Address-Block Extraction by Bayesian Rule

Tatsuhiko Kagehiro, Masashi Koga, Hiroshi Sako, and Hiromichi Fujisawa
Central Research Laboratory, Hitachi, Ltd., Tokyo 185-8601, JAPAN
E-mail: kagehiro@crl.hitachi.co.jp

Abstract

A method for extracting a recipient address-block from a mail image has been developed. The method is composed of two steps: nomination of address-block candidates and evaluation of these candidates by using the Bayesian rule according to each of address-block type. Accordingly, the proposed method can cope with various types of address-blocks. The effectiveness of the method was confirmed in several address extraction experiments. These experiments show that the top-five extraction results include one correct address-block in 94% of total number of printed-mail cases and 89% in of handwritten-mail cases.

1. Introduction

In order to read a recipient address of mail correctly, it is firstly necessary to accurately extract recipient address lines. Because the lines of a recipient address of Japanese mail are written in various ways, extracting them is very difficult. Moreover, the existence of the sender’s address lines and advertisements on pre-printed mail sometimes makes the extraction much more difficult. In this paper, we call an area occupied by the recipient address lines the address-block.

There are several conventional methods of extracting the address-block [1...5]. For example, in the case of handwritten-mail, there is a technique of estimating the address-block type by analyzing the positional relationship between character lines and ruled lines [1]. When the address-block has ruled lines, this method is very useful, but it essentially has no power to extract an address-block without ruled lines. In the case of large-sized mails, another technique extracts an area having a uniform grey value is extracted as an address-block [4,5]. This technique is effective to extract an address-block on attached address label. As for these techniques, it is not enough to apply them to various specific types of address-blocks only.

The proposed method, that will be described later, can be adapted to various address-block types. This method extracts several candidates of the address-block by using multiple hypotheses on types and sorts them according to a confidence value. To calculate the confidence value of a candidate, an evaluation based on a Bayesian rule is executed by using a type-hypothesis. Address-block candidates are read by address character recognition one by one until one candidate is accepted as an address, the result is then regarded as the recipient address of the mail. This recognition procedure can effectively read the recipient address of mail [6].

2. Variation of Address-block Type in Japanese Mail

There are various types of address-block in Japanese mail. We have classified address-blocks into several types by human observation. That is, the type of mail envelope is either portrait or landscape, the direction of the written address is either vertical or horizontal, the address is either written in printed or handwritten. Note that a landscape mail with vertical writing is very uncommon. The total number of types is therefore six and the abbreviation of each type is defined as follows:

- P-PV : (Printed, Portrait, Vertical writing)
- P-PH : (Printed, Portrait, Horizontal writing)
- P-LH : (Printed, Landscape, Horizontal writing)
- H-PV : (Handwritten, Portrait, Vertical writing)
- H-PH : (Handwritten, Portrait, Horizontal writing)
- H-LH : (Handwritten, Landscape, Horizontal writing)

The six types of address-blocks are shown in Table 1. We assume that a mail image is scanned in portrait orientation so that the stamp is located in the upper left. The resolution of the image is 200 dpi.

The rectangles in Table 1 show the address-blocks. It is difficult for a simple method to extract the address-block, because its position and area are different in the case of each type (as shown in Table 1). Since the type of address-block cannot be known in advance, it is not possible to adjust the extraction process for each type. Moreover, a sender address and pre-printed advertisements on the mail sometimes make address extraction even more difficult.
3. Mail Recipient Address Reading System

To read a recipient address of mail, two operations must be performed: extraction of the address-block and the reading of the address lines in the address-block. The address-block extraction should be a bottom-up approach, because it is performed without the knowledge of the address-block type. The address-block extraction is composed of the following four steps:

1. Removal of noise by pre-processing
2. Extraction of objects
3. Creation of candidates of address-block
4. Evaluation and sorting of the address-block candidates

Figure 1 shows the flow of these steps. In the pre-processing step, noise and underlines are removed from the input binary image by a simple image processing. Next, connected components are derived from the image. All processing after this step is performed using connected components.

Table 1: Variation of Address-block Types

<table>
<thead>
<tr>
<th></th>
<th>PV</th>
<th>PH</th>
<th>LH</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td><img src="PV_P.png" alt="Image" /></td>
<td><img src="PH_P.png" alt="Image" /></td>
<td><img src="LH_P.png" alt="Image" /></td>
</tr>
<tr>
<td>H</td>
<td><img src="PV_H.png" alt="Image" /></td>
<td><img src="PH_H.png" alt="Image" /></td>
<td><img src="LH_H.png" alt="Image" /></td>
</tr>
</tbody>
</table>

In the object extraction step, the objects (a stamp, a postal code, character lines, etc.) that may commonly exist in the mail image are extracted. The extraction uses common knowledge of all address-block types such as the knowledge of the position and the size of each object.

In the address-block candidate creation step, several character lines extracted in the object extraction step are combined, and several candidates of the address-block are generated. This step creates several proper candidates for each mail type, because the position and the area of the correct address-block are different for each type. In the next step, one correct answer is selected from the candidates even if there is no knowledge of the address-block types.

In the address-block candidate evaluation step, a confidence value expressing the likelihood that a candidate is the correct address-block is calculated according to a Bayesian rule. The confidence value is calculated by using features detected from the candidate. One candidate has several confidence values corresponding to all address-block types. The number of confidence values for one candidate is therefore the same as that of types. Next, the candidates are sorted in order of their maximum confidence values.

In the address-lines reading step, the sorted candidates are analyzed one by one by character recognition. When a recognition result is an address, it can be regarded as the recipient address of the mail.

In the next chapter, the candidate evaluation using Bayesian rule is described somewhat in detail.

4. Address-Block Evaluation by Bayesian Rule

4.1 Calculation of Confidence Value

In the address-block candidate evaluation step, the confidence value is calculated from the following features that can be observed from the candidate.

1. Averages of height and width of character lines
2. Variances of height and width of character lines
3. Area of address-block candidate
4. Position of the candidate

Since the standard values of the features change according to the address-block type, a single set of parameters for the candidate cannot be determined. To cope with this problem, several address-block dictionaries are prepared according to each type as shown in Figure 2. The optimal confidence values of all candidates in each type can thus be calculated. However, because the address-block type is unknown in advance, a dictionary corresponding to a certain type cannot be selected. Accordingly, an address-block candidate for each of type
is assumed, and a confidence value for each type is calculated independently as each dictionary is selected and referenced. Several confidence values per candidate are thus computed and are compared each other to select maximum one among them. This maximum value is regarded as a representative confidence value for the candidate, because the confidence value computed from the dictionary of the corresponding type should be the highest and all confidence values for an incorrect candidate should be low.

4.2 Address-block Dictionary and Bayesian Rule

A Bayesian rule is applied to calculate the confidence value for the address-block candidate. The confidence value is calculated using the candidate’s features according to the Bayesian rule given following formula:

$$P(H_c | e_1, e_2, e_3, ... e_n) = \frac{P(H_c) \prod_{k=1}^{n} L(e_k | H_c)}{1 + \frac{P(H_H) \prod_{k=1}^{n} L(e_k | H_H)}}$$

$$L(e_k | H_c) = \frac{P(e_k | H_c)}{P(e_k | \overline{H_c})}$$

where $c$ is address-block type, $H_c$ is the hypothesis of correct candidate for type $c$, $e_1, e_2, e_3, ... e_n$ are the observed features of a candidate, $P(e_k | \overline{H_H})$ is a posterior probability of $e_k$ supposing hypothesis $H_H$, and $L(e_k | H_H)$ is the likelihood ratio for $e_k$ supposing hypothesis $H_H$.

We define $P(H_{H_c} | e_1, e_2, e_3, ... e_n)$ as a confidence for a candidate, where $e_1, e_2, e_3, ... e_n$ are observed features. To obtain $P(H_{H_c} | e_1, e_2, e_3, ... e_n)$, likelihood ratio $L(e_k | H_{H_c})$ is required in advance. The likelihood ratio is estimated from learning data, and is stored as shown in Figure 2. It is referenced while the confidence values are calculated. Besides, the likelihood ratio should be calculated according to each type independently. The learning data for each type of address-blocks is therefore collected from real mail images in advance.

To estimate the likelihood ratio $L(e_k | H_H)$, $P(e_k | H_H)$ and $P(e_k | \overline{H_H})$ must be known. $P(e_k | H_H)$ is a posterior probability of $e_k$ supposing hypothesis $H_H$. We consider that a posterior probability can be estimated generally from the occurrence frequency of $e_k$ in the case of $H_H$. Thus, to estimate $P(e_k | H_H)$ and $P(e_k | \overline{H_H})$, in the address-block candidate creation step, lots of address-block candidates are created from a leaning set of mail images. Secondly, all candidates are judged to be correct or incorrect by human observation. Features of candidates are extracted from the above address-block candidates, and the frequencies of features in correct and incorrect candidates are created respectively. These frequencies are used to calculate probabilities, $P(e_k | H_H)$ and $P(e_k | \overline{H_H})$.

Note that $P(H_H)$ and $P(\overline{H_H})$ are prior probabilities of $H_H$ and $\overline{H_H}$, and they can be obtained by calculating the ratios of the number of correct and incorrect candidates to the total number of candidates, respectively.

5. Experimental results

5.1 Evaluation of Address-block Extraction

To evaluate the proposed address-block extraction method, we measured extraction accuracy of the address-block. Two datasets were prepared: a set of 500 pieces of printed-mail and a set of 500 pieces of handwritten-mail. The number of learning samples is 1341, including both types of mail. This dataset is different from the evaluation dataset as mentioned above. Two methods were compared in terms of extraction accuracy of address-block: (A) a conventional heuristic method and (B) the proposed Bayesian rule method.

In method (A), the parameter set is adjusted heuristically based on according to the learning data. In
the evaluation of both, the type of the test mail was not known in advance. We compared the cumulative extraction rate for the address-block in mails of the P-LH and H-PV type mail.

Figure 3 shows the cumulative extraction rates for the two methods. Cumulative extraction rate is calculated from frequency of candidate ranking including correct ones. Hereinafter, a cumulative extraction rate form the top to 5th candidate is called a top-five correct rate.

In the case of the P-LH type, the top-one correct rate for method (B) was improved by 16pt up to 79% in comparison with that for method (A). The top-three, top-four, top-five correct rates for methods (A) and (B) became equivalent, that is, 95%.

In the case of type H-PV, the top-one correct rate for method (B) was improved 11pt up to 90% in comparison with that for method (A). The top-five correct rate for (B) is still larger, that is, 91%, than for method (A).

A general experiment with all types of mail confirms that the proposed method can extract the top-five candidates including the correct address-block in 94% of printed-mail cases and 89% in handwritten-mail cases.

5.2 Evaluation of Address Reading

We measured the read rate of a recipient address by using the proposed method. The number of test mail sample was 13778 for printed-mail and 7265 for handwritten-mail. In the recipient address reading system mentioned in Chapter 3, the read rates of two systems using method (A) and (B) respectively were compared.

In the case of the printed-mail, the read rates for methods (A) and (B) were equivalent, the rate was 84.5%. In the case of the handwritten-mail, the read rate was improved by 0.61pt to 67.5%.

As shown by this evaluation, the improvement in read rate improvement by using method (B) for the handwritten-mail is larger than that for the printed-mail. This is because the parameter set for method (A) is heuristically adjusted for printed-mail but not for the handwritten-mail. If we use a parameter set adjusted for handwritten-mail, this improvement for handwritten-mail would be larger than that for printed-mail. Therefore if it were possible to judge in advance whether the mail is printed or handwritten, each optimal parameter set for method (A) could be selected according to the judgment. However, it is difficult to judge the mail type in advance. On the other hand, method (B) can cope with both printed and handwritten cases without knowledge of types.

6. Conclusion

A method for extracting the address-block from a mail image in various formats was developed and tested. In terms of address-block extraction accuracy of this method, the top-five correct rate is 94% for printed-mail and 89% for handwritten-mail. In the case of an address reading system using this method, the read rate for the printed-mail is 84.5% and that for handwritten mail is 67.5%. It is thus concluded the developed method can be applied to future areas such as signboard extraction and document analysis, if a lot of learning data is collected in advance.

7. References