基于知识模型的手写中国地址识别系统

王春恒 堀田悦伸* 諏訪美佐子* 直井聡*

（富士通研究开发中心，北京 100016）

（* 富士通研究所，日本川崎）

E-mail: wangchh@frdc.fujitsu.com

摘 要：中国手写地址识别是一个具有广泛应用场合的大类别识别问题，针对这一问题，本文给出了一种新的基于中国地址知识模型的识别方法。方法中强调中国地址固有的树状分层结构信息，通过抽取较少的关键字和词语的整体识别，避免了传统识别方法中单字分割所带来的分割误差，体现出较高的分类性能，对一般书写的地址字符串识别率达到 93.80%，单个字符的识别率达到 96.45%.

关键词：手写汉字识别 手写地址识别 层次结构模型 关键字 整体识别

Handwritten Chinese Address Recognition Based on Knowledge Model

C.H. Wang, Y. Hotta*, M. Suwa*, S. Naoi*

(Fujitsu R&D Center Co. Ltd, Beijing 100016)

(* Fujitsu laboratory Co. Ltd, Kawasaki, Japan)

E-mail: wangchh@frdc.fujitsu.com

Abstract: Handwritten Chinese address recognition is large category classification problem with much actual application foreground. In order to solve this problem effectively, a new knowledge based recognition approach is proposed in this paper. A hierarchical multilayer tree model of structure is defined and applied to the recognition procedure. According to the model, a little number of key characters and relative word can be searched and recognized, which reduces the segmentation errors inevitable in traditional method. The new approach is proved to be very effective in experiments. For 600 handwritten address images of test, the string recognition result is 93.80%, character recognition rate is 96.45%.

Keywords: Handwritten Chinese character recognition, handwritten address recognition, hierarchical multi-layer model, key character, holistic word matching

1 介绍

虽然 OCR 技术的研究取得了很多进展，然而没有任何约束的手写体汉字识别技术与实

际应用之间还有相当长的距离。手写汉字识别是字符识别中最为困难的问题之一，具有类别大、相似字多和不规则变形严重等特点[1]。现在，专用领域中手写汉字识别技术的研究与应用备受关注，其中手写地址识别(Handwritten Chinese Address Recognition, HCAR) 就是一个典型的专项识别问题。地址信息是各种表格和信封中重要的信息，也是最难识别的信息。中国地址中包含的字符种类超过了 4,500 类，所以说手写地址识别也是一个大类别识别问题，是表格自动处理的关键。

中国地址具有一定的层次结构，例如“北京市海淀区中关村”中有三个层次，也就是三段。其中，“市”、“区”和“村”代表着段的属性和段与段之间的区别，称之为关键字；“北京”、“海淀”和“中关”标志着段的内容，称为词。本文地利用了地址层次结构知识，定义并建立了中国地址关键字树状层次结构模型、地址串语言模型和中国有效地址知识库，并将这些信息应用到识别的整个过程，如图 1 所示。首先，抽取地址字符串图像中的关键字和关键字的有效组合；之后，利用关键字组合信息，查询出与关键字组合相符的词语候选集，对于关键字之间的图像，在词语候选集中以词为单位进行词语识别[2,3]。最终得到可信度最高的识别结果。实验证明，本文的中国手写地址识别系统能够达到较高的和实用的效果，具有广泛的应用前景。

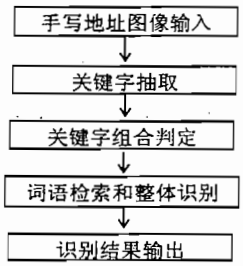

图 1. 中国手写地址识别过程

本文的第二节介绍了关键字结构模型定义和关键字抽取技术；第三节介绍了词语整体识别技术。实验与分析在第四节；第五节是结束语。

2 关键字抽取

2.1 关键字定义

中国地址在层次结构上有着固有的规律。其中有一个具有不变性和标志性的字符集。这一字符集中的字符频繁出现在地址中，标志着地址逐级精确的层次结构和地址中每段的属性。这一字符集在本文中叫做关键字字符集，字符集中的字符称为关键字。如图 2 所示，

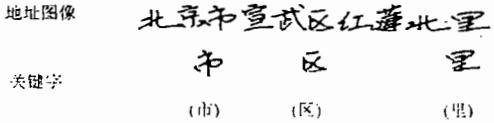

图 2. 关键字定义

在“北京市宣武区红莲北里”这一地址中，“市”、“区”和“里”就是三个关键字。

2.2 中国地址关键字树状层次结构模型

地址中的关键字具有一定的层次结构关系，也就是上下级关系。图3所示为中国部分地区的地址关键字层次关系。定义并统计关键字层次结构非常重要。这种层次结构关系将引导关键字的抽取，为每个关键字确定合法的候选集，为地址中所有关键字抽取提供合法的组合信息。

关键字的这种层次关系呈现出一种树状结构。其中，处于第一层的关键字是树的根节点，图3中的根节点只有一个关键字“市”，可以用集合的方式表示为{市}；处于第二层的关键字组成树的二层子节点，表示为{区，县，路，街，道，村，巷，里，同，镇，乡，弄，湾，楼}；以此类推，中国地址关键字最多只有四层。与普通树状结构不同的是，每个节点都是末节点，这意味着停止于任何关键字的地址都被看作有效地址来处理。例如以下地址都是有效地址：

（1）北京市

（2）北京市西城区

（3）北京市西城区展览路

（4）北京市西城区展览路北四巷

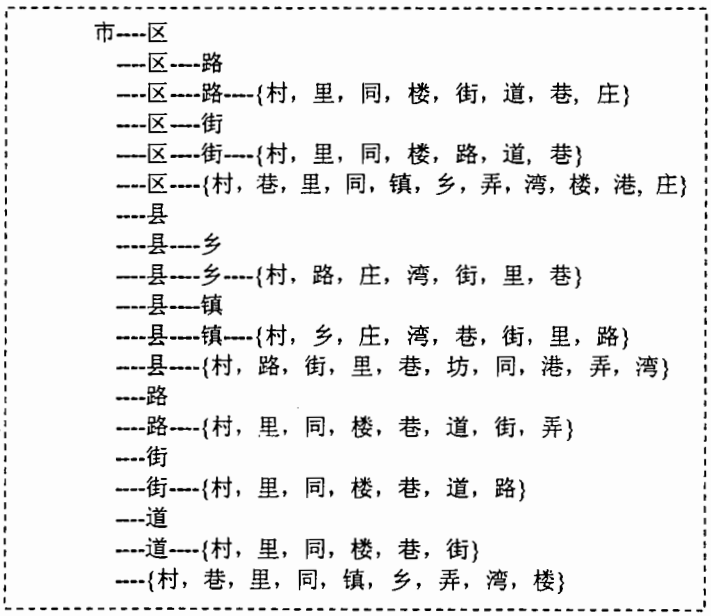

图 3. 中国部分地区关键字树状层次结构关系模型

在这个树状层次结构中，前一节点是后一节点父节点，后一节点是前一节点的子节点。这种父子关系在关键字抽取中非常重要，前一节点的关键字决定着后面待识关键字的类别集。例如，如果识别得到前一关键字为“乡”，则接下来只需要在{村，路，庄，湾，街，

里，巷}中寻找正确的关键字，有效的提高了识别结果的准确性。

2.3 关键字抽取

图 4 所示为关键字抽取的过程。首先，要对原始图像进行字分割。不同于传统的 OCR 中的字分割，这里只考虑关键字字符集范围内的分类，也就是 22 类分类问题，这要比 6,763 类所有二级汉字分类问题简单的多。

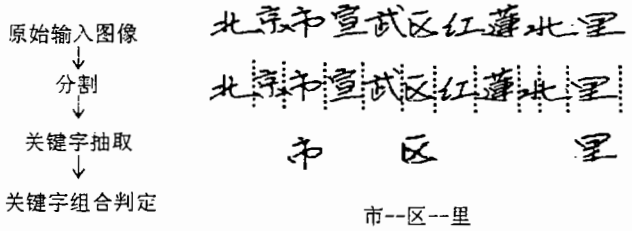

图 4. 关键字抽取

在抽出关键字之后，要对关键字的组合进行判断。组合判断的依据是(1)识别结果中关键字候选的相似度信息；(2)中国地址关键字树状层次结构关系模型。图 4 中最终确定的有效关键字组合的第一候选为{市 — 区 — 里}。

3 地址字符串的分层递阶识别

3.1 词语定义

个地址中位于关键字之间的信息称为词或词语，代表着地址中每一段的具体内容。如图 5 所示，“北京”，“宣武” 和“红莲北”是地址中的词。

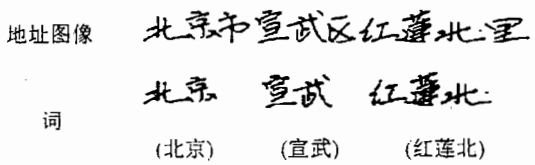

图 5. 词定义

3.2 地址语言模型

至此可以看出，每一条地址由关键字和词组成，关键字和词遵循着如下关系：

$$(\text{词 } 1 - \text{关键字 } 1) - (\text{词 } 2 - \text{关键字 } 2) \cdots\cdots (\text{词 } N - \text{关键字 } N) \qquad (\text{式 } 1)$$

第 1 层　　第 2 层　　第 N 层

式 1 可以被看作地址语言模型。当然，这不是严格意义上的语言模型，没有主谓宾属性和关系。但是每层之间具有严格的上下级关系，每层单元的属性由关键字确定，内容由词来确定。整体识别的整个过程都是由这个地址语言模型来引导的。

3.3 分层递阶识别

在关键字抽取的基础上，对整个地址分层次进行递阶识别。图 6 所示为地址识别的完全过程。这一识别方法的特点就是以词为单元进行整体处理，将词作为一个传统方法中的单字进行识别。首先，对输入的词语图像块进行预处理；然后从预处理后的图像中抽取特征向量，本文中采用广泛应用的轮廓方向属性特征[4,5]；来依据地址语言模型检索地址知识库，合成词语字典特征；最后，将特征与合成的词语字典相比较，得到最终的词识别结果。

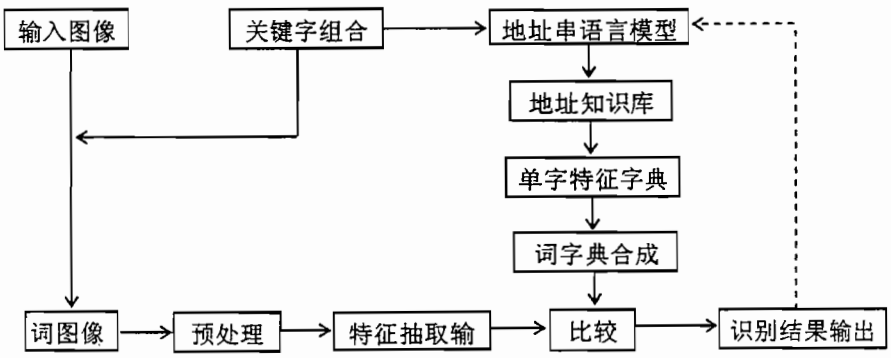

图 6 词语整体识别

下面结合具体地址介绍这一识别过程。

对于图 2 中的地址图像，其关键字抽取结果为“市—区—里”，在此基础上，识别过程如下：

(1) 根据关键字组合和地址语言模型，确定地址第一层单元的属性为“市”。

(2) 在地址知识库中检索所有属性为“市”，同时层次为第一层的词，形成如下一个词语列表

{北京，上海，天津，重庆}

如果具有父节点关键字，则要根据上层词，父节点关键字和本节点关键字检索地址知识库，确定词语列表。

(3) 在单字特征字典中检索词语列表中每个字符的特征，合成如下词字典

{VD(北京),VD(上海),VD(天津),VD(重庆)}，其中 VD(*)是一个特征矢量。

(4) 由关键字抽取结果可以确定并截取地址图像中的每块词图像。这里应该有四块词图像被抽出。

(5) 对第一块词图像进行预处理，抽取特征 VX.

(6) 将 VX 与词字典中的每一个词特征进行比较，最为接近词特征对应的词就被认为是

词图像的识别结果。这里正确的识别结果为“北京市”。

(7) 至此得到第一层次单元的识别结果为“北京市”。

反复上述步骤，直至所有层次单元都得到识别，从而得到整个地址图像的识别结果。

在词语字典特征的合成过程中，主要考虑两方面的信息。第一，关键字层结构组合信息；第二，词语列表信息。词语列表是根据关键字层次结构由地址词库中检索得出的。这个地址词库涵盖了全国所有地区的有效地址，其层次深入到村、胡同等较小的级别。

在得到关键字抽取结果和词语整体识别结果之后，按照式 1 中的关系组合成完整的地址识别结果，进行合法性判断，得到最终的识别结果。

4 实验与分析

为了验证方法的有效性，对系统进行了实际的测试，并与传统的基于单字分割的方法进行了比较。

实验中统计了两个主要性能指标，一个是字符串识别率，另一个是字符识别率。其中，

字符串识别率 =（识别正确的地址字符串数/地址字符串数总数）×100%

字符识别率 =（识别正确的字符数/字符总数）×100%

统计地址字符串识别正确与否的标准是，如果有一个或一个以上的字符识别错误，则认为整个地址字符串识别错误；只有所有字符完全正确，才认为地址字符串的识别结果正确。

4.1 实验数据

图 7 所示为部分用于测试的地址字符串图像。整个测试集包括 600 条手写中国地址，分别由不同文化背景和工作环境的人书写而成。我们将这 600 条手写地址按照书写质量分成好、中和差三类，每类包括 200 条地址，分别进行测试，同时也统计了总体识别性能。

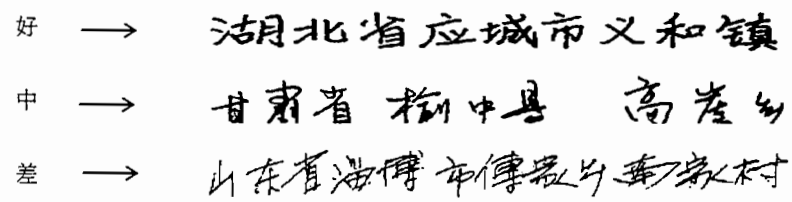

图 7. 部分测试样本

4.2 实验结果及分析

表 1 所示为实验结果。每一项子测试的识别率都远高于传统的模板匹配方法的识别率。其中总体字符串识别率达到 93.80%，总体单个字符的识别率为 96.45%。这一识别率使得本文中手写中国地址识别技术的实际应用成为可能，也使得在实际应用中的人工校正修改工作量大大减少，提高了信息录入和管理的效率。

表 1. 实验结果

测试指标和方法 / 样本种类	字符串识别率		字符识别率	
	本文方法	传统方法	本文方法	传统方法
好（200）	100.00%	42.11%	100.00%	83.01%
中（200）	95.78%	21.87%	98.31%	65.96%
差（200）	85.62%	2.40%	91.04%	39.84%
总体（600）	93.80%	22.13%	96.45%	62.60%

5 结束语

中国手写地址识别是一个具有广泛应用场合的大类别识别问题，针对这一问题，本文给出了一种新的基于地址树状层次结构知识的识别方法。方法中强调中国地址固有的分层结构信息，通过抽取较少的关键字和词语的整体识别，避免了传统识别方法中单字分割所带来的分割误差，体现出较高的分类性能。

参考文献

[1] Hao Hongwei. Xiao Xuhong and Dai Ruwei. "Handwritten Chinese character recognition by metasynthesis approach". *Pattern Recognition*. 30(8), 1321—1328, 1997

[2] 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

[3] 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

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

[5] 戴汝为，郝红卫，肖旭红，《集成型汉字识别方法与系统》，浙江科学技术出版社，1998.