [9] T.Yamaguchi, S.Tsuruoka, T.Yoshikawa, et al., "A segmentation System for Touching Handwritten Japanese Characters", Proceedings of IWFHR'02, Ontario, Canada: IEEE Computer Society, 2002, pp.407-412.
- [10] T. Yamaguchi, T. Yoshikawa, T. Shinogi, et al., "A Segmentation Method for Touching Japanese Handwritten Characters Based on Connecting Condition of Lines", Proceedings of ICDAR'01, Seattle, WA, USA: IEEE Computer Society, 2001, pp. 837-841.