A Method for Determining Address Format in the Automated Sorting of Japanese Mail

T. Tsuchiya, N. Nakajima and T. Kamimura

Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216 JAPAN

Abstract. This paper presents a new method for determining the address format used on mail that has been hand-written in Japanese, a task made particularly difficult by the variety of formats possible in Japanese. We classify possible formats into six types, and identify the distinguishing features of each. In the proposed method, features characterizing any of the six types are identified for a given address, and from this a list of format-candidates is generated. Character lines are then determined for any format candidate, and one candidate is subsequently selected on the basis of the statistical likelihood of any address being written in that particular way, given the location and size of the character lines. We also demonstrate the effectiveness of the new method experimentally.

1 Introduction

A lot of research has been done related to ways of automating mail sorting work at post offices to handle greater volumes of mail smoothly and to improve the postal service. In particular, there is a strong need to automize delivery sorting, i.e. sorting letters and postcards to delivery zones which are assigned to individual carriers. Many ideas for delivery sorting have been proposed[1][2].

Delivery sorting includes four processes, Address Block Finding (ABF), character segmentation, character recognition, and address recognition. Because ABF is the first step in the delivery sorting process, and because it is not easy to recover from ABF errors, it should be as stable and accurate as possible.

ABF detects macroscopic address information, i.e. the address format and the location of the address lines. Here, we define "address format" as the direction of address lines (vertical or horizontal) and the direction of the written characters. The location of an address line may be expressed in the form of the coordinates of a rectangle circumscribing that address line. All of this macroscopic address information is necessary for delivery sorting of Japanese mail because without it individual characters cannot be segmented out and recognized.

It is difficult to determine the address format being used, because Japanese mail has many format variations. One of the most common methods for determining format does so on the basis of the location and size of detected address lines. However, some classes of address formats are hard to determine by this method.

We propose an address format determination method that uses features that characterize Japanese address formats. First, we define specialized terms, then

point out the difficulties involved in format determination, and then descirbe our method. Finally, we present details of experiments with mail images which were carried out to demonstrate the effectiveness of our method.

2 Japanese Address Formats and The Difficulty in Determining Them

2.1 Japanese Address Formats

In an address recognition machine, letters and postcards are conveyed by a belt and their images are obtained when they run past scanner line sensors. The origin and the coordinates for a mail image are defined according to the running direction, as shown in Fig. 1. The figure also shows the top, bottom, forward and back positions relative to the mail image.

Addresses written in Japanese have many more variations than ones written in English. (For example, English is almost never written vertically.)

Japanese address formats are broadly classified in terms of two items, the postcard or envelope format (portrait or landscape) and the address line direction (vertical or horizontal) (See Fig. 2).

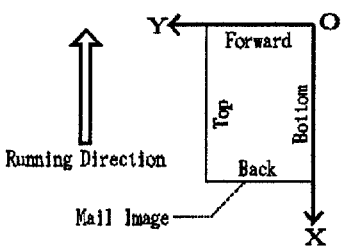

Fig. 1. Origin and Coordinates

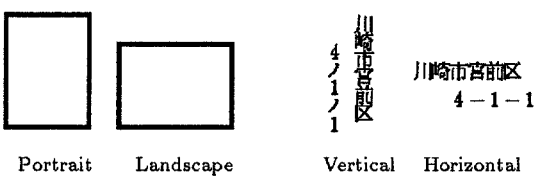

Fig. 2. Classification of Japanese mail

We thus defined six formats as shown in Fig. 3. The "vertical plus landscape" combination is not considered, because this combination is almost never used in actual mail. The term "normal" means that the characters are written toward the bottom or back edge and the term "reverse" means that they are written toward the top or forward edge.

2.2 The Difficulty in Determining Address Formats

In one conventional method, the address format is determined based on the location and size of the detected address lines. Specifically, it assumes that there are several formats as format candidates, and detects address lines for each format candidate and compares the likelihood scores of the detected lines. The score is calculated based on rank tables where approximate probability densities for the location and size of the address lines are listed. The tables are determined

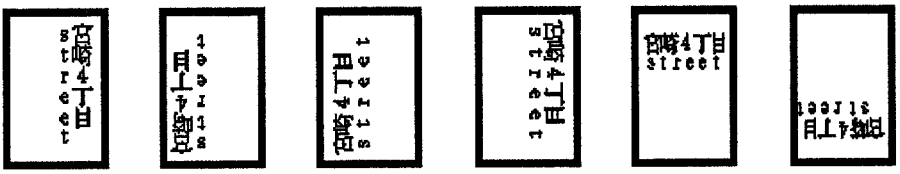

Fig. 3. Mail Format Definition

in advance based on mail image databases for each format[3]. The method is effective when the locations and sizes of address lines are for the most part fixed in each format.

However, some classes of address formats are hard to determine with this method. Figure 4 shows two examples of format determination errors that occur with this method.

Figure 4(a) shows VP_normal format mail and Fig. 4(b) shows HL_normal format mail. In Fig. 4(a), the sender's name was detected wrongly as an address line when the format candidate was HL_normal format. In a case like this, it is difficult to distinguish VP_normal format from HL_normal format with this method because both the location and size of the detected address line are almost identical.

This indicates that there are some types of mail for which it is difficult to determine address formats by using only the location and size of address lines.

In order to develop a better format determination process, it is necessary to use other information that is present in mail images.

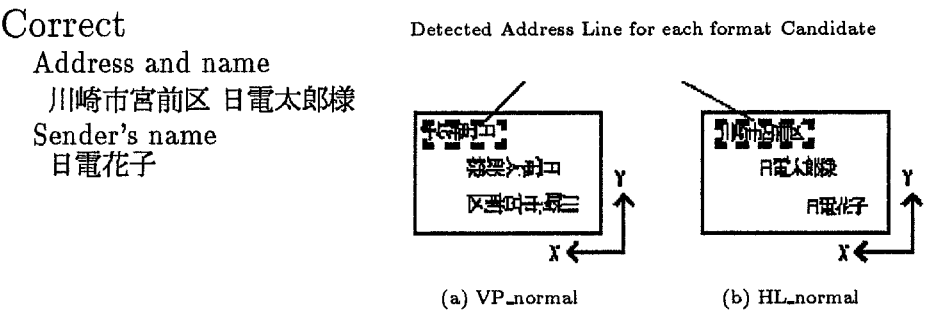

Fig. 4. Example of Difficult-to-Determine Format Using Only Detected Address Lines

2.3 Supplemental Properties of Japanese Handwritten Mail

To extract the features of Japanese address formats, we consider the supplemental properties of Japanese handwritten mail.

Japanese mail include elements other than address lines. There is almost always a stamp and there are frequently printed red boxes for entering the postal code. There are sometimes line segments as guidelines for the addrss. And sometimes there are printed character lines, for example the corporation name and address, near the edge. These may be printed within a box frame. Detecting special elements such as these is one effective way to determine the address format.

3 Format Determination Process Using Mail Features

3.1 Outline of Format Determination Process

The format determination is composed of two processes, Format Candidates Selection (FCS) and Final Determination.

FCS detects specific features on a mail image and generates format candidates. If FCS generates several format candidates, Final Determination selects the final format from the candidates by means of a conventional method using the location and size of the address lines.

To solve the problem described in Section 2, we introduce FCS into the format determination. FCS helps to make format determination more accurate because it makes it possible to determine address formats that cannot easily be determined by the conventional method. How this is done is explained in the next section.

3.2 Format Candidates Selection

FCS is mainly composed of two parts, Feature Detection and Selection Using Rules.

Feature Detection This consists of several feature detectors, such as stamps, red boxes, mail item length, direction of address lines, line segments, box frames, and printed character lines.

Selection Using Rules This generates format candidates using detected features. It has three rule-based processes as follows.

A. Conversion

This process converts detected features into attribute values. Examples of attributes are "Existence of stamps", "Existence of red boxes", and "Existence of pattern on forward-back position".

For example, the attribute "Existence of pattern on forward-back position" has five values: "box frame on forward position", "box frame on back position",

"printed character line on forward position", "printed character line on back position" or "not clear".

B. Matching

This process selects address formats that match each attribute value. For example, the value "printed character line on back position" can match the VP_normal and HP_normal formats in Japanese mail.

All the formats that match all the attribute values become format candidates as the result of the matching process.

C. Adjustment

This process adjusts conflicts in matching rules. Conflicts occur if no format candidates are left after applying the matching rules.

4 Experiment

We have developed a simulator of a handwritten address recognition system on EWS4800 (NEC workstation with MIPS risc chips). The proposed method was applied to 455 mail images by the simulator. We compared these simulator results with the correct results, which were prepared manually in advance. The format determination results are shown in Table 1. The total rate of correct format determination was 93.4%, confirming that the proposed method is effective.

Table 1. Results of Format Determination

		Results						Total	Correct rate
		VP_n	VP_r	HL_n	HL_r	HP_n	HP_r		
True	VP_n	274		6	1	4		285	96.1%
	VP_r							0	
	HL_n	7	1	49	3			60	81.7%
	HL_r	2			7	1		10	70.0%
	HP_n	5				95		100	95.0%
	HP_r							0	
								455	93.4%

Table 2 shows the rate of correct format determination that was calculated from the number of format candidates. The number is 1 or 2 and the two cases are mutually irreconcilable. Format determination is regarded as correct when the macroscopic address information, which is the ABF output, includes the correct address format.

The rate of correct determination was 95.8% when the number of format candidates was 1 and 91.4% when the number of format candidates was 2. This compares with 87.9% and 87.1% respectively for the conventional method.

Thus the proposed method yields better results than the conventional method. Its performance is significantly better when the number of format candidates is 1, in which case the format is determined only by FCS. Hence, it is clear that FCS is effective.

Table 2. Rate of Correct Format Determination by Number of Format Candidates

Number of Format Candidates	Number of Mail [percent]	Rate of Correct Format Determination by Proposed Method	Rate of Correct Format Determination by Conventional Method
1	306 [67.3%]	95.8%	87.9%
2	139 [30.5%]	91.4%	87.1%

5 Conclusion

Our address format determination method for mail sorting is more effective than the conventional method. It introduces Format Candidates Selection (FCS), which detects features that characterize each format, and generates format candidates using these features, before address lines are detected. This makes it possible to determine address formats that are hard to determine using only results of address line detection.

When the number of format candidates is 1, the rate of correct format determination is especially improved, proving that FCS is effective.

We would like to improve the ABF process by using the results of other processes such as character recognition and address recognition.

References

1. A. Gardin (CGA-HBS): A Real Time Address Block Location System for Handwritten Letter and Flat Mail. USPS Advanced Technology Conference 4 1992
2. Sargur N. Srihari, Ching-Huei Wang, Paul W. Palumbo and Jonathan J. Hull (State University of New York): Recognizing Address Blocks on Mail Pieces: Specialized Tools and Problem-Solving Architecture. AI Magazine, Winter (1987) 25–40
3. T. Ishikawa, Y. Nishijima, Y. Tsuji, I. Kaneko and T. Bashomatsu: Address Block Location and Format Identification of Japanese Address Reading Letter Sorting Machine. NEC RESEARCH & DEVELOPMENT Vol.33, No.2, April (1992) 217–225,
4. T. Bashomatsu: Address Block Location and Format Identification. USPS Postal Service Advanced Conference 5 (1992) 1295–1303