

# Handwritten Chinese Address Recognition

Chunheng Wang  
Fujitsu R&D Center Co. Ltd.  
Beijing, China 100016  
[wangchh@frdc.fujitsu.com](mailto:wangchh@frdc.fujitsu.com)

Yoshinobu Hotta, Misako Suwa, Satoshi Naoi  
Fujitsu Laboratories Ltd.  
Kawasaki, Japan  
[{y.hotta, suwa, naoi.satoshi}@jp.fujitsu.com](mailto:{y.hotta, suwa, naoi.satoshi}@jp.fujitsu.com)

## Abstract

A handwritten Chinese address recognition (HCAR) system is proposed in this paper. Handwritten Chinese address recognition is a difficult problem. Handwritten Chinese characters are characterized by large vocabulary, complicate structure, irregular distortion and touching characters etc. Proposed approach takes good advantage of Chinese address knowledge, and applies key character extraction and holistic word matching to solving the problem. Different from conventional approach, proposed approach can avoid the character segmentation error successfully. Experimental results show the proposed approach is very effective.

## 1. Introduction

Handwritten Chinese character recognition (HCCR) is one of the most difficult problems in character recognition field. Large category, irregular distortion and complicate character structure prevent it from actual application.<sup>[1, 2]</sup> In recent years, more and more research effort has been put on some specific application fields. Among them, handwritten Chinese address recognition (HCAR) is a typical problem and research direction.

In conventional recognition way, the address string image is firstly segmented into single character images, and then be recognized one by one. Segmentation error is unavoidable in this way, and cause sharp drop of recognition rate. Then how to avoid segmentation error becomes a key problem. Here we solve the problem by applying knowledge to recognition. Like most addresses in the world, there is a hierarchical multi-layer structure in Chinese addresses. For example, “北京市海淀区中关村” is a Chinese address. It has three layers or sections in, they are “北京市”, “海淀区” and “中关村”. “市”, “区” and “村” define the difference and hierarchical relationship between sections, are called key characters. “市” is the upper level key character of “区”, and “村” is low level key character of “区”. “北京”, “海淀” and “中关村” are the address section names, and called words. In this paper, the Chinese address hierarchical structure is applied to recognition. Firstly, key characters are extracted; then words are recognized. In stead of being segmented into

single characters, word is recognized in a holistic way.<sup>[3, 4]</sup> So segmentation error can be avoided in word recognition.

Before word recognition, key character should be extracted. If key character can not be extracted correctly, word recognition result will be false. Our previous approach used in Japanese key character extraction<sup>[5]</sup> is successful to deal with Japanese address recognition. But for Chinese character, it is no more effective. Chinese character has more complicate structure. Some characters include two or three radicals, and are often wrongly segmented into two or three single characters. In order to solve the problem, an enhanced key character extraction approach is proposed. In the enhanced approach, one, two and three radicals are processed and recognized, the most possible and similar combination is regarded as the final result. Figure 1 shows the whole recognition procedure.

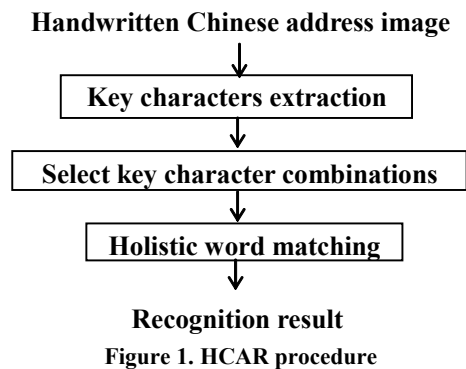


Figure 1. HCAR procedure

Chinese address is analyzed in section 2. Key character extraction is introduced in section 3. Section 4 describes holistic word matching. Experiments and analysis are provided in section 5. Section 6 is conclusion remark.

## 2. Analysis of Chinese address

HCAR is a large vocabulary recognition problem. There are totally 4,569 different characters in Chinese addresses. In order to apply knowledge into HCAR, a thorough analysis of Chinese address is necessary.

As introduced above, there is definite hierarchical multi-layer structure in Chinese addresses. Figure 2

shows the Chinese string model. There are 4 levels at most in Chinese addresses.

It should be pointed out that there is other information after the last key character in a real address. For example, number of building and number of door. These characters after the last key characters are not dealt with in this paper. They are processed by conventional approach in the real system.

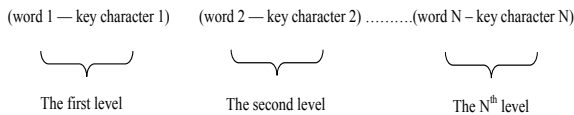


Figure 2. Chinese address string model

Key characters that define the difference and hierarchical relationship between sections or levels are very important information in addresses. There are 22 key characters in Chinese addresses, they are “市,省,区,弄,路,街,村,乡镇,港,湾,县,道,里,同,巷,楼,州,旗,胡,庄,坊.” And there are only 9 in Japanese addresses, they are “都,道,府,县,市,区,郡,町,村”, less than the number of Chinese key characters. Key characters have definite relationship and structure. Key character sets according with the relationship and structure are regarded as legal sets. For example, {市 (CITY) – 路 (ROAD)} is a legal key character set, and {路 (ROAD) – 市 (CITY)} is illegal in China because road belongs to a city. Figure 3 shows part of the key character structure in China.

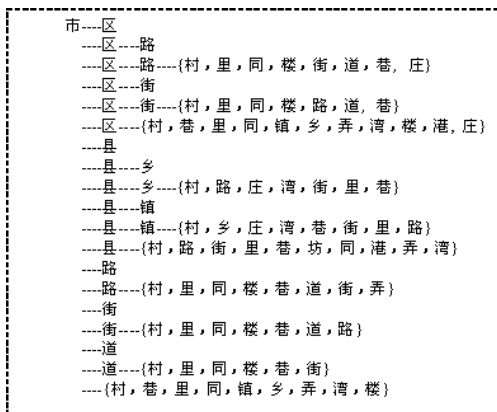


Figure 3. Part of the hierarchical multilayer structure of Chinese addresses

The structure is a tree structure. Each path from the start to end is a legal key character set. There are totally 220 legal key character sets in Chinese address, much more than the 18 legal sets in Japanese Addresses. Address complexity lies most on the number of key characters and legal sets. Chinese addresses are much more complicate than Japanese addresses.

From above analysis, we can see that definite hierarchical multi-layer structure in Chinese addresses, each address is made up of key characters and words. Key characters and words keep definite relationship. It is the basis for key character extraction and holistic word recognition. Compared with Japanese addresses, Chinese addresses have more complicate structure, more key character, more complicate character structure, more word number and more characters in word. Aiming to solve the problems, we developed the Chinese address recognition approach and system introduced in this paper.

### 3. Key character extraction

#### 3.1. Definition

As introduced in above section, characters that indicate or define different section or level in an address, and indicate the hierarchical relationship between sections are called key characters. As shown in figure 4, “市”, “区” and “里” are the three key characters in the address “北京市宣武区红莲北里”.

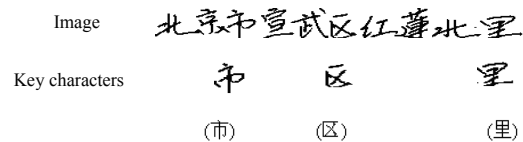


Figure 4. Definition of key character

Shape structure of Chinese key characters is more complicate than that of Japanese. Japanese key characters have no more than 2 radicals. But there are three radicals in Chinese key characters. Such as “街”. More key character number and more complicate character structure prevent our previous key character extraction approach no more so effective here. In order to extract key characters from a Chinese address string image correctly, an enhanced key character extraction scheme is proposed and introduced in section 3.3.

#### 3.2 Conventional key character extraction

Key character extraction is the first step in the address recognition procedure. It is also a very important step.

Figure 5 shows the key character extraction procedure. First of all, segment the image into radical images. Then recognize each single radical image. There are only 22 key characters in Chinese addresses. Although it is much easier to do such a little category classification work, segmentation error still can not be avoided in the conventional way.

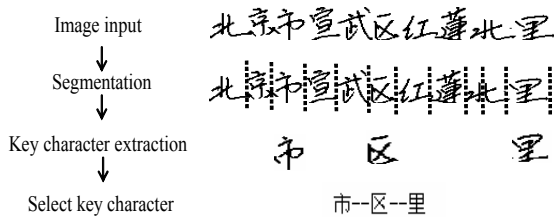


Figure 5. Key character extraction

After key character extraction, there will be lots of groups of key characters. According to the hierarchical multi-layer structure of Chinese addresses, the legal key character groups can be selected from all group candidates. In figure 5, the final legal key characters g {市---区---里} , it can be found in figure 3.

### 3.3 Enhanced Key character extraction

In above address shown in figure 5, all the three key characters only include one radical. But there are two or three radicals in some key characters, such as “街”, which includes three radicals. There are many Chinese characters with two or three radicals. Large vocabulary and complicate structure make it is very difficult to extract correct key characters by conventional extraction approach. In order to extract all the key characters correctly, the enhanced key character extraction approach is proposed. In this approach up to three radical combinations are considered and processed. Figure 6 shows the enhanced key character extraction procedure.

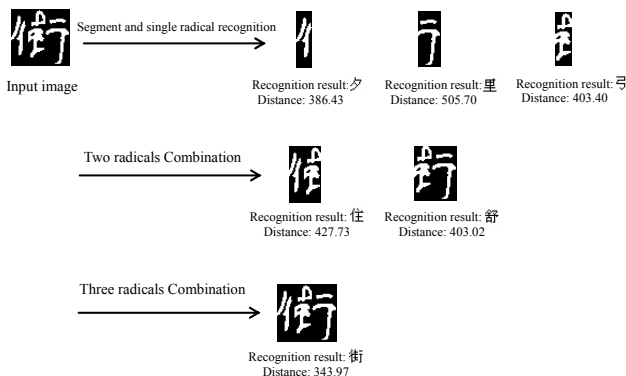


Figure 6. Three radical combinations recognition

Firstly, each single segmented radical is recognized. Then two radicals are recognized as a character. Last three radicals are combined together and recognized. After recognition, we can select correct recognition result according to the first candidate and matching distance. In figure 6, there are six recognition results. Among the six results, only “里” and “街” are key characters. So, we

only consider the two candidates. Matching distance of “街” is 343.97, which is less than 505.70, the matching distance of “里”. We select the candidate with less matching distance, so the final recognition and extraction result is “街”, it is the correct result.

By considering three radical combinations, the key characters can be extracted more accurately. Compared with conventional approach, enhanced key character extraction approach considers more possible segment situations. It also benefits touching character segmentation because some touching strokes can be broken without hesitation. And then the broken images are regarded as radicals in the radical combination strategy. So the enhanced approach can avoid many segmentation errors.

### 4. Holistic word matching

The characters between two key characters are regarded as a single unit and called word. As shown in figure 7, “北京”, “宣武” and “红莲北” are the words in the address.

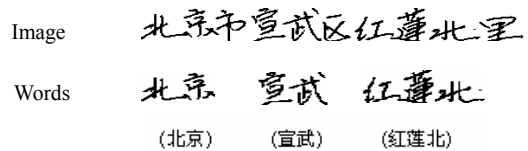


Figure 7. Definition of word

After key character extraction, words are recognized in holistic way. No matter how many characters in the word, it will be recognized as one single character. This avoids character segmentation error greatly. In our scheme, dictionary of word is synthesized on line in a dynamic way, so the dictionary size is not so large. Detail information about the holistic word matching is available in reference.<sup>[3]</sup>

### 5. Experiments and analysis

In order to verify proposed approach, a series of experiments have been carried out, including compare experiments with conventional approach and our previous approach used in the Japanese address recognition. Feature used in proposed approach, conventional approach and previous approach is contour chain code feature proposed in reference [6].

In the conventional approach, each address string image is segmented into single characters and recognized. City block distance is applied to the classification. In the conventional approach, no address knowledge is used; there is no post processing procedure in the recognition.

Two kinds of recognition rate are calculated. One is string recognition rate (SRR); the other is character recognition rate (CRR).

$SRR = (\text{Correctly recognized address number} / \text{Total test address number}) \times 100\%$

$CRR = (\text{Correctly recognized character number} / \text{Total test character number}) \times 100\%$

### 5.1 Experiment data

Figure 8 shows part of the test data. The whole test data includes 600 test address images, which were written by different person from different social strata. The test data is categorized into three subsets according to writing quality, which are good, normal and bad. Each subset contains 200 images respectively. The proposed approach and conventional approach have been tested with each subset.

Good → 湖北省应城市义和镇  
 Normal → 甘肃省榆中县高崖乡  
 Bad → 山东省淄博市博山区源泉村

Figure 8. Experiment data

### 5.2 Experiment results and analysis

Table 1 and table 2 show the comparison experiment results of proposed approach and conventional approach. Compared with conventional approach, the approach proposed in this paper show much higher recognition rate. For string recognition, recognition rate of proposed approach is much higher than conventional approach. It verified the fact that character segmentation brings unavoidable error into recognition. Proposed approach can avoid segmentation error effectively.

Table 1. Experiment results of proposed approach

Recognition rate / Test data	SRR	CRR
Good (200)	100.00%	100.00%
Normal (200)	95.78%	98.31%
Bad (200)	85.62%	91.04%
Total (600)	93.80%	96.45%

Table 2. Experiment results of conventional approach

Recognition rate / Test data	SRR	CRR
Good (200)	42.11%	83.01%
Normal (200)	21.87%	65.96%
Bad (200)	2.40%	39.84%
Total (600)	22.13%	62.60%

Figure 9 shows one of the samples correctly recognized by proposed approach but miss-recognized by conventional approach.

福建省福州市打铁土当新村

(a). Test sample

福建省福州市打铁土当新村

福建省 福州市 打铁土当 新村

(b). Correct recognition result of proposed approach

福建省福州市打铁土当新村

末 懒 福 叫 步 打 铁 与 删

(c). Miss-recognition result of conventional approach

Figure 9. Sample 1 and recognition results

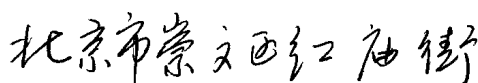
From above recognition results, it can be seen that segmentation error is the main obstacle preventing conventional approach from correctly recognizing. In fact, even segmentation result is right, it is also a difficult problem to recognize handwritten Chinese single character, it is because of large vocabulary, irregular distortion and complicate structure etc. On the other hand, proposed approach avoids the obstacles by enhanced key character extraction and holistic word matching. There are only 22 key characters; it is easier to solve such a vocabulary classification problem. For word recognition, there are only about one hundred words corresponding to each key character, so high recognition rate can be reached.

Table 3 shows the experiment results of our previous approach. Compared with our previous approach used in Japanese address recognition system, proposed approach is more effective in key character extraction by considering three radical combinations. So, it show more accuracy in Chinese address recognition than previous one.

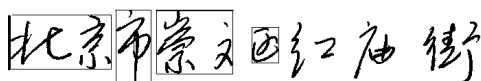
**Table 3. Experiment result of previous approach**

Recognition rate / Test data	SRR	CRR
Good (200)	88.00%	93.16%
Normal (200)	90.35%	94.75%
Bad (200)	83.20%	88.87%
Total (600)	87.18%	92.26%

Figure 10 shows one of the samples correctly recognized by proposed approach but miss-recognized by our previous approach.

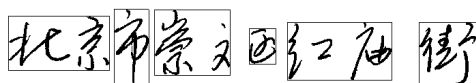


(a). Test sample



北京市 崇文 区

(b). Correct recognition result of proposed approach



北京市 崇文 区 红庙 街

(c). Miss-recognition result of previous approach

**Figure 10. Sample 2 and recognition results**

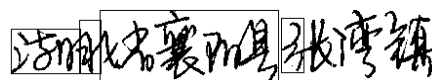
There is key character “街” in the address shown in figure 10. Previous key character extraction approach does not consider three radical combinations, so the last key character and word are lost. The enhanced key character extraction approach considers the three radicals situation, so it can extract correct key character set, and final correct string recognition result.

Figure 11 shows one of the samples miss-recognized by proposed approach.

Both key character extraction and word matching results of above image are wrong. It is because the writing quality of the address image is very bad. In fact, it is difficult for human to recognize. This address image is also miss-recognized by the conventional approach and the previous approach.



(a). Test sample



洛阳市 操场 街

(b). Miss-recognition result of proposed approach



湖北省 襄阳县 张湾 镇

(c). Correct result

**Figure 11. Sample miss-recognized by proposed approach**

## 6. Conclusion remark

Handwritten Chinese address recognition is large vocabulary recognition problem with wide application fields. In this paper, a recognition approach based on address knowledge is proposed. Key character, word and hierarchical address model are defined. By key character extraction based on enhanced key character extraction and holistic word matching, character segmentation error is avoided effectively. This approach shows high recognition performance, and has been applied to actual document auto-processing business.

## References

- [1] Hao Hongwei, Xiao Xuhong and Dai Ruwei, "Handwritten Chinese character recognition by metasynthesis approach", *Pattern Recognition*, 30(8), 1321—1328, 1997
- [2] Ruwei Dai, Hongwei hao and Xuhong Xiao, "Integrated Chinese character recognition approach and system", Zhejiang science and technology press, 1998
- [3] Y.Hotta, H.Takebe and S.Naoi, "Holistic Word Recognition Based on Synthesis of Character Features, "Fourth IAPR International Workshop on Document Analysis Systems(DAS) pp.313-324, 2000
- [4] S.Naoi, M.Suwa, and Y.Hotta, "Recognition of Handwritten Japanese Addresses Based on Key Character Extraction and Holistic Word Matching," Third IAPR

International Workshop on Document Analysis Systems(DAS) pp.149-152, 1998

[5] N.Babaguchi, M.Tsukamoto and T.Aibara "An Improvement of the Segmentation of Handwritten Characters by Introducing Recognition Process", IEICE Trans.J69-D, No.11, pp.1774-1782, 1986

[6] M. Shridhar, F. Kimura "Segmentation-Based Cursive Handwriting recognition", Handbook of Character Recognition and Document Image Analysis, pp 123~156, 1997