第 36 卷 第 7 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.7 Computer Engineering April 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)07—0170—03 文献标识码：A 中图分类号：TP391

手写汉语拼音的融合识别系统

朱 萌[1,2]，刘长松[1,2]，陈御天[1,2]，邹燕明[3]
(1. 清华大学智能技术与系统国家重点实验室，北京 100084；2. 清华大学电子工程系清华信息科学与技术国家实验室，北京 100084；
3. 诺基亚北京研究院，北京 100176)

摘 要：手写设备用户容易忘记特定中文单字写法，需要为其提供拼音输入法。采用分类器融合方式构筑拼音单词识别系统，通过隐马尔可夫模型分类器获得拼音单词的切分点，利用统计特征识别模块进行识别后融合，研究并改进拼音单词基线提取方法。实验结果表明，该方法对 17 745 个测试样本的识别率达 91.37%。

关键词：中文信息处理；字符识别；基线

Combined Recognition System for Handwritten Pinyin

ZHU Meng[1,2], LIU Chang-song[1,2], CHEN Yu-tian[1,2], ZOU Yan-ming[3]
(1. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084;
2. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084; 3. Nokia Research Center, Beijing 100176)

【Abstract】 Handwritten device users are easy to forget how to write a certain Chinese character. It is necessary to provide Pinyin input method for them. This paper constructs a Pinyin word recognition system through classifier fusion style. It obtains the cutting point of Pinyin word by Hidden Markov Model(HMM) classifier, accomplishes after-recognition fusion by using recognition module for statistic characteristic, studies and improves the base line extraction method for Pinyin word. Experimental results show that this method can recognize 91.37% test samples from 17 745 ones.

【Key words】 Chinese information processing; character recognition; base line

1 概述

手写识别为掌上产品用户提供了较自然的文本输入接口。由于中文字符集规模庞大(GB-2312-80 标准定义 6 763 个简体中文字符，Big5 标准定义 13 053 个繁体中文字符)，其中一些字符的形状复杂[1]，因此用户可能暂时忘记一个字符的结构但仍然记得该字符的发音。另外，一些多笔划字符的书写耗时较大。拼音是普通汉语中最常用的注音符号系统。对掌上设备用户来说，若在忘记某个中文字符的写法时可以切换到拼音输入，就能节省一定时间。因此，对汉语拼音识别系统的研究具有实际意义，但目前该领域的成果较少[1]。拼音单词以声母和韵母组合的形式出现，笔者采用 467 个拼音单词组成的词典，研发了一个手写汉语拼音识别系统。用户可以使用手写板输入汉语拼音数据，识别系统会对输入的拼音单词进行识别并返回匹配程度较高的前 10 位侯选单词。汉语拼音识别系统运行界面如图 1 所示。

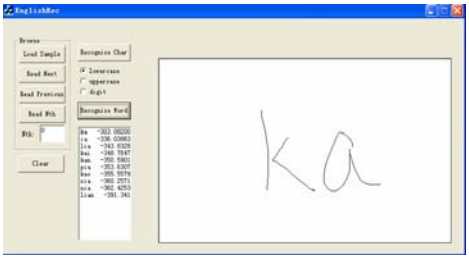

图 1 汉语拼音识别系统运行界面

2 系统架构

拼音单词识别系统架构见图 2。

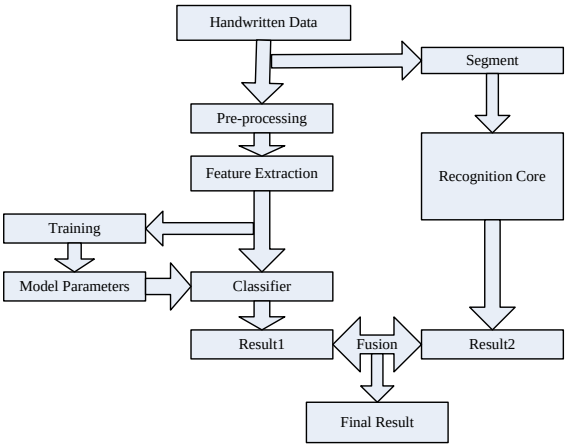

图 2 拼音单词识别系统架构

当系统接收到一个拼音单词输入数据时，拼音单词隐马尔可夫模型(Hidden Markov Model, HMM)模块会产生单词的切分信息，拼音单词输入被切分成分开的单字数据并被送入单字识别核心，以产生一个匹配度分数。拼音单词 HMM 本身可以产生识别的后验概率。最终结果由这 2 个结果融合产生。

在单词 HMM 模块部分，手写拼音单词要先经过预处理

基金项目：国家“973”计划基金资助项目(2007CB311004)；国家自然科学基金资助项目(60772049)

作者简介：朱 萌(1984－)，男，硕士研究生，主研方向：文字识别；刘长松，副教授、博士；陈御天，博士；邹燕明，研究员

收稿日期：2009-11-02 **E-mail**：zhumeng030@gmail.com

过程，包括去噪、归一化、重采样和平滑滤波，在归一化过程中采用的一些独特方法将在第 3 节中介绍。预处理步骤的目的是实现对手写数据进行特征提取[2]。在建模阶段，选用异型体作为基本建模单元[1]。dx 和 dy(图 3)被提取为单词 HMM 的特征。在训练过程中，样本在提取特征后用来训练单词隐马尔可夫模型的模型参数。在识别过程中，拼音样本会被送到分类器中，以产生后验概率。

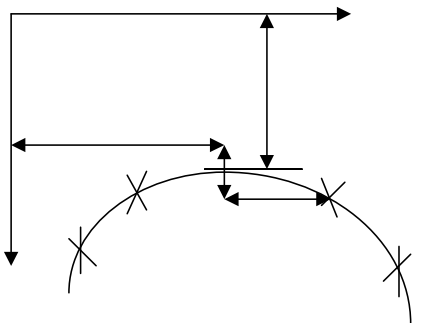

图 3　拼音识别的特征提取

本系统中的文字识别核心是一个基于提取字符统计特征和多模版欧式距离分类器的单字识别系统。特征由格点方法从字符结构和时间空间信息中得到。实验结果表明，按此方法得到的特征具有较强的分辨能力[3]。

3　基线估计

手写字符占据的空间可以划分成 3 个区域，即中间区、上升区和下降区，如图 4 所示。

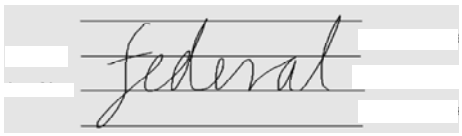

图 4　归一化过程所需的区域划分

在不对书写者施加特殊限制的情况下，当削减尺寸或倾斜角度发生变化时，先要检测出上述 3 个区间的位置。因为特征最密集的中间区需要在归一化过程中保持稳定，所以对中间区的检测很重要。

在文献[2,4]中，中间区由水平方向投影的峰值决定(图 5)。处理较短的单词时，该方法效果不理想，此时要找到一个足够明显的峰值难度较大。

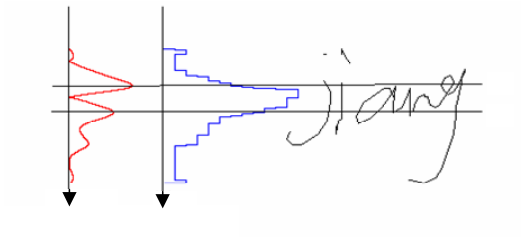

图 5　中间区的检测

本文对文献[2,4]的方法进行融合和改进，简要描述如下：

(1)判定水平投影中是否有明显的峰值，估计基线和中线位置的上下界。

(2)当峰值明显时，用导数判决中线和基线。

(3)没有明显的峰值时做如下处理：

1)找到 y 轴方向的极大值点和极小值点，利用 K 均值聚类方法得到 4 条基线。

2)当处理只占据 2 个区间的单词时，步骤 1)得到的分割线顺序不一定准确，此时可以使用上下界划分区间。

3)如果聚类结果在上下界构成的区间外，则使用水平极点的平均值作为基线和中线。

4　拼音的HMM

在隐马尔可夫理论中，每个字符可以视为一个马尔可夫链(图 6)，字符的 HMM 从左至右连接起来形成拼音单词的 HMM[4]。

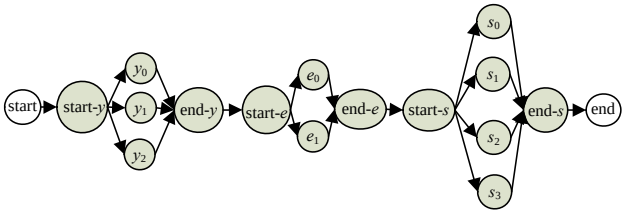

图 6　拼音单词的 HMM

与字符识别相似，单词识别的基本原理是找到适当的单词，使通过其 HMM 模型，λ*对应最佳路径的后验概率最大，即

$$\lambda^* = \arg\max_{\lambda} P(O \mid Q^*, \lambda), \qquad Q^* = \arg\max_{Q} P(O \mid Q, \lambda)$$

Viterbi 算法[5]可以找到一个马尔可夫链的最佳译码顺序和最佳匹配分数。若对词典里所有单词最佳路径的后验概率进行计算，则时间复杂度会很大，即

$$O(f \times w \times \bar{l} \times \bar{a} \times \bar{s})$$

其中，f 代表帧数[2]；w 代表单词个数；$\bar{l} \times \bar{a}$ 表示字符的平均异型体数目；s 代表每个 HMM 的平均状态数；$w \times \bar{l}$ 代表词典中的字符数目。处理稍大词典的识别问题时，该方法是不现实的。因此，本文采用如下策略：连接词典中所有单词的 HMM，组成一个更大的 HMM，所有马尔可夫链组成一个有限状态机，图 7 给出了此状态机的结构。

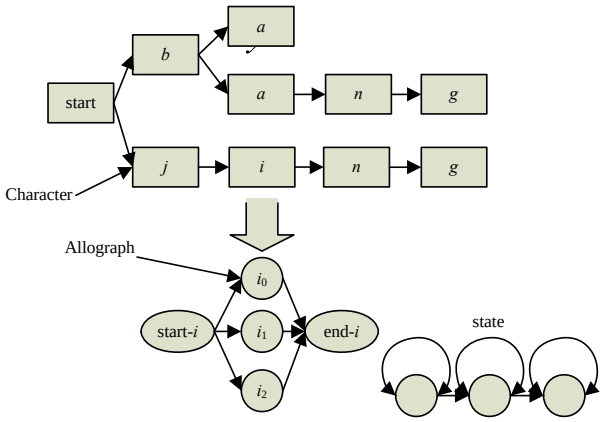

图 7　有限状态机

有限个状态构成异型体 HMM，有限个异型体 HMM 构成单字 HMM 模型。单字 HMM 模型结合在一起构成拼音单词的 HMM。所有拼音单词 HMM 构成整个词典。对此状态机结构采用 Viterbi 译码来获得最佳的 N 个候选单词。本文对词典的储存在字符级别上采用树状结构(图 7)，可以使拼音词典的节点数减少 64%。

5　分类器融合

为了提高识别效率，本文融合了不同分类器的识别结果。由于没有其他拼音识别系统，因此为区别于直接将手写数据

输入其他分类器，本文采用如下策略：先使用拼音单词 HMM 模块产生的分割结果将拼音单词切分成分开的单字，单字会被放入字符核心(基于统计方法)，字符识别结果的累加作为拼音单词的识别距离。

本文工作采用下式计算融合后的匹配分数：

$$S_w = P_w + v\sum_i(-D_{c_i})$$

其中，P_w 代表由候选单词 w 的单词 HMM 模块产生的后验概率；D_{c_i} 代表由统计识别核心对单词中第 i 个字符产生的识别距离；权值 v 可通过实验取令样本错误率最低的值而得到。

6 实验结果

本文用实验室收集的拼音样本(表 1)来评估以上工作，采用的训练集包括来自 41 个书写者的 44 832 个无限制手写拼音样本。测试集包括来自 16 个书写者的 17 745 个无限制手写拼音样本。本系统采用的拼音词典共包括 467 个单词。

表 1 本文采用的样本数

	书写者	样本
训练集	41	44 382
测试集	16	17 745

统计识别核心性能如表 2 所示。

表 2 统计识别核心性能 (%)

字符集	首选识别率	五选识别率
大写字母	94.01	99.28
小写字母	97.08	99.79

实验结果表明，与笔者曾开发的基于规则切分与统计方法的拼音单词识别系统相比，本文方法得到了更高识别率。Motorola 公司将 HMM 分类器和基于模板的分类器直接结合起来[1]，获得了 87.15%的识别率。表 3 给出了本文系统和原有系统的准确率。由于对拼音识别没有公用标准，因此很难将本文工作与上述其他工作做出有效比较，但 91.37%的首选识别率是可接受的并能付诸实际应用。

表 3 拼音单词识别系统的准确率 (%)

系统	首选识别率	五选识别率
HMM 模型与统计方法融合	91.37	97.55
规则切分与统计方法融合	81.06	92.01

7 结束语

本文实现了一个具有一定实用意义的手写拼音单词识别系统。在下一步工作中，笔者将尝试找到鲁棒性更强的帧划分方式，并充分利用手写笔迹断点的信息(在 Viterbi 译码中没有充分利用)来改进识别率。本文采用的一些方法也可以应用于手写英文单词识别。

参考文献

[1] Ge Yong, Guo Fengjun, Zhen Lixin, et al. Online Chinese Character Recognition System with Handwritten Pinyin Input[C]//Proc. of 2005 International Conference on Document Analysis and Recognition. Seoul, Korea: [s. n.], 2005: 1265-1269.

[2] Homayoon S M B. Pre-processing the Dynamics of On-line Handwriting Data, Feature Extraction and Recognition[C]//Proc. of the 5th Int'l Workshop on Frontiers in Handwriting Recognition. Colchester, England, UK: [s. n.], 1996: 255-258.

[3] Bai Zhenlong, Huo Qiang. A Study on the Use of 8-directional Features for Online Handwritten Chinese Character Recognition[C]//Proc. of 2005 International Conference on Document Analysis and Recognition. Seoul, Korea: [s. n.], 2005: 262-266.

[4] Biem A. Minimum Classification Error Training for Online Handwriting Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1041-1051.

[5] Homayoon S M B, Nathan K, Clary G J, et al. Size Normalization in On-line Unconstrained Handwriting Recognition[C]//Proc. of ICIP'94. Austin, Texas, USA: [s. n.], 1994: 169-173.

编辑 陈 晖

(上接第 169 页)

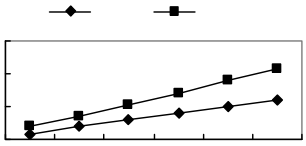

图 4 CHC 与 SGA 的运行时间

表 1 为实验所用 6 个数量属性分别用 2 种方法得到的近似最优解及相应的适应度值。

表 1 遗传进化结果

数量属性	遗传模型	近似最优解	适应度值
湿度	CHC	69,22,69,82,69,82,86,85	1.47
	SGA	58,22,58,74,58,74,86,78	1.24
降水量	CHC	71,0,71,152,130,152,166,152	0.97
	SGA	97,0,97,157,143,157,179,166	0.94
气压	CHC	10 291, 9 951, 10 291, 10 333, 10 315, 10 333, 10 342, 10 340	1.53
	SGA	10 154, 9 951, 10 155, 10 257, 10 157, 10 257, 10 342, 10 285	1.26
日照时数	CHC	98,27,98,113,111,113,114,113	1.54
	SGA	79,27,79,96,79,96,114,101	1.44
气温	CHC	164,-76,165,284,165,284,296,290	1.29
	SGA	123,-76,123,252,123,252,296,255	1.16
风速	CHC	39,9,39,46,43,46,47,46	1.46
	SGA	28,9,28,38,28,38,47,39	1.35

4 结束语

本文基于实数编码的 CHC 遗传模型对数量属性进行模糊划分，减少了遗传进化所需时间，且产生的模糊 1-频繁项个数和最优个体的适应度值较高。下一步工作将对数量属性的模糊预处理方法与其他预处理方法进行比较，并将其应用于基于模糊关联规则的分类系统。

参考文献

[1] Hong Tzung-Pei, Chen Chunhao. A GA-based Fuzzy Mining Approach to Achieve a Trade-off Between Number of Rules and Suitability of Membership Functions[J]. Soft Computing, 2006, 10(11): 1091-1101.

[2] Hong Tzung-Pei, Chen Chunhao. Genetic-fuzzy Data Mining with Divide-and-conquer Strategy[J]. IEEE Transactions on Evolutionary Computation, 2008, 12(2): 252-264.

[3] Alcalá-Fdez J, Alcalá R, Gacto M J, et al. Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms[J]. Fuzzy Sets and Systems, 2009, 160(7): 905-921.

[4] Ballester P J, Richards W G. A Multiparent Version of the Parent-centric Normal Crossover for Multimodal Optimization[C]// Proc. of IEEE Congress on Evolutionary Computation. Vancouver,

Canada: [s. n.], 2006: 2999-3006.

编辑　陈　晖