一个新的脱机手写汉字数据库模型及其应用

郭 军, 蔺志青, 张洪刚
(北京邮电大学信息工程系, 北京 100876)

摘 要: 本文提出一个新的脱机手写汉字数据库模型. 此模型的特点在于将汉字样本信息与其书写者信息结合起来, 因而既可为开发手写汉字识别算法提供训练和测试样本, 也可用于研究各类人员的文字书写特征, 探讨文字书写的相关因素. 本文还介绍了一个应用此模型的实例 HCL2000, 并利用 HCL2000 研究了影响识别率的相关因素, 并获得了一些令人感兴趣的结果.

关键词: 手写汉字数据库; 文字识别; 书写特征
中图分类号: TP391. 1 文献标识码: A 文章编号: 0372-2112 (2000) 05-0115-02

A New Database Model of Off-line Handwritten Chinese Characters and Its Applications

GUO Jun, LIN Zhi-qing, ZHANG Hong-gang
(*Dept. of Information Engineering of Beijing University of Posts & Telecomm. Beijing* 100876, *China*)

Abstract: A new database model of off-line handwritten Chinese characters is proposed. The new model not only contains the samples of Chinese characters, but also contains the information of their writers. So it can provide the learning and testing samples for developing the recognizing algorithm, meanwhile it also can be used for researching the writing features of different people. HCL2000, a database developed according to this model, is discussed.

Key words: database of handwritten Chinese characters; recognition of characters; writing features

1 引言

建立手写汉字数据库是研究和开发手写汉字识别技术的基础. 一个有影响的手写汉字数据库可以为开发手写汉字识别算法提供公共的训练和测试样本, 使各家的实验结果具有可比性, 促进技术的竞争和发展. 日本的 ETL8、ETL9[1] 以及中国科学院自动化所[2] 研制的数据库便起到了这样的作用. 但是, 现有的数据库只有手写汉字样本信息, 随着研究的深入, 这样的数据库已经不能满足研究者的需求. 因为无法用它深入研究人的哪些因素与汉字书写及识别有关, 以及有多大的相关性. 这样的问题不仅令人感兴趣, 也对开发更高水平的识别技术有重要意义. 为此, 本文提出了一个新的脱机手写汉字数据库模型, 在这一模型中, 既存储手写汉字样本信息, 也存储它们的书写者信息, 两种信息可以随时互查. 从而为研究上述问题提供了方便. 采用这一模型, 在国家 863 计划的资助下, 我们研制开发了一个大规模的脱机手写汉字数据库系统 HCL2000, 它包含 3755× 1300 个手写汉字样本和 1300 个书写者的个人信息. 应用 HCL2000, 我们可以研究各类人员的文字书写特征及影响识别率的相关因素.

2 系统模型

图 1 为本文所提出的脱机手写汉字数据库系统模型.

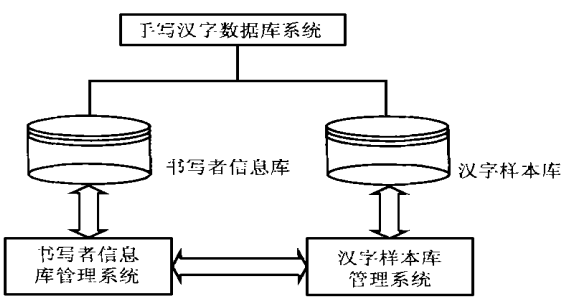

图 1 脱机手写汉字数据库系统模型

系统由汉字样本库和书写者信息库两大部分组成. 汉字样本信息和书写者信息分开存放, 既有利于相对独立地使用这两类信息, 也便于分别采用普通的高级语言和数据库语言开发. 为了实现汉字样本信息和书写者信息间的互查, 需要提供汉字样本库管理系统与书写者信息管理系统之间的相互调用机制.

3 信息组织方式和互查方法

3.1 汉字样本信息的组织

汉字样本的组织既要考虑到便于使用, 还要考虑便于与书写者信息互查. 为此可将汉字样本信息以书写者为单位按

收稿日期: 1998-12-28; 修订日期: 1999-03-16
基金项目: 国家 863 计划资助课题

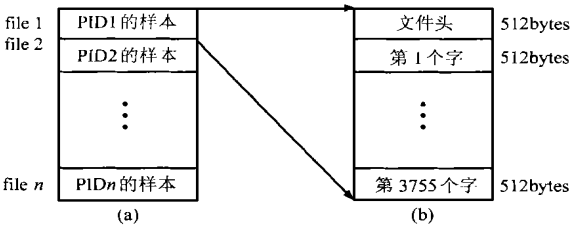

图 2 汉字样本信息的组织

文件形式组织与存放. 即同一个书写者所写的样本按区位码的顺序放在一个文件中. 全部汉字样本文件便构成了汉字样本信息库, 如图 2(a)所示. 每个文件有一个 512 字节的文件头, 用来存放书写者标识(PID)、水平和垂直方向扫描的精度、整个文件的大小等信息. 每个汉字样本用 64×64 个二值像素描述, 占用 512 个字节. 如图 2(b)所示.

3.2 书写者信息的管理

书写者信息除书写者标识信息外, 主要应包括与文字书写有关的信息, 如性别、年龄、职业、文化程度、书写工具等. 为了便于统计分析, 这些信息利用数据库管理系统进行管理.

3.3 两种信息的互查方法

系统按如下方法实现互查: 在浏览手写汉字样本信息时查看它的书写者信息的方法是, 通过样本文件的文件名得到该书写者的标识(PID), 调用书写者信息库管理系统打开书写者信息 DBF, 用该 PID 将书写者信息读取出来. 相反, 在浏览书写者信息时查看某个书写者所写的字的方法是, 将该书写者的 PID 取出, 调用汉字样本库管理系统打开并显示该书写者的汉字样本文件.

4 一个实例: HCL2000

按照第 2 节中提出的模型, 在国家 863 计划资助下, 我们建立了一个大规模的脱机手写汉字数据库 HCL2000. HCL2000 面向 GB2312-80 一级汉字, 包含 3755× 1300 个手写汉字样本和 1300 个书写者的个人信息. 表 1、表 2 和表 3 中的数据描述了 HCL2000 中书写者的职业、年龄和文化程度的分布情况.

表 1 职业分布 (单位: 百分比)

工人	17.1	农民	2.7	学生	25.0	行政人员	11.7
科技人员	4.7	商业人员	4.5	教师	5.5	军人	13.3
文艺类	0.9	医生	2.2	其它	12.4		

表 2 年龄分布 (单位: 百分比)

20 岁以下	25.7	20~ 29 岁	41.9	30~ 39 岁	12.4	40~ 49 岁	12.7
50~ 59 岁	5.0	60~ 69 岁	1.8	70~ 79 岁	0.5		

表 3 文化程度分布 (单位: 百分比)

研究生	2.2	大学	13.9	大专	17.0	中专	17.5
高中	23.5	初中	23.2	小学	2.7		

为了研究影响文字书写和识别的相关因素, 我们对各类人员所写文字的识别率分别进行了统计, 通过比较, 获得了表 4 至表 7 的结果. 所用的识别方法是文献[3]中提出的.

表 4 男性与女性的平均识别率

男性	87.50	女性	89.49

表 5 各种职业的平均识别率

工人	90.16	农民	89.11	学生	86.68	行政人员	89.42
科技人员	91.07	商业人员	90.39	教师	90.55	军人	87.67
医生	85.49	其它	87.33				

表 6 各年龄段的平均识别率

20 岁以下	87.81	20~ 29 岁	87.35	30~ 39 岁	89.89	40~ 49 岁	90.65
50~ 59 岁	89.11	大于 59 岁	89.21				

表 7 各种文化程度的平均识别率

初中以下	88.24	高中	88.66	中专	88.95	大专	89.83	大学以上	86.97

通过表 4 至表 7, 我们发现手写汉字的识别率与书写者的性别、年龄和职业相关性较大, 而与文化程度的相关性相对较小. 这些结果中, 有些是与常识相符的, 比如女性字的识别率较高, 医生字的识别率低. 而有些结果却是富有新意和耐人寻味的, 比如大学以上学历的人的字的识别率低, 中年人的字的识别率比青年人和老年人都高等. 当然, 本文这些结果还是初步的, 还需要进一步深入地加以分析和研究. 但我们相信, 定量地揭示相关因素对识别率的影响, 会对我们开发高精度的识别系统提供新思路和新方法. 除此之外, 利用 HCL2000, 我们还可以研究不同书写者的文字书写特征, 为笔迹鉴别等应用领域提供参考. 由此可见, 与现有的手写汉字数据库相比, 按照本文提出的模型所研制的 HCL2000 具有更大的用途.

参考文献

[1] T. Saito, H. Yamada, T. Yamamoto. On the Data Base ETL9 of Hand-printed Characters in JIS Chinese Characters and Its Analysis. IEICE of Japan Trans., 1985, J68-D(4): 757~ 764

[2] 刘迎建, 戴汝为等. 几个大规模文字识别样本库. 第四届全国汉字及汉语语音识别学术会议论文集, 1992, 5: 35~ 43

[3] J. Guo, N. Sun, Y. Nemoto. Recognition of Handwritten Characters Using Pattern Transformation Method with Cosine Function. IEICE of Japan Trans. 1993, J76-D-Ⅱ(4): 835~ 842

郭 军 北京邮电大学教授、博士生导师. 中国电子学会高级会员、IEICE 会员. 日本东北学院大学博士. 主要研究领域: 模式识别、中文信息处理、网络的智能控制.

蔺志青 北京邮电大学副教授. 主要研究领域: 计算机应用、移动卫星通信.