Pattern Recognition Letters 73 (2016) 68–75

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Recognition of handwritten Chinese address with writing variations☆

Xiaohua Wei[a], Shujing Lu[b], Ying Wen[a], Yue Lu[a,b,*]

[a] Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China

[b] ECNU-SRI Joint Lab for Pattern Analysis and Intelligent System, Shanghai Research Institute of China Post, Shanghai 200062, China

ARTICLE INFO

*Article history:*
Received 30 May 2015
Available online 30 January 2016

*Keywords:*
Word-level-tree
Character classification
Candidate address word
Handwritten Chinese address recognition
Writing variations

ABSTRACT

Handwritten Chinese address recognition is a challenging task, not only because of the large quantity of Chinese characters and unconstraint of handwriting, but also due to irregularities of various address formats. The existing techniques generally solve the problem by transforming the address database to a large scale character-level-tree (CLT) and then utilizing the nodes of the generated CLT to match with the candidate patterns. However, the CLT is unable to cover all the variations of address formats. A more compact tree is proposed in this paper to cover the variations of address formats as many and complete as possible by building the structure tree at word level. Specifically, the segment candidate patterns are firstly recognized by a character classifier, then are mapped to candidate address words by matching with the proposed word-level-tree (WLT) address database. Finally, the address recognition result is obtained in the path matching phase by summing the scores of candidate address words in each match path. The proposed scheme was tested with real mail address images captured by an automatic letter sorting machine. Experimental results have demonstrated that the performance of the proposed WLT based method outperforms the four benchmarking methods.



## 1. Introduction

Handwritten Chinese address recognition plays a critical role in automatic sorting of postal letters or parcels (e.g., [2,17]). In mail processing centers, the processing of mail items reliably requires the addresses to be recognized quickly and accurately. Although some progress in Chinese/Japanese address recognition (e.g., [4,6,11,14]) has been achieved, it is still an open issue due to the large quantity and unconstrained writing style of the handwritten characters. In particular, it is quite difficult to handle the various or irregular address formats of the handwritten Chinese addresses.

Huang et al. [3] extracted the key characters, such as province (‘省’), city (‘市’), district (‘区’) etc., to locate the address words and then resort to the address knowledge to recognize the words holistically. To improve efficiency of address retrieval and reduce memory requirements, a trie structure (e.g., [9]) was proposed. In the trie, each node corresponds to a character of address and each path from the root node to a leaf node corresponds to an address phrase (the entry of address list or character string). Furthermore, [11] introduced a beam search scheme for the trie to achieve real-time and accurate recognition. These methods are sensitive to noise since they require that all the characters in the address image are sequentially recognized. Hence, [7] proposed a suffix tree to store the addresses. The structure can access an address from any character, and it can well deal with the document noise and some variations of address formats. The common point of these above mentioned trees is that they are character-level-trees (CLT), and are established on a predefined address list. If the address list is incomplete, the accuracy of recognition will be significantly degraded. [4] used a context-free language for description of the addresses and then transformed the language model into a graph representation. The use of the context-free language formality enabled the target addresses to be included in the language model without enumerating all variants. But its graph search did not consider the relative position of recognized characters. The word matching will fail in the case that the search encounters miss-recognized patterns.

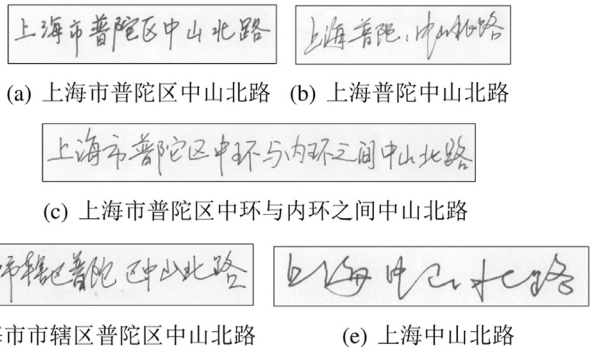
In general, an address is composed of several address words which are defined as the basic administration units. For instance, the address ‘上海市普陀区中山北路’ in Fig. 1(a) includes the

☆ This paper has been recommended for acceptance by Umapada Pal.

* Corresponding author at: Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China. Tel.: +86 21 54345109.

*E-mail address:* ylu@cs.ecnu.edu.cn, luyuetri@aliyun.com (Y. Lu).

http://dx.doi.org/10.1016/j.patrec.2015.12.018
0167-8655/



(a) 上海市普陀区中山北路 (b) 上海普陀中山北路

(c) 上海市普陀区中环与内环之间中山北路

(d) 上海市市辖区普陀区中山北路 (e) 上海中山北路

**Fig. 1.** Writing variations of the address '**上海市普陀区中山北路**'.

address words '上海市', '普陀区', and '中山北路'. The last character in each address word is defined as key character, such as '省' (province), '市' (city), '区' (district), '路' (road), and so on. However, addresses on envelopes are very complicated in real life as people don't always strictly follow the guidelines of addressing formats. Fig. 1 shows writing variations of the address '上海市普陀区中山北路'. Fig. 1(a) is an image with the normative writing format of the address and Fig. 1(b–e) illustrate the writing variations of the same address. Although some information of them is omitted or redundant, these various address formats are still considered as valid. In our study, we group the handwriting variations of Chinese address into three categories, as listed below:

1. *The omission of key characters.* Some key characters may be omitted due to writing habit. For example, '上海' and '上海市', '普陀区' and '普陀' share the same meaning and they are all valid. In this way, the address in Fig. 1(a) may be written as '上海普陀中山北路' (Fig. 1(b)), '上海市普陀中山北路' or '上海普陀区中山北路'.
2. *The omission of address words.* An address can still be valid when some address words are omitted. For example, the address in Fig. 1(a) may be written as '上海中山北路' (Fig. 1(e)).
3. *The redundant words.* Most of these information is often not an address word. They are useless and redundant, but may cause the interference for recognition. For instance, the words '中环与内环之间' in Fig. 1(c) and '市辖区' in Fig. 1(d) are redundant words.

In Chinese handwritten addresses, only about 35% are written in normative format, while about 55% of addresses omit some key characters or address words, and 26% of addresses contain redundant information (omission and redundancy may occur in the same case). Overall, the writing in Chinese addresses is various and it is a heavy and impractical task for manual acquisition of all the variations.

In this paper, we propose a word-level-tree (WLT) based approach for the recognition of handwritten Chinese address with writing variations. In the proposed tree, each node is corresponding to an address word, such that a path from the root to the leaf node is able to give a normative address format. Compared to the CLT, the proposed WLT has three advantages in the practical applications. Firstly, all the writing variations of an address can be mapped to its corresponding normative address format. Secondly, the WLT is more compact and easier to be constructed than the CLT, since the number of nodes in the WLT is significantly reduced. Thirdly, the WLT provides the information of the inherent administrative relationship among different address words.

The block diagram of the proposed system is shown in Fig. 2. The input is an address line image obtained by pre-processing of

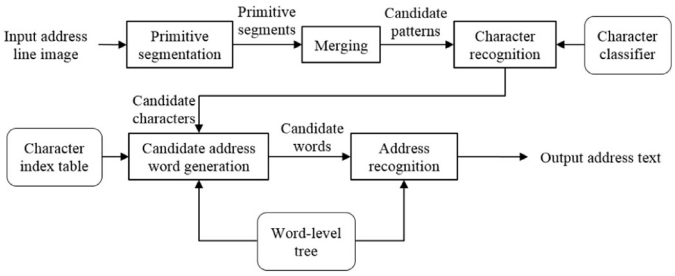



**Fig. 2.** Block diagram of the proposed handwritten Chinese address recognition system.

address block location, text line segmentation and binarization. This system adopts the segmentation-and-recognition framework which is commonly utilized to recognize handwritten character strings (e.g., [5,10,16]). The input address line image is first divided into primitive segments. The consecutive primitive segments are then merged into candidate patterns. After that, the candidate patterns are recognized by a character classifier and are mapped to the address words in the proposed WLT address database to generate the candidate address words. In addition, each candidate word is labeled by a score based on its recognition confidence and segmentation cost. Finally, in the path matching phase, the address recognition result is obtained by summing the scores of candidate address words in each path.

The remainder of this paper is organized as follows: In Section 2, the organization of WLT is presented. Section 3 describes the recognition of handwritten Chinese address. Section 4 shows the experimental results and Section 5 presents a briefly conclusion.

## 2. Organization of word-level-tree

The normative address in China is top-down hierarchical structure, which typically consists of four layers representing province, city, district and road/street, respectively.

Fig. 3 illustrates the normative addresses represented by WLT. For clarity, we give the definitions of the node and path in WLT as follows:

**Definition 1** (Word-level node)**.** Each node is corresponding to an address word rather than a character.

**Definition 2** (Path of normative address format)**.** A path from the root node to a leaf node gives a normative address format.

To deal with missing key characters, the key characters (except the key characters of road name) are set to be optional in the address words (In Fig. 3, the character in the parenthesis is optional). Since the address format is variable, an address may start from any basic administration unit or skip some nodes of certain layers. However, the road name is a requisite in address as it is the most important information for destination. Once the road name is given, the candidate addresses containing that road name can be easily found out. For example, if the address word '中山北路' is identified, the address words '上海市', '普陀区' , '浙江省', '杭州市', '下城区' , etc., can be predicted via the WLT and the address '上海市普陀区中山北路' and '浙江省杭州市下城区中山北路' , etc., will be considered as the candidate addresses.

Table 1 shows the index of characters of all address words. In the table, the column of *Character* lists all characters involved in the address words. Each character is indexed by its GB2312-80 code. The third column lists the correlative address words. In this manner, when a character is recognized, a series of associated