IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 3, Issue 6, Dec -Jan, 2016
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

Handwritten Address Text String Extraction from Mail Images

Zhimin Huang[1], Ju Wu[2], Yujie Xiong[3]

[1,2]The 3rd Research Institute of Ministry of Public Security of China, Shanghai 200031, China
[3]East China Normal University, Shanghai 200241, China

Abstract

In this paper, we present a novel method for extracting handwritten text string from mail images which contain various types of handwritten characters, machine-printed characters, stamps and noises. Firstly, address text strings in mail images are detected based on connected component analysis and heuristic rules. Then features are extracted from each text string. Adaboost algorithm is applied for feature selection. Afterwards, fisher classifier is applied for the classification of machine-printed and handwritten address text strings. Experimental results show the proposed method can extract handwritten address text string from different kinds of mail images properly. The approach is tested on two databases. The precision and recall are both more than 95%.

Keywords—text string extraction; mail images; connected component analysis; adaboost algorithm; fisher classifier

I. INTRODUCTION

Address block location is of vital importance for postal automation. A lot of articles for postal automation are available [1-3]. Most of them gave equal treatment to machine-printed address blocks and handwritten address blocks. But in practice, recognition technique for machine-printed text and handwritten text is different. It is essential to distinguish machine-printed text and handwritten text before text recognition. Over the last two decades, many studies have been published. Fan et al. [4] described a method for classifying machine-printed and handwritten text blocks, which used character block layout variance as the main feature. In [5], Pal and Chaudhuri proposed a method to detect handwritten text-line from legal instruments. Previous methods mainly performed at three levels: the text line [5], word [6], or character level [7]. It seems that the classification at character level would achieve the goal effectively, but the performance is sensitive since characters may be curved and overlapped. Considering that the address information always appears as rows on the mail images, the method of this paper extracts handwritten characters at the text line level. It leads to a big benefit: the extracted futures of the proposed method are robust.

The mail images are complicated with machine-printed text, handwritten text, and noise mixed together. At first, we eliminate noises and non-text components such as stamps, background texture and so on. Then connected components are extracted, we merge them at the text line level based on heuristic rules. Each text line is regarded as the smallest unit, which is called as candidate zone. Meanwhile, several features are extracted from each candidate zone, and fisher classifier is used to classify them into machine-printed or handwritten text strings.

This paper is organized as follows: In section II, the overall system is presented. Section III is the experimental result and section IV summarizes the conclusions drawn from this study and future work directions are given.

II. SYSTEM OVERVIEW

The system of extracting machine-printed text string from mail images can be divided into three stages. Fig. 1 illustrates the overall framework. Firstly, stamps and noises are removed by heuristic rules after binarization. Then text strings are detected from the mail image by connected component analysis. Features are extracted for each text string. After that, features are selected by adaboost algorithm [8]. Finally, with a fisher classifier [9], selected features are used to classification.

A. Address text string extraction

The inputs of our system are real-world mail images. There are some stamps and banners in the images. Some approaches [10-11] are proposed in order to perform extraction of text-lines. To remove those non-text contents, binarization method is utilized. First, the input image I is converted into binary image B_1 using the Otsu's binarization method [12]. By utilizing morphological dilation operation with a 5*5 operator, image B_1 is converted into binary image B_2. Meanwhile, the input image is also converted into binary image by the Bernsen's binarization method [13], and the output is called as binary image C. Finally, the binary image B_2 and the binary image C are mix together to create the ultimate binary image D by mathematical "OR" operation. Namely, when $d(x, y)$ is a pixel of D, $b_2(x, y)$ is a pixel of B_2, and $c(x, y)$ is a pixel of C, $0 \le x \le w, 0 \le y \le h$ and w, h are image's width and height, respectively, then $d(x, y)$ can be calculated by the following formula:

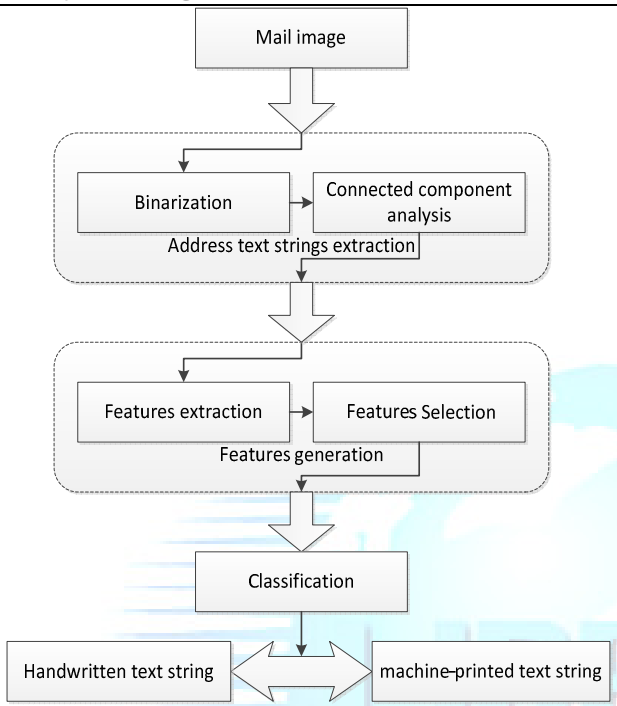


Fig. 1 Block diagram of handwritten text string extraction

$$d(x,y)=\begin{cases}0, if\ \ b_2(x,y)=0, c(x,y)=0\\ 255, otherwise\end{cases} \quad (1)$$

Fig. 3 shows a sample of binary image D. Then connected component labeling is applied to binary image D. It is assumed that the images of mail are all collected in the same orientation. In consideration of the particularity of mail image, some heuristic rules are used to eliminate the stamps, graphics and banners. For example, stamps always appear in the right-top of the image, so the area of the right-top hand corner of the image is wiped off when the connected component labeling is applied. The size of the removed area is confirmed by experiments. Afterwards, the method of text line extraction proposed by [14] is applied, and then the candidate zones are extracted, as shown in Fig. 4.

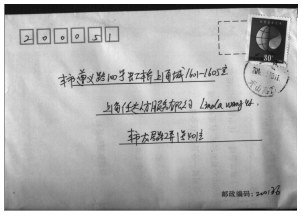


Fig. 2 Input image

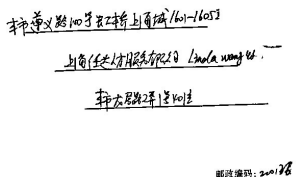


Fig. 3 Binary image D

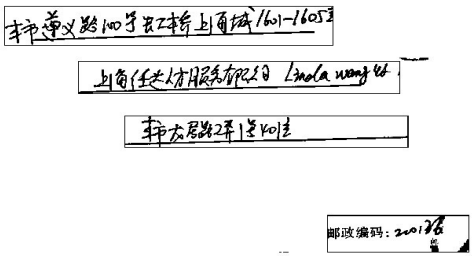


Fig. 4 Candidate zones in the image

B. Features generation

i.Features extraction

For each candidate zone, four kinds of features are extracted including structural feature, run-length feature, cross-count feature and bi-level co-occurrence feature. These features are similar to those used in [15].

ii.Features Selection

As shown in section II. B. i, we extract a total of 50 features from the candidate zone. These features are used to distinguish between different strings, and some features may more stable than others. For the purpose of improving the performance, feature selection is necessary. Another advantage of features election is a small set of features can also reduce the time of feature extraction and classification.

It is generally known that adaboost algorithm can be used to reduce the error of learning algorithm. Suppose that each feature can be used to train a simple classifier, and then we obtain 50 weak classifiers. Using adaboost algorithm, weak classifiers are combined into a single composite classifier, and the weight distribution of the weak classifiers represents the importance of the corresponding feature.

We collected 350 text strings from each class for feature extraction, and applied adaboost algorithm to combine 50 weak classifiers into a single classifier. Fig.5 shows that the relationship of error rate and the number of features used in the classification. As showed in Fig.5, the more features are

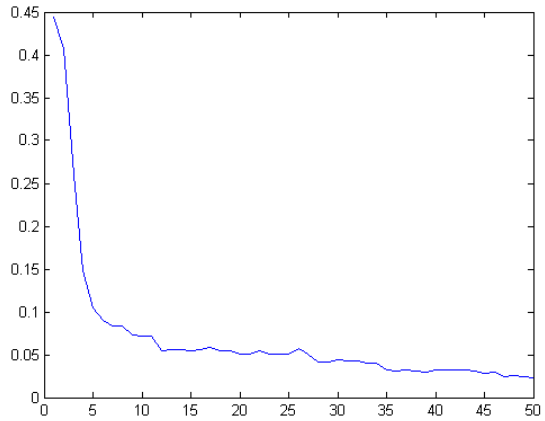


Fig.5 Feature Selection

included, the lower error rate it is. But when the number of features is larger than 15, the decline of the error rate is very slow, and even rebounded. Therefore, in the following processing, we use only 15 features which are selected by the adaboost algorithm.

C. Classification

A fisher classifier is trained for the classification. The fisher classifier has been widely used as an effective classifier for pattern recognition. Compared with Neural Network (NN) and the Support Vector Machine (SVM), the fisher classifier is easy to train, it does not require model selection, and the performance of fisher classifier is non-sensitive to the parameters. Beyond that, the speed of fisher classifier is much faster than SVM and NN, this property meets our requirements of the system. Therefore, we choose fisher classifier for classification. More information can be found in [9].

III. EXPERIMENTAL RESULT

We use two datasets to evaluate the performance of the proposed method. At first, we use 100 mail images as dataset 1 for the test of handwritten address text string extraction. The images are collected with resolution of 96 dpi (average size: 2149*2048 pixels) by the image scanner of the postal sorting machine in real-world. 50 images contain handwritten address text string mainly, and few machine-printed characters. The others contain machine-printed address text string mainly, but also with some handwritten text strings. Whether handwritten text strings or machine-printed text strings, most of them (more than 95%) are Chinese.

We also collected 100 images from IAM-database [16] as dataset 2, in order to test the performance of the approach for classification of machine-printed and handwritten English text strings. The images of IAM-database are English and collected with resolution of 300 dpi (average size: 2479*3542 pixels).Each image contains handwritten text and is labeled by machine-printed text.

For each Dataset, We use 30 of these 100 images to train the classifier and left images to test. We use precision and recall as measures to evaluate the result:

$$precision(i) = \frac{the\ number\ of\ zone\ is\ classified\ as\ class\ i\ correctly}{the\ number\ of\ zone\ is\ classified\ as\ class\ i} \quad (2)$$

$$recall(i) = \frac{the\ number\ of\ zone\ is\ classified\ as\ class\ i\ correctly}{the\ number\ of\ zone\ is\ class\ i} \quad (3)$$

Sometimes, the text strings are not surrounded by the candidate zone completely, so we suppose that if more 85% area of the text strings is surrounded by the text strings, we believe that the candidate zone represents the text strings. If a candidate zone contains both machine-printed and handwritten text strings, it will be considered as a mis-classified zone.

The result of the experiments on dataset 1 is shown in Table 1. Fig.6 is a sample of dataset 1, and Fig.7 is the classification results (the red numbers represent classification results, 1 means handwritten and 0 means machine-printed). This dataset is challenging because text is present in different of styles, the variety of fonts is larger and the images are noisy. The primary reason of misclassification is that the handwritten text strings are overlapped with machine-printed text strings, they are segmented erroneously. In this case, the candidate zone is considered as a mis-classified zone. It can also explain that why the precision of classification for handwritten text is lower than the precision of classification for machine-printed text, but the recall is much higher. Since a mixed zone is always classified as the handwritten text string.

TABLE I. RESULTS OF DATASET 1

	Handwritten	machine-printed
#of zones	165	227
#of zones classified	186	198
#of zones classified correctly	154	185
precision	82.79%	93.43%
recall	93.33%	81.49%

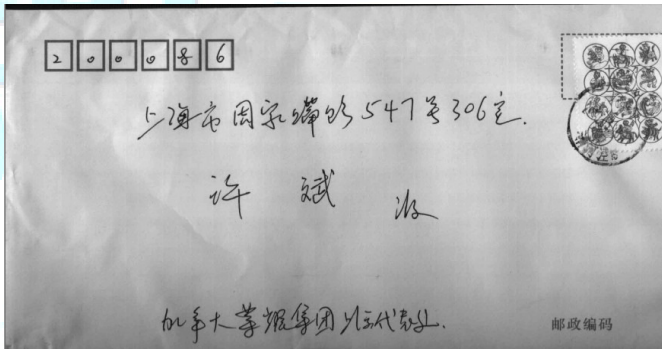


Fig.6 Sample of dataset 1

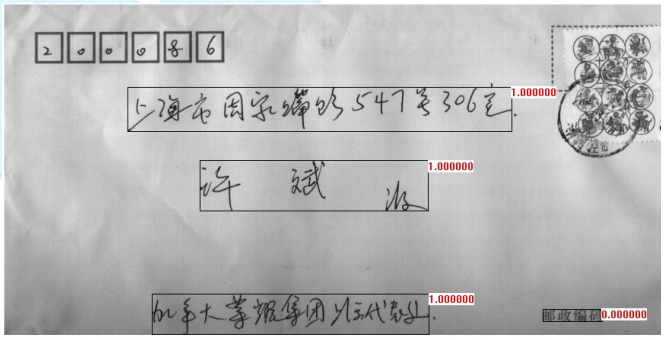


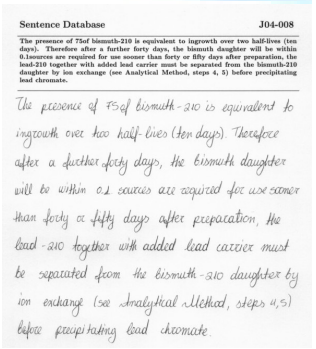
Fig.7 Classification results 1

Owing to regular layout of the inputs which is easy to be segmented correctly, the performance of the proposed method in dataset 2 is quite well. It also indicates that most of misclassifications are caused by bad segmentation. In dataset 2, the classification results of handwritten text are better than

machine-printed text's classification results. The cause of this phenomenon is that the number of the character in machine-printed text string is smaller than the number of the character in handwritten text string, features is more stable when the candidate zone contains more character. On the other hand, the language of two datasets is different, considering that Chinese is more complex than English and the features that we extracted are proposed for English character recognition at first, the difference of the performance in two datasets is acceptable.

TABLE I. RESULTS OF DATASET 2

	Handwritten	machine-printed
#of zones	763	638
#of zones classified	757	654
#of zones classified correctly	748	622
precision	98.81%	95.10%
recall	98.03%	97.49%



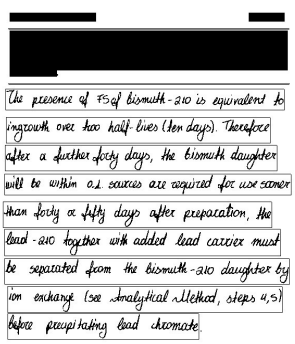


Fig.8 Sample of dataset 2 Fig.9 Classification results 2

IV. CONCLUSION

A system of handwritten text string extraction is presented. It is able to distinguish from machine-printed and handwritten text. The proposed method can deal with from mail images, to find the address areas which are handwritten text. According to the experiments, a good way to reduce the misclassification is to improve the accuracy of segmentation, and the proposed features can reflect the differences between machine-printed and handwritten text in English very well, but facing to Chinese character, we need to find more powerful features. Our future work is to design a new method of segmentation, and create powerful features for classification of machine-printed and handwritten Chinese text.

V. ACKNOWLEDGMENTS

This work is supported by the Research Project of Ministry of Public Security of China (Grant No. 2015JSYJA006).

REFERENCES

[1] S. Srihari, E. Keubert, "Integration of handwritten address interpretation technology into the United states postal service remote computer reader system," The Proceedings of the International Conference on Document Analysis and recognition, pp. 892-896, 1997.

[2] S. Lee, K. Kim. "Address block location on handwritten korean envelopes by merging and splitting method," Pattern Recognition, vol.27, No. 12, 1994, pp. 1641-1651.

[3] J. Bhattacharjee, A. Jain and S. Bhattacharjee, "Address block location on envelopes using Gabor filters," Pattern Recognition, vol. 25, No. 9, 1992, pp. 1459-1477.

[4] K. Fan, L. Wang and Y. Tu, "Classification of machine-printed and hand-written texts using character block layout variance," Pattern Recognition, vol. 31, No. 9, 1998, pp. 1275-1284.

[5] U. Pal, B. Chaudhuri, "Automatic separation of machine-printed and hand-written text lines," The Proceedings of the International Conference on Document Analysis and recognition, pp. 645-648, 1999.

[6] Y. Song, G. Xiao, Y Zhang, and L Yang, "A Handwritten Character Extraction Algorithm for Multi-language Document Image," The Proceedings of the International Conference on Document Analysis and Recognition, 2011, pp. 93-98.

[7] Y. Zheng, C. Liu, X. Ding, "Single character type identification," The Proceedings of SPIE Conference on Document Recognition and Retrieval, 2002, pp. 49-56.

[8] Y. Freund, R. Schapire, "Experiments with a new boosting algorithm," The Proceedings of the International Conference on Machine Learning, pp.148–156,1996.

[9] C. Bishop, Pattern recognition and machine learning, vol. 4, New York: Springer, pp.186-189, 2006.

[10] E. Kavallieratou, N. Dromazou, N. Fakotakis and G. Kokkinakis,"An Integrated System for Handwritten Document Image Processing," International Journal of Pattern Recognition and Artificial Intelligence, vol. 17, No. 4, 2003, pp. 101-120.

[11] E.Kavallieratou, D.Balcan, M.F.Popa, and N.Fakotakis, "Handwritten Text Localization in Skewed Documents," The Proceedings of the International Conference on Image Processing, 2001, pp.1102-1105.

[12] N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, No. 1, 1979, pp. 62-66.

[13] J. Bernsen, "Dynamic thresholding of gray-level images," The Proceedings of the International Conference on Pattern Recognition, 1986, pp. l251 - l255.

[14] V.Pal, B.Chaudhuri, "Machine-printed and handwritten text lines identification", Pattern Recognition Letters, vol. 22, 2001, pp.431-441.

[15] Y. Zheng, H. Li, and D. Doermann, "Machine Printed Text and Handwriting Identification in noisy document Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, vol. 26, No. 3, pp. 337-353.

[16] U. Marti, H. Bunke, "The IAM-database: an English Sentence Database for Off-line Handwriting Recognition," International Journal on Document Analysis and Recognition,2002,vol. 5,No. 1, 39 - 46.