A Large Scale Japanese Handwritten Address Recognition System Using Rough and Fine Classification

Nei KATO, Kazutaka Tokumoto, Yoshiaki NEMOTO
Tohoku University
Sendai, 980-8579,Japan
Email: kato@nemoto.ecei.tohoku.ac.jp

Masato SUZUKI
Tokyo National College of Technology
Tokyo, 193-8610, Japan
Email: suz@pr.tokyo-ct.ac.jp

Abstract

Post-preprocessing is an important part in handwritten character recognition system. In post-preprocessing, the knowledge about address is used to verify the result of segmentation and recognition. As a result, the best-matched output of segmentation and recognition is selected. The main problem in post-preprocessing is there are too many addresses needed to be handled, so that the processing time is tremendously long. In this paper, we propose a post-processing technique with rough and fine classification. Especially in rough classification, we introduce the priority allotting method(PAM) for selecting address candidates with accuracy and speed. We show the effectiveness of this method and the recognition system constructed by applying the proposed method.

1 Introduction

Since the document recognition system can realize user friendly interface between human and computer, many research have been employed in recent years. In Japan, the handwritten address recognition technology is a hot topic. For maintaining postal service efficiently, parceling machine with handwritten recognition system becomes much important these days. According to statistics, more than 40 billion postal items were distributed last year, and there is an increase of 20% every year. Especially at the end of December of every year, millions of Japanese new year greetings, most of them with addresses written by hand, are sent out almost in two weeks, all of these cards must be delivered before the coming new year.

With the increasing computational power of computers, recognizing handwritten address in real time has become realistic and some research have been done[1, 2, 3, 4, 5]. However, due to the more than 3,000 Japanese characters(KANJI and KANA) used in daily life in the first class of the Japanese Industrial Standard(JIS), constructing a handwritten address recognition system with high accuracy is a challenging problem. Furthermore, the separate characters and touching characters make the problem complicated.

To construct a handwritten address recognition system with high accuracy, recognition precision with respect to individual characters is critical. However, since the shapes of characters in handwritten style vary greatly, the recognition rate even with the handwritten recognition system with highest accuracy in Japan, can only reach around 80%[6]. This poor ability of individual character recognition limits the performance of address recognition system significantly.

To cope with this problem, we propose a hand-

0-7803-5747-7/00/$10.00©2000 IEEE.

written address recognition system with rough and fine classification for post-processing. The characteristic part of this system is its rough classification algorithm called priority allotting method. By using this technique, number of addresses needed to be taken into consideration in fine classification can be reduced largely.

The recognition experiments have been conducted with the largest handwritten address database in Japan, which offered by Institute for Posts and Telecommunication Policy, Ministry of Posts and Telecommunications[7]. The experimental results have shown the system have achieved a very high recognition rate.

The rest of the paper is organized as follows. Section 2 describes the whole recognition system briefly. In section 3, the details of the proposed method is explained. In section 4, the experimental results are shown, and the effect of the proposed system is discussed.

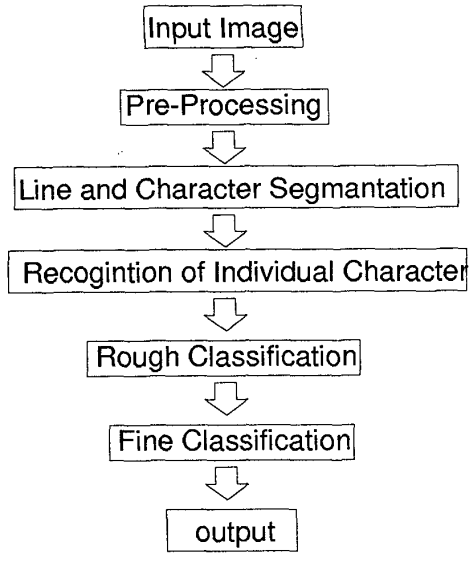

Figure 1: Scheme of handwritten address recognition system

2 Handwritten Address Recognition System

The construction of handwritten address recognition system is shown in Fig.1.

The recognition object is the part except street numbers of vertical writing. Firstly, noise is removed. Then, lines are extracted according to histogram of black pixels. Basic circumrectangles are drawn around connected black pixels. The average vertical length $\hat{L}$ and the standard deviation σ of the basic circumrectangles are calculated. The basic circumrectangles those with longer length than $(\hat{L}+\sigma)$ are segmented according to the method of [8]. Finally, the segmented circumrectangles are checked whether they need to be combined with the following two conditions, $S \leq (\hat{S} + \sigma_s)$, and $L_{int} < (L_{max} \times 1.5)$, where S is an interval between circumrectangles, $\hat{S}$ is the average and σ_s is the standard deviation. L_{max} is the max of lengths of circumrectangles, and L_{int} is the length of combination of any two circumrectangles. An example of segmentation is shown in Fig.2.

For the image in each circumrectangle, individual character recognition is conducted. ETL9B is the largest handwritten character database in Japan, it includes 3,036 kinds of character. There are two hundred samples per character. For the frequently-used one hundred kinds of characters in addresses, considering the features of handwritten addresses, samples written in writing brushes, sign pens or ball-point pens are collected to make the standard patterns. For the other 2,936 kinds of characters, the standard patterns are calculated from the samples in ETL9B.

Rough classification is a process of selecting address candidates from address database for an input image using the recognition results of circumrectangles. Rough classification will be explained in details next section.

In fine classification, output addresses of rough classification is used. The best-matched address will be selected finally as first candidate.

3 Priority allotting method

3.1 How to make character table

Considering that address may not be written from prefecture, we expanded the whole address

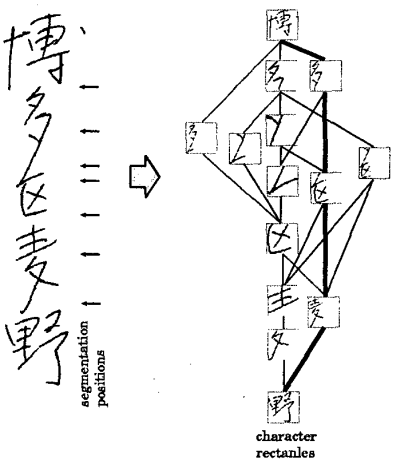

Figure 2: Example of character segmentation

set into 632,942 including addresses written from city or village. As shown in Fig.3, each address is given a specific ID. Then according to this index, we generate another table for priority allotment. As shown in Fig.4, all characters of address are registered by using their ID number and position appeared in addresses.

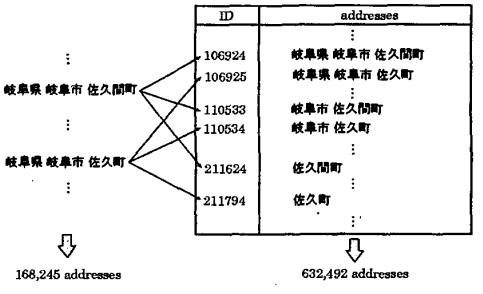

Figure 3: Structure of address dictionary

3.2 Priority allotment

Priority allotment is organized as follows.
[step 1] set the initial point of each address candidate to zero
[step 2] refer to character table shown in Fig.4, if the character is registered in the table, set the flag of this rectangle to 1, otherwise, to 0. Note that the recognition result used here is the top 10 candidates.
[step 3] count up the flags for each address candidate as F_i, where i indicates the address number, priority of each address P_i is calculated as follows.

character	ID	position
岐	106924	1
	106925	1
	106924	4
	106925	4
	110533	1
	110534	1
佐	106924	7
	106925	7
	110533	4
	110534	4
	211624	1
	211794	1

Figure 4: Structure of character table

$$P_i = \frac{\sum_{j=1}^{length(i)} F_i(j)}{length(i)} \quad (1)$$

where j is the length of each address candidate.

An example is shown in Fig.5, recognition results of image of character "麦" are "麦, 素, 章...", this character can take the position for 4, 5, and 6. in table. As the result, "麦" in address "博多区麦野" and "章" in address "佐賀市成章町" had the flag set to 1. After applying this process to all rectangles, "博多区麦野", the correct one, had 4 flags among 5 rectangles, this equals to priority 0.8. The other one "佐賀市成章町" had only 1 flag among 6 rectangles. This gave the priority to 0.17. In our system, threshold 0.5 is used to select the high priority ones. By this means, the number of addresses needed to conduct pattern matching in fine classification can be reduced largely.

4 Experimental Results

Recognition experiments are carried out with the largest handwritten address database in Japan,

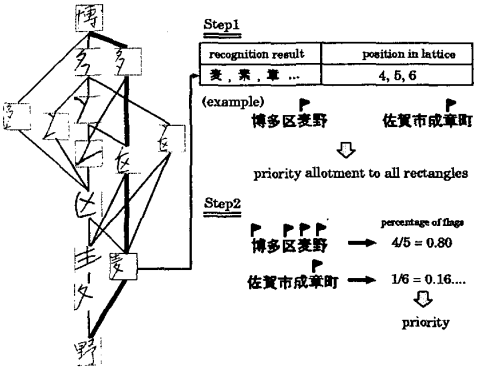

Figure 5: Example of applying priority to address candidates

Table 1: Recognition results

	rough	fine	total
Correct	1,539	103	1642 (89.2%)
Error	8	5	13 (0.7%)
Reject	169	16	185 (10.0%)

offered by Institute for Posts and Telecommunication Policy, Ministry of Posts and Telecommunications. 1,840 handwritten addresses samples are used as test samples. The dictionary size is 632,942.

Experimental results of the proposed method are shown in Table 1. This is a very high recognition rate for this database. Also since the proposed method is effective, the number of addresses reduced by rough classification is more than 99%.

5 Conclusion

In this paper, we proposed a new post-processing technique with rough and fine classification. Especially in rough classification, we introduce priority allotting method(PAM) for selecting address candidates with accuracy and speed. We showed the effectiveness of this method and the recognition system constructed by applying the proposed method.

References

[1] J.Higashino, H.Fujisawa, Y.Nakano and M.Ejiri: "A Knowledge-based Segmentation Method for Document Understanding", Proc. of 8th ICPR, pp.745-748, 1986.

[2] K.Banno, T.Kawamata, K.Kobayashi and H.Nambu: "Text Recognition System for Japanese Documents", Proc. of 9th ICPR, pp.176-180, 1988.

[3] H.Murase: "Online Recognition of Free-format Japanese Handwritings", Proc. of 9th ICPR, pp.1143-1147, 1988.

[4] T.Fukushima et al.: "A Word-sequence Search Algorithm for A hand-written Character Reader", Trans. Information Processing Society of Japan, vol.37, No.4, 1996[in Japanese].

[5] E.Ishidera, D.Nishiwaki and K.Yamada: "Unconstrained Japanese Address Recognition Using a Combination of Spatial Information and Word Knowledge", Proc. of 4th ICDAR, pp.1016-1022, 1997.

[6] M.Suzuki, N.Kato, H.Aso, and Y.Nemoto: "A Hand-printed Character Recognition System Using Image Transformation Based on Partial Inclination Detection", IEICE Trans. on Information & System, Vol.E79-D, No.5, pp.504-509,1996.

[7] F.Kawamata, T.Wakahara, T.Matsui, T.Noumi, I.Yamashita, T.Tshutsumida: "The Results of the Third IPTP Character Recognition Competition for Handwritten Kanji Characters on Post Cards", Proc. of 1994 IEICE autumn conf. D-321, 1994.

[8] H.Ino, K.Saruta, N.Kato and Y.Nemoto: "Handwritten Address Segmentation Algorithm Based on Stroke Information," Transactions of Information Processing Society of Japan, Vol.38, No.2, pp.280-289, 1997.