脱机手写体汉字识别综述

赵继印[1],郑蕊蕊[2],吴宝春[1],李 敏[1]
(1.大连民族学院机电信息工程学院,辽宁大连 116600;2.吉林大学通信工程学院,吉林长春 130025)

摘 要: 脱机手写体汉字识别是模式识别领域最具挑战性的课题之一.本文分析了近年来脱机手写体汉字识别的最新进展,讨论了脱机手写体汉字分割、特征提取和分类器设计等关键技术的各种主流方法,介绍了 3 种典型的汉字识别数据库,并提出了脱机手写体汉字识别的难点问题和今后发展的趋势,为该领域的研究者指明研究方向,共同促进脱机手写体汉字识别技术的发展.

关键词: 脱机手写体汉字识别; 字符分割; 特征提取; 分类器设计; 汉字识别数据库
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2010) 02-0405-11

A Review of Off-Line Handwritten Chinese Character Recognition

ZHAO Ji-yin[1], ZHENG Rui-rui[2], WU Bao-chun[1], LI Min[1]
(1. *College of Electormechanical and Information Engineering, Dalian Nationalities University, Dalian, Liaoning* 116600, *China*;
2. *College of Communication Engineering, Jilin University, Changchun, Jilin* 130025, *China*)

Abstract: Off-line handwritten Chinese character recognition is one of the most challenging problems in pattern recognition field. This paper analyzed the latest developments of off-line handwritten Chinese character recognition in recent years. Main methods of the key technologies such as Chinese characters segmentation, feature extraction and classifier design were discussed. This paper also introduced 3 typical off-line handwritten Chinese character recognition databases. Finally, remain difficult issues and future trends of off-line handwritten Chinese character recognition were proposed. This paper will guide researchers in this field and promote development of off-line handwritten Chinese character recognition technology.

Key words: off-line handwritten Chinese character Recognition; characters segmentation; feature extraction; classifier design; Chinese recognition database

1 引言

汉字识别是模式识别的一个重要分支,也是文字识别领域最为困难的问题之一,它涉及模式识别、图像处理、统计理论等学科,呈现出综合性的特点,在办公和教学自动化、银行票据自动识别、邮政自动分拣、少数民族语言文字信息处理等技术领域,都有着重要的理论意义和实用价值[1].汉字识别技术可分为印刷体和手写体汉字识别两大类.手写体汉字识别又可分为联机(on-line)和脱机(off-line)手写体汉字识别.脱机手写体汉字识别可分为受限和非受限两种情况,如图 1 所示.

清华大学、中科院自动化所等著名高校和科研院所都致力于汉字识别的研究,以汉王科技股份有限公司为首的科技企业也推出了一系列成熟的商业产品[2].目前,很多论文提出的脱机手写体汉字识别的方法在不同的字符数据库试验中,取得了 95% ~ 99% 的识别率,但是对真正的手写文档的识别效果却难以达到实际应用的要求.目前脱机手写体汉字识别仍处于实验室研究阶段,成功的商业产品仍未发布[2~4].本文着重讨论脱机手写体汉字识别的现状和存在的问题,明确今后的发展趋势,为脱机手写体汉字识别领域的广大研究人员提供参考和借鉴.

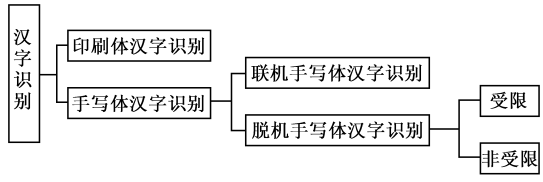


图1 汉字识别的分类

2 手写汉字字体特点

从识别的角度分析,汉字具有如下 4 个特点.

2.1 汉字类别多

汉字的个数很多,国家标准 GB18030-2000《信息交换用汉字编码字符集基本集的扩充》收录 27484 个汉字[5].汉字个数在模式识别问题中体现为汉字的类别,因此汉字识别问题属于超大规模数据集的模式识别问题.

收稿日期:2009-02-16;修回日期:2009-06-13
基金项目:大连民族学院科研基金(No.20086201);吉林省科技厅科技引导计划(No.20090511)

2.2 字体结构复杂

汉字基本笔画(stroke)分为:横、竖、撇、点、折[6].笔画的组合方式分为相离、相接和相交三种.特征结构笔画和相应笔画的组合方式在以笔画为特征的汉字识别中起到关键作用.

汉字的组合方式包括独体字和合体字.合体字又包括上下结构、左右结构、品字结构等多种结构.以部件(radical)为基础的手写体汉字识别中,需根据汉字的组合方式对已提取的部件进行重新组合.

2.3 字形变化多

手写体汉字字形总的来说可以分为:手写印刷体(hand-print fashion scripts)、行书(fluent scripts)和草书(cursive scripts).对于相同的字形又因不同人书写风格的差异造成手写汉字的变形.脱机手写汉字在日常生活中以行书为主.对于行书和草书等笔迹相连的情况,字符分割是识别的关键环节,如果出现分割错误,将影响后续识别结果的精度.

2.4 相似字多

汉字集合中相似字较多,由于手写体汉字变形的存在,使得手写体中相似字的区分比印刷体要困难得多[1].比如,在手写体中的一点,可能会因为不当的预处理而消失,从而造成字符的误识.因此要求预处理方法能够针对手写汉字的特点,做到尽量不丢失笔画信息.在识别过程中,对于相似的字体,可以采用更精确的细分类过程进行鉴别.在识别后处理阶段,通常采用相似字符集作为候选字符集的主体.

3 识别过程

对于脱机手写体汉字识别而言,其识别过程通常如图 2 所示.

原始的手写文稿通过扫描仪等 OCR(Optical Character Recognition)设备,转换成灰度图像或者二值图像,并进行预处理.字符特征分为统计和结构特征两种模式.识别阶段,根据提取的特征,选择相应的分类器及其组合形式进行模式分类.识别后处理根据前后文字的上下文关系选择最合乎逻辑的字词,能进一步提高识别准确率,最后输出分类结果.

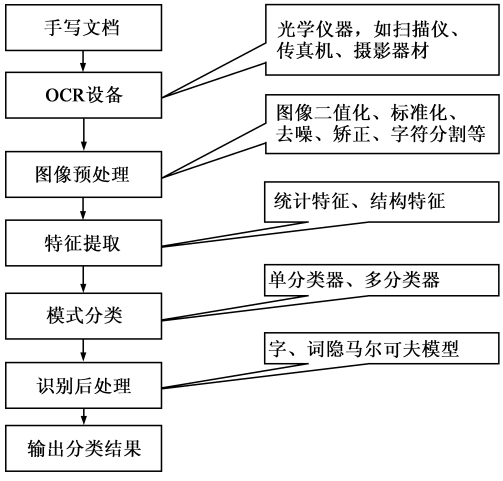


图2 脱机手写体汉字识别流程图

4 图像预处理

很多图像处理技术可以应用于脱机手写体汉字图像,包括:(对灰度图像)二值化、(对二值图像)伪灰度化、去噪、骨架化、边缘提取、倾斜矫正等.本节主要介绍字符图像分割的关键技术和方法.

基于切分的汉字识别方法是目前汉字识别的主流方法.汉字的分割通常首先对整篇文档做行切分,再在行分割的基础上进行单个字符的分割.图 3 显示了手写体汉字分割处理的一般流程[2].

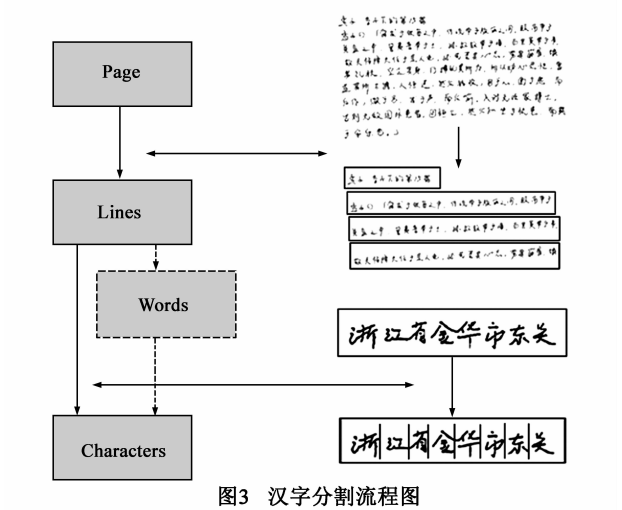


图3 汉字分割流程图

只有当每一单个字符的图像都能正确地从整个文本页面图像中分割出来,才有可能进行正确的文字识别[5].然而,手写体汉字的书写随意性很大,相邻汉字之间的位置关系也复杂多样.手写体汉字的书写可能产生如下 4 种基本位置排列情况[7,8],如图 4 所示.

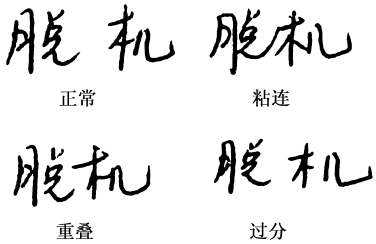


图4 手写体汉字4种书写位置关系

(1)正常:汉字各自分开独立为整体;单个汉字中各个部件间的距离远小于字间距离.(2)粘连:汉字的某一笔在一点或几点与相邻汉字接触;(3)重叠:汉字间无接触,但无法用垂直分割线分割;(4)过分:汉字左右部分间距过大或汉字内部出现笔画断裂.

真实手写文档上述情况往往同时出现,这是造成无法正确分割汉字的主要因素,对这种手写体汉字的切分是今后研究的重点和难点问题[8].目前手写体汉字分割广泛采用的方法如下:

4.1 投影法

投影法(Project Profile Histogram,PPH)通过统计图像中每一列(行)黑像素的个数得到投影直方图.在直方图中字符区域对应于波峰,字符间隔对应于波谷.投影法简单,速度快,对印刷体汉字和手写印刷体汉字的分割有相当好的效果,但是会将粘连或重叠的字符识别为一个字符,出现弱分割现象;或将过分字符识别为若干字符,产生过分割现象.

4.2 连通域分析法

连通域分析法[8](Connected Component Analysis,CCA)是在整个字符图像中寻找相连的像素作为连通元,分析这些连通元本身的图像属性,判断它们是否属于同一个字符图像,然后利用先验知识对它们进行拆分和合并.连通域分析法对于重叠字符和倾斜字符能够取得理想的分割效果.但使用该方法时连通元容易过碎,使严重断裂的字符图像无法重新合并,真正粘连的字符也不能通过连通元切分开,需在后续的识别模块中加入粘连字符模板或者通过其它方法进行再切分.

4.3 Viterbi 算法

字符分割路径可视为一个自上而下的 m 层单向图,建立一个隐马尔可夫模型(Hidden Markov Model,HMM)来表示该有向图[8,9].图中的每个节点对应隐含状态,有向边表示状态的转移方向,用节点轨迹组成观测序列,其概率分布为分割路径穿过结点的几率大小,m 是观测序列的长度.采用 Viterbi 算法[10~12]寻求分割路径,相当于在图中沿着有向边方向找出所有路径中的最大概率者,组成顺向首尾相接的一串有向边的集合,即得到非线性的分割路径.Viterbi 算法对于交错、单处笔划粘连等字符能够得到较好的分割效果,但并未从根本上解决多种粘连方式的分割问题.

4.4 基于识别的方法

将字符分割与识别截然分开,分割将是手写体汉字识别误差的主要来源,基于识别的统计分割方法是汉字分割的新出路[5].基于识别的方法首先将字符分成若干组成部分,并采用合并策略在多条候选的合并路径中通过识别结果选择一条最佳路径[13,14].基于识别的字符分割方法通过识别模块来指导切分,识别结果对分割起着决定性的作用,分割是识别的副产品[9],分割结果依赖于识别分类器的性能[13].

图像预处理会给字符图像带来干扰或形变,引入新的误差.改进的二值化、细线化、字符归一化、字符分割等图像预处理算法[15~17],能够减少预处理带来的字体变形等不利影响,但不能从根本上解决预处理带来的干扰.由于目前尚不能完全实现字符的正确分割,所以,对于基于分割的脱机手写体汉字识别,字符分割的精度直接决定后续汉字识别的精度,是手写体识别系统精度的瓶颈.文献[18]提出了一种无分割的手写体汉字识别方法,并通过实验证明了该方法的可行性.这种方法实质上是对文本进行行分割,再在行分割的基础上提取字符特征,而非精确到单个字符的分割.行分割相对字符分割简单,计算量小,引入误差更小.无分割脱机手写体汉字识别更符合人类识别字符的习惯,将是未来汉字手写体识别的新趋势.

5 特征提取

手写体汉字识别特征提取方法可分为基于结构特征、统计特征和将结构特征和统计特征相融合的方法.

5.1 结构特征

结构特征是汉字识别研究初期的主流方法,需要先抽取结构基本单元,再由这些基本单元构成来描述汉字特征.结构特征比较直观,符合人们书写汉字的过程,能较好地反映汉字的结构特性;缺点是对结构基本单元提取困难,各结构元素之间的拓扑关系复杂,抗干扰性较差.同时,由于汉字的结构特征通常都要利用细化算法提取,不仅计算量大而且会出现形变问题,给汉字识别带来新的噪声影响.

5.1.1 基于特征点

特征点是反映汉字形体特征整体分布状况的关键点.通常对大多数结构稳定的汉字,一旦获得了正确的特征点集,就可能顺利地按一定的策略和步骤(连接笔划、结构匹配等)将汉字形体划归为正确的字类.根据不同的研究思路,研究人员对特征点的定义也不尽相同[19~21].

5.1.2 基于笔画

一个汉字区分于其它汉字的主要特征就是笔画及其所在的位置,“横”、“竖”、“撇”、“捺”四种笔画的数量及其相对位置唯一地确定了一个汉字[22].基于笔画的特征提取方法将字符分解成笔画,并根据笔画的数量、顺序和位置进行识别[23~26].“横”、“竖”、“撇”、“捺”是构成汉字的四种基本笔画,所占比重大,并且提取容易,因而在识别系统中常采用它们作为识别特征.

5.1.3 基于部件

部件是一个居于笔画和单字之间的中间层次,相当于西文的字母.把若干个部件按照一定规则加以组合就可构成方块汉字.我国语言文字工作委员会对 GB130001 字符集中的 20902 个汉字逐个进行拆分、归纳与统计后,制定《汉字基础部件表》,共有 560 个可供独

立使用的部件.这 560 种部件并不都适用于汉字识别,通常从中选用若干部件作为识别特征[27].文献[28]提出的基于部件的汉字分解示意图,如图 5 所示.图中的 4 个汉字具有相同的 3 个部件,可根据最后一级分解部件来进行识别.

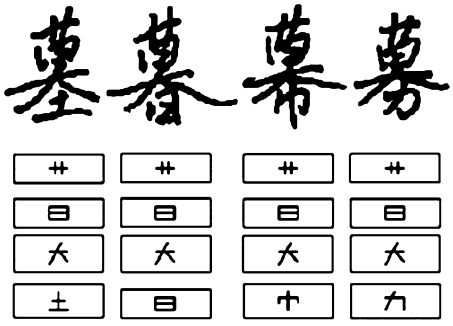


图5 部件分解示意图

5.2 统计特征

统计特征一般针对单个汉字,即整字(Holistic),提取方便,抗干扰能力强.文献[5]指出,汉字结构的复杂,在统计识别方法中,不仅不是缺点,而且使得汉字具有比其他西方文字具有更强的鉴别能力,不仅可以识别成千上万个超多类汉字,而且具有高抗干扰和高鲁棒识别性能,这是结构分析方法无法达到的.统计特征的缺点是没有充分利用汉字的结构信息.本节针对脱机手写体汉字主流的统计特征方法进行介绍.

5.2.1 弹性网格特征

弹性网格特征(Elastic Mesh,EM)用一种弹性网格将汉字图像分块,对每一块内的像素进行变换或者分析后产生特征向量[29~32].对字符进行弹性网络的划分能有效地反映汉字的结构细节和字符的共同特征,避免手写体汉字中因个人书写风格差异引起的字体变形和因数据采集、非线性变换等因素导致的样本变形等问题.但该方法各个块之间互不关联,不能体现汉字的整体结构信息.

5.2.2 方向线素特征

方向线素特征(Directional Element Feature,DEF)首先抽取汉字的轮廓,并考察轮廓点像素的 8 邻域内的黑像素点在水平、垂直、+45°、-45°四个方向上的分布情况.如有符合四个方向上的任一种情况,则该像素对应方向上的方向线素值加一个常数[33~35].方向线素特征同时反映了字符的结构和统计特征,比较全面地代表图像信息,是汉字识别领域一种成熟的特征提取方法.但方向线素特征的特征维数多,在进行特征匹配之前要对特征向量进行降维处理,增加了识别算法的复杂度.

5.2.3 Gabor 特征

Gabor 滤波器是窄带带通滤波器,有明显的方向选择和频率选择特性,能在空域和时域同时达到最优联合分辨率[31],因此 Gabor 滤波器在脱机手写体汉字识别中提取特征方面得到了广泛应用[36~38].Gabor 变换提取汉字特征充分反映了笔画结构在空间上的局域性,笔画的方向性以及在频域上笔画与干扰的可分性等重要特性,提高了识别算法的鲁棒性和对细节的分辨率.Gabor 滤波器缺点在于特征提取时间较长且提取的特征数据存在冗余性,需通过主成分分析等方法进行压缩.

5.2.4 矩特征

脱机手写体汉字识别中采用 Hu 不变矩、Legendre 矩、Zernike 矩、Krawtchouk 矩、小波矩[3946].Hu 矩为非正交矩,含有大量冗余信息.正交矩对模式具有位移、旋转和变换不变性,在应用中最具代表性的是 Legendre 矩和 Zernike 矩.CHO-HUAK THE 和 ROLAND T. CHIN 对[47] Legendre 矩和 Zernike 矩在噪声敏感性、信息冗余和图像表示能力三方面进行了实验对比和理论分析,结论表明 Zernike 矩的效果在各方面都优于 Legendre 矩.Zernike 矩可以任意构造高价矩,因而包含更全面的图像信息,所以 Zernike 矩识别效果更好.与 Zernike 矩和 Legendre 矩等连续正交矩特征相比,Krawtchouk 矩是数字域的离散正交矩,不存在数字化过程中所带来的近似误差问题,在计算过程中不需要进行坐标转换,而且构造简单,更加适合用来描述数字图像[46].小波矩能同时得到图像的全局特征和局部特征,因而在识别相似形状的物体时有更高的识别率[48,49].

对于手写体汉字识别,单独运用结构特征和统计特征中的任何一种单一的特征,必然存在识别的盲区.将汉字结构特征和统计特征等多种特征相结合,可以实现各种特征的优势互补,能够更全面地反映汉字的特征.特征融合后的脱机手写体汉字通常具有多维的特征,增加了识别算法的计算复杂度,因此普遍采用 PCA,LDA 和 FDA 等方法[50~52]对特征向量进行降维处理后再送入分类器分类.多特征融合的方法成为手写体汉字识别特征提取的主流方法[53~56],是未来发展的必然趋势.如果能够借鉴相关领域的研究成果,引入更适于手写体汉字的特征描述方法,特别是能够直接从原始字符图像提取的特征,将简化图像预处理步骤,减少因预处理带来的误差,进一步提高脱机手写体汉字的识别精度.

6 分类器设计

手写体汉字识别的对象是几千个(种)汉字,脱机手写体汉字识别常用的分类器可分为单分类器和多分类器集成两种.多分类器集成的方法是目前的主流技术,同时也是未来的发展趋势.

6.1 单分类器

6.1.1 改进的二次判别函数

改进的二次判别函数(Modified Quadratic Discrimina-

tion Function,MQDF)分类器以一个 Gauss 分布去描述每个类的样本分布,直接采用常数代替偏小特征值,有效地缓解了小特征值估计误差所带来的系统性能下降[5].基于统计模型的 MQDF 分类器便于设计与实现,且具有很好的鲁棒性和较高的识别准确率,因此在脱机手写体汉字识别中得到广泛的应用[57,58].

6.1.2 支持向量机

支持向量机(Support Vector Machine,SVM)根据 Vapnik 提出的结构风险最小化原理,通过最大化分类间隔,使学习机的泛化性能尽量提高,其优越性在理论和实验方面都得到了深入地研究和验证.SVM 是一个两类问题的判别方法,在对多类问题实现分类时,采用一对一、一对多、SVM 决策树和有向无环图支持向量等分解策略,因此 SVM 的计算复杂度和时间复杂度较大,一般不用于直接分类[52].针对这个问题,目前研究者[15,59~63]提出了如下解决方法:(1)采用 SVM 作细分类;(2)将大规模字符集划分成小的子集;(3)采用多种 SVM 算法的改进形式.引入各种改进的快速 SVM 多分类算法[64~66]到脱机手写体汉字识别领域中,也能够提高识别的速度.

6.1.3 人工神经网络

人工神经网络(Artificial Neural Network,ANN)具有并行处理、自组织、自适应和学习能力,被广泛的应用于脱机手写体汉字识别领域,包括:BP 网络[67]、多层感知器网络、模块径向基神经网络[50]、自适应振荡神经网络[68,69]、Hopfield 网络、自组织特征映射网络等[1].ANN 用于大字符集分类时,训练时间和分类时间太长,一般不用于直接分类[52].文献[70]针对这个问题,提出了在预分类结果的基础上,采用简化的样本集对 ANN 进行训练的策略,在不降低识别效果的基础上,大大缩短了 ANN 在大样本集上的收敛时间.

6.1.4 隐马尔科夫模型

常用的统计语言模型是建立在将文本语言看作为字或词的不同阶的马尔可夫链的基础上,语言相关模型的参数可以通过大型语料库的学习而获得.语言模型和单字识别结果的可信度结合,利用 Viterbi 算法,获得在考虑上下文信息的语言模型条件下的最优文本识别结果.由于实际资源的限制,实际系统中往往采用字或词的一阶或二阶马尔科夫模型[71,72].隐马尔科夫模型适合于大规模分类,缺点是尚缺乏公认权威的语言模型.目前广泛应用的是对某种特定领域进行小规模的建模,如邮政地址系统,银行手写支票金额的模型.

6.2 多分类器集成

多分类器集成算法通过特定的组合方式,能够对单分类器取长补短,发挥各个组成分类器的最大优势.多分类器集成算法中每一个组成的分类器称为元分类器,可以采用 6.1 节介绍的任何一种单分类器的形式.集成算法根据其结构可分为串行和并行结构两类.

6.2.1 串行结构

串行结构的集成算法[73]是根据汉字识别特点对整个识别过程进行分级,或分阶段处理.前一级的输出结果是后一级的输入,后一级识别是对前一级识别的细化和延续,实现多特征多方法的互补以及多识别级间信息的利用,以进一步提高汉字识别率.

6.2.2 并行结构

并行结构的集成算法首先构造多个分类器,这些分类器基于不同特征、不同分类器形式或是不同训练样本集合,每个分类器独立训练,相互之间没有影响.针对各分类器的输出结果,采取一定的规则进行融合或表决,得到最终的输出结果.常用的表决策略有投票法、D-S(Dempster-Shafer)法、行为知识空间法、综合集成法、基于置信度的神经网络集成法等[1].从模式识别的观点来说,汉字识别是一种超多类的模式集合,已有的适用于模式类别较少的识别方法和理论已不完全适用[27].应选择针对大规模数据集的分类方法或者对汉字类别进行合理的划分,以适应目前的分类方法.采用结合了串、并行结构的混合结构多分类器集成对脱机手写体汉字进行分类是未来的发展趋势.串行分类器具有分类递进,后级分类器能够弥补前一级识别的不足,实现细节上的互补的优点;并行分类器能够在全局的分类器输出结果间取得整体上的平衡.因此,采用混合结构的多分类器集成策略,能够实现细节与整体上的双保险,从而提高脱机手写体汉字识别的精度.

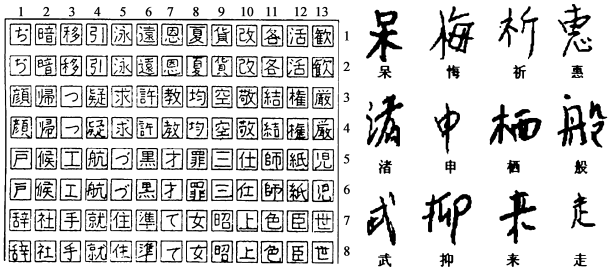
7 数据库

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.目前国内外一些研究团体已建立并公开了大规模的字符识别数据库.脱机手写体汉字识别的结果在这些数据库上实验,更有利于公正客观地对比实验结果,促进汉字识别技术研究的深入与发展.目前,具有典型代表性的数据库有以下几种.

7.1 ETL 字符数据库

ETL 字符数据库由日本电子工业发展协会(Japan Electronic Industry Development Association,现在的 Japan electronics and information technology industries association)、大学和研究机构联合协助的电工技术实验室(Electrotechnical Laboratory,现在的 Tsukuba central 2,national institute of advanced industrial science and technology,AIST)收集[74].ETL 数据库包含了 120 万手写和机器印刷字符图片,涵盖了用于识别研究的日文,中文,拉丁文和数字字符.数据库图片分别有 60×60,64×63,72×76,和

128×127 不同像素规格.字符图片文件包含不止一个记录,每个记录有一个字符图片和对应的 ID 信息的编码.该数据库不包含书写者信息.图 6 是 ETL8 中的字符样本[24]和 ETL9B 数据库中的部分字符[56].



A ETL8字符样本 B ETL9B字符样本

图6 ETL数据库样本

7.2 HCL2000 数据库

HCL2000 数据库[75]是由北京邮电大学信息工程系在国家 863 计划的资助下研发的一个大规模脱机手写汉字数据库系统.该数据库面向一级汉字,包含了 3755×1300 个手写汉字样本和 1300 个书写者的个人信息,可实现汉字样本信息和书写者信息间的互查,为研究各类人员的文字书写特征及影响识别率的相关因素提供了方便.每个汉字样本采用 64×64 个二值像素描述,占用 512 字节.书写者信息除书写者标识信息外,还包括性别、年龄、职业、文化程度、书写工具等.图 7 是两幅来自于 HCL2000 数据库[76]的字符图片,编号分别为 Hh451 和 Hh453.HCL2000 数据库是目前我国汉字识别领域被广泛采用的数据库.

Hh451 Hh453

图7 HCL2000数据库样本

7.3 HIT-MW 数据库

HIT-MW 数据库[3,18]由哈尔滨工业大学计算机科学与技术学院开发.该库由 780 多个书写者在无监督的情况(无监督情况是指书写参与者与数据库收集者并不发生正面接触,而是通过邮寄等方式将数据库页面交与书写者,书写者按照自己习惯的书写规则在一块未经分格的区域书写题签上标注的内容,允许出现涂改、文本行倾斜和交叠等复杂手写现象)下书写完成,优化出合格的手写样本 853 份.HIT-MW 数据库字量为 186444 字(包括标点、字母和汉字),涵盖了大部分 GB2312-80 一级汉字,一定量的 GB2312-80 二级汉字,甚至 GB2312-80 字符集以外的少量汉字.图 8 和图 9 是两幅来自 HIT-MW 数据库的样本,编号分别为 b04090303 和 b04090902.

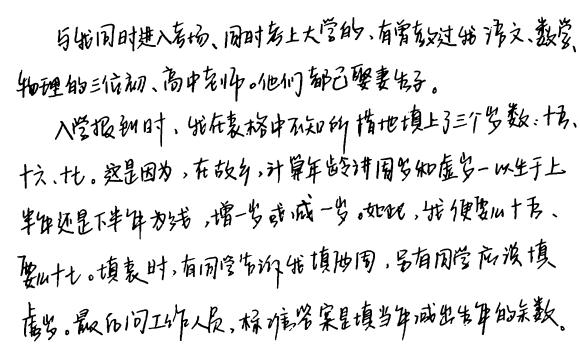


图8 HIT数据库b04090303号样本

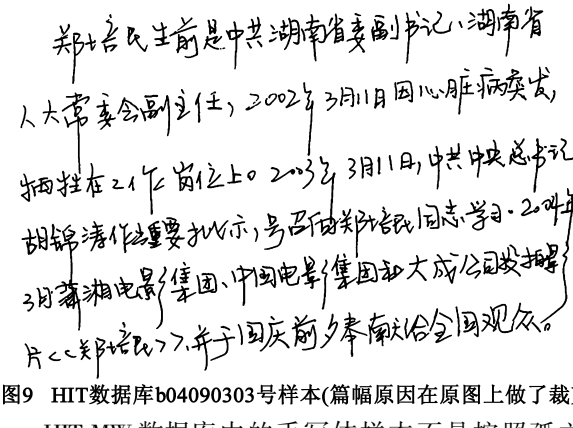


图9 HIT数据库b04090303号样本(篇幅原因在原图上做了裁剪)

HIT-MW 数据库中的手写体样本不是按照孤立的汉字书写,而是按照一定的规则从《人民日报》上随机抽取的一段 200 字左右具有一定含义的文字,因此可以看作是真实的手写体样本.迄今为止,HIT-MW 数据库已被美国 U C Berkeley,日本 Tokyo 大学,清华大学,吉林大学和华南理工大学等多家科研院所采用,应用领域主要集中在中文文档的行切分、汉字的切分识别、中文文本的无切分识别、笔迹鉴别和签名验证等方面.

8 识别后处理及评价准则

手写体汉字识别后处理一般是根据上下文关系对单字的识别进行处理.利用后处理技术,能够实现对单字识别结果的确认或者纠错,进一步提高整个汉字识别系统的正确率.目前主流的后处理技术包含以下 3 个步骤:(1)根据上下文关系建立基于词或字的 N 元语法(N-gram)统计语言模型,即 N-1 阶 Markov 模型.实践中最常见的是 bi-gram 或 tri-gram 模型[77~81];(2)确定并调整候选字的相似字集,作为候选字符集;(3)在候选字符集上,根据统计语言模型,以句子为处理单元,采用 Viterbi 算法选择具有最大概率的句子路径,从而确定相

应的汉字.

由于汉语语法的复杂性与灵活性,对通用的手写体汉字识别做出符合语法规则的模型很难.语言模型包含规则模型和统计模型两类.N-gram 语言模型是基于统计的语言模型,目前应用中占有绝对优势,研究较成熟.将两种模型相融合形成综合模型能够相互补充,同时也是未来的发展趋势[82,83].随着语言模型的不断完善,汉字识别后处理技术的精度必然能够实现新的突破;研究者还可以尝试将综合语言模型运用到汉字识别后处理中,以进一步提高后处理的效果.文献[84]提出一种结合传统统计语言模型和特定语言模型的自适应语言模型,能够充分利用已校对信息自动修正候选字符集,提高了后处理的正确率.

识别率、误识率和拒识率是识别系统的三个性能指标,它们之和应该等于 100%[27].

9　总结与展望

脱机手写体汉字识别技术发展迅速,特定场合的脱机手写体汉字识别系统的研究也逐步走向实用.本文分析总结了近年来脱机手写体汉字识别的最新进展,讨论了脱机手写体汉字分割、特征提取和识别分类器设计等关键技术的各种主流方法,介绍了汉字识别典型数据库,明确了脱机手写体汉字识别的核心技术呈现如下发展趋势:(1)改进图像预处理技术并简化图像预处理步骤,减少由于预处理引入的字体变形;(2)研究基于无分割的脱机手写体汉字识别技术,减少因字符分割引入的误差;(3)融合汉字的结构和统计特征,引入新的汉字特征描述方法,选择能够直接从原始字符图像提取的新特征;(4)采用混合结构的集成分类器,实现细节与整体上的双保险;选择针对大规模数据集的分类方法;对汉字数据集进行合理的优化,以适应目前的分类方法.

汉字识别经历了 40 余年的发展,目前在印刷体和联机汉字识别方面都取得了长足进步,商业产品趋于成熟,但脱机手写体汉字识别仍不能满足用户的实际要求.其难点集中于脱机手写体汉字的正确分割、特征提取和对超大规模数据集的分类问题.本文明确了脱机手写体汉字识别的难点和今后发展趋势,能够为研究者在该领域的研究指明方向,共同促进脱机手写体汉字识别技术的发展.

参考文献:

[1] 陈友斌,丁晓青,吴佑寿,等.非特定人脱机手写汉字识别[OL].中国计算机报,1997-06-23. http://media.ccidnet.com/media/ciw/663/01350001.htm,2008-06-27.

[2] Sargur N. Srihari, Xuanshen Yang, Gregory R. Ball. Offline Chinese handwriting recognition: an assessment of current technology[J]. Front Computer Science of China, 2007, 1(2): 137 – 155.

[3] Tonghua Su, Tianwen Zhang, Dejun Guan. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text[J]. International Journal on Document Analysis and Recognition, 2007, 10(1): 27 – 38.

[4] Tong-Hua Su, Tian-Wen Zhang, Hu-Jie Huang, et al. HMM-based recognizer with segmentation-free strategy for unconstrained Chinese handwritten text[A]. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)[C]. Curitiba, Brazil, IEEE Computer Society, 2007. 133 – 137.

[5] 丁晓青.汉字识别研究的回顾[J].电子学报,2002,30(9):1364 – 1368.
Ding Xiaoqing. Chinese character recognition: a review[J]. Acta Electronica Sinica, 2002, 30(9): 1364 – 1368. (in Chinese)

[6] 汉字结构[OL]-百度百科. http://baike.baidu.com/view/1137679.htm,2008-7-16.

[7] 高彦宇,杨扬.无约束手写体汉字切分方法综述[J].计算机工程,2004,30(5):144 – 146.
Gao Yanyu, Yang Yang. Survey of unconstrained handwritten Chinese character segmentation[J]. Computer Engineering, 2004, 30(5): 144 – 146. (in Chinese)

[8] 邵洁,成瑜.关于手写汉字切分方法的思考[J].计算机技术与发展,2006,16(6):184 – 190.
Shao Jie, Cheng Yu. A survey of methods in handwritten Chinese character segmentation[J]. Computer Technology and Development, 2006, 16(6): 184 – 190. (in Chinese)

[9] 马瑞.非限制手写字符分割中相关技术与算法的研究[D].南京:南京理工大学,2007.
Rui Ma. Research on segmentation of unconstrained handwritten characters[D]. Nanjing : Nanjing University of Science and Technology, 2007. (in Chinese)

[10] Zhizhen Liang, Pengfei Shi. A metasynthetic approach for segmenting handwritten Chinese character strings[J]. Pattern Recognition Letters, 2005, 26(10): 1498 – 1511.

[11] Yi-Hong Tseng, Hsi-Jian Lee. Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm[J]. Pattern Recognition Letters, 1999, 20(8): 791 – 806.

[12] 马瑞,杨静宇.一种有效的手写汉字多部分割方法[J].中国图像图形学报,2007,12(11):2062 – 2067.
Ma Rui, Yang Jing-yu. An effective multi-stage segmentation method for handwritten Chinese characters[J]. Journal of Image and Graphics, 2007, 12(11): 2062 – 2067. (in Chinese)

[13] Wuyi Yang, Shuwu Zhang, Haibo Zheng, et al. A recognition-based method for segmentation of Chinese character in images and videos[A]. 2008 International Conference on Audio, Lan-

guage and Image Processing, Proceedings(ICALIP2008)[C]. Shanghai, China, IEEE Computer Society, 2008.723 – 728.

[14] Guohong Fu, Chunyu Kit, Jonathan J. Webster. Chinese word segmentation as morpheme-based lexical chunking[J]. Information Sciences 2008, 178(9): 2282 – 2296.

[15] Jian-xiong Dong, Adam Krzyzak, Ching Y. Suen. An improved handwritten Chinese character recognition system using support vector machine[J]. Pattern Recognition Lettersm2005, 26(12): 1849 – 1856.

[16] 王建平,钱自拓,王金玲,等.基于数学形态学的图像汉字笔画细化和提取[J].合肥工业大学学报(自然科学版),2005,28(11):1431 – 1435.
WANG Jian-ping, QIAN Zi-tuo, WANG Jin-ling, et al. Chinese characters stroke thinning and extraction based on mathematical morphology[J]. Journal of Hefei University of Technology, 2005, 28(11): 1431 – 1435. (in Chinese)

[17] 方树名,张媛媛.免细化过程的脱机手写体汉字的动态信息提取[J].科技信息,2008(1):87 – 88.

[18] 苏统华.脱机中文手写识别-从孤立汉字到真实文本[D].哈尔滨:哈尔滨工业大学,2008.
Su Tonghua. Off-line recognition of Chinese handwriting: from isolated character to realistic text[D]. Harbin: Harbin Institute of Technology, 2008. (in Chinese)

[19] 周昌乐,张雄伟.一种基于段化的手写汉字特征点提取方法及其实现[J].电子学报,1997,25(5):57 – 60.
ZHOU Chang-le, ZHANG Xiong-wei. An abstracting method and its implementation for feature-points in handwritten Chinese[J]. Acta Electronic Sinica, 1997, 25(5): 57 – 60. (in Chinese)

[20] 耿强,马珏.手写体汉字识别笔画提取方法的研究[J].江苏广播电视大学学报,2006,1(17):41 – 43,81.
GEN Qiang, MA Jue. Stroke Extracting method for handwritten Chinese character recognition[J]. Journal of Jiangsu Radio & Television University, 2006, 1(17): 41 – 43, 81. (in Chinese)

[21] 刘伟,朱宁波,李德鑫,等.基于模糊子笔画统计特征的手写体汉字识别[J].计算机工程与应用,2007,43(1):239 – 241,244.

[22] 杨玲,毛以芳,吴天爱.基于弹性网格和方向线素特征的脱机手写汉字识别[J].辽宁省交通高等专科学校学报.2008,10(1):38 – 39.
Yang Ling, Mao Yifang, Wu Tianai. Off-line handwrtitten Chinese character recognition research based on elastic meshes and directional line element feature[J]. Journal of Liaoning Provincial College of Communications, 2008, 10(1): 38 – 39. (in Chinese)

[23] Ruini Cao, Chew Lim Tan. A model of stroke extraction from Chinese character images[A]. Proceedings of 15th International Conference on Pattern Recognition[C]. Barcelona, Spain, IEEE Computer Society, 2000.4368 – 4371.

[24] Cheng-Lin Liu, In-Jung Kim, Jin H. Kim. Model-based stroke extraction and matching for handwritten Chinese character recognition[J]. Pattern Recognition, 2001, 34(12): 2339 – 2352.

[25] 王建平,蔺菲,陈军.基于手写体汉字笔画提取重构的识别方法[J].计算机工程,2007,33(10):230 – 232,248.
Wang Jianping, Lin Fei, Chen Jun. Recognition method based on handwritten Chinese characters stroke extraction recombined[J]. Computer Engineering, 2007, 33(10): 230 – 232, 248. (in Chinese)

[26] Jia Zeng, Zhi-Qiang Liu. Markov random fields for handwritten Chinese character recognition[A]. Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)[C]. Seoul, Republic of Korea, IEEE Computer Society, 2005.101 – 105.

[27] 吴佑寿.教电脑识字:浅谈汉字识别[M].北京:清华大学出版社,广州:暨南大学出版社,2000,12.

[28] Shi D M, Damper R I, Guan S R. Offline handwritten Chinese character recognition by radical decomposition[J]. Association for Computing Machinery Transactions on Asian Language Information Processsing (TALIP), 2003: 2(1): 27 – 48.

[29] 何浩智,朱宁波,刘伟.基于霍夫变换和弹性网格的手写汉字识别方法[J].计算机仿真,2008,25(1):240 – 243.
He Hao-zhi, Zhu Ning-bo, Liu Wei. Handwrtten Chinese character recognition based on Hough transformation and elastic mesh[J]. Computer Simulation, 2008, 25(1): 240 – 243. (in Chinese)

[30] 金连文,徐秉铮.手写体汉字识别中的一种新的特征提取方法[J].电路与系统学报,1997,2(3):7 – 12.
Jin Lian-wen, Xu Bin-zheng. Directional cellular feature extraction with elastic meshing for handwritten Chinese character recognition[J]. Journal of Circuits and Systems, 1997, 2(3): 7 – 12. (in Chinese)

[31] 金连文,覃剑钊.手写汉字识别弹性网格 Gabor 特征提取方法的研究[J].计算机应用研究,2004,21(12):163 – 165.
Jin Lian-wen, Qin Jian-zhao. Study on Gabor filter-based handwritten Chinese character feature extraction[J]. Application Research of Computers, 2004, 21(12): 163 – 165. (in Chinese)

[32] 陈光,张洪刚,郭军.一种新的加权动态网格汉字特征抽取方法[J].中文信息学报,2007,21(2):89 – 93.
Chen Guang, Zhang Hong-gang, Guo Jun. Feature extraction for handwritten Chinese character by weighted dynamic mesh based on nonlinear normalization[J]. Journal of Chinese information Processing, 2007, 21(2): 89 – 93. (in Chinese)

[33] 张睿,丁晓青,方驰.脱机手写汉字识别的最优采样特征新方法[J].中国图象图形学报,2002,7(2):176 – 180.

Zhang Rui, Ding Xiao-qing, Fang Chi. New method of optimal sampling features for offline handwritten Chinese character recognition[J]. Journal of Image and Graphics, 2002, 7(2): 176 – 180. (in Chinese)

[34] Nei Kato, Masato Suzuki, Shin ichiro Omachi, et al. A handwritten character recognition system using direction element feature and asymmetric mahalanobis distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(3): 258 – 262.

[35] 马少平,夏莹,朱小燕.基于模糊方向线素特征的手写体汉字识别[J].清华大学学报(自然科学版),1997,37(3):42 – 45.
Ma Shaoping, Xia Ying, Zhu Xiaoyan. Handwritten Chinese characters recognizing based on fuzzy directional line element feature[J]. Journal of Tsinghua University (Sci &Tech), 1997, 37(3): 42 – 45. (in Chinese)

[36] 王学文,丁晓青,刘长松.基于 Gabor 变换的高鲁棒汉字识别新方法[J].电子学报,2002,30(9):1317 – 1322.
Wang Xue-wen, Ding Xiao-qing, Liu Chang-song. Gabor filters based feature extraction for robust Chinese character recognition[J]. Acta Electronica Sinica, 2002, 30(9): 1317 – 1322. (in Chinese)

[37] 陈蓉,邓洪波,金连文.一种基于局部 Gabor 滤波器组的手写体汉字识别方法[J].计算机应用,2007,27(5):1222 – 1224.
Cheng Rong, Deng Hong-bo, Jin Lian-wen. Handwritten Chinese character recognition based on local Gabor filter bank [J]. Computer Applications, 2007, 27(5): 1222 – 1224. (in Chinese)

[38] Kai Ding, Zhibin Liu, Lianwen Jin, et al. A comparative study of Gabor feature and gradient feature for handwritten Chinese character recognition[A]. Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR'07) [C]. Beijing, China, IEEE Inc., 2007. 1182 – 1186.

[39] 崔金魁,杨杨,颉斌.一种基于集成 BP 网络的手写汉字识别方法[J].微电子学与计算机,2006,23(8):121 – 124.
Cui Jin-kui, Yang Yang, Xie Bin. A handwritten Chinese character recognition method based on integrated BP network[J]. Microelectronics & Computer, 2006, 23(8): 121 – 124. (in Chinese)

[40] 徐赵辉,杨杨,颉斌.基于弹性网格和 Legendre 矩的手写体汉字识别方法[J].计算机工程与应用,2006,42(17):163 – 165.
Xu Zhaohui, Yang Yang, Xie Bin. Handwritten Chinese character recognition based on elastic mesh and Legendre moment [J]. Computer Engineering and Applications, 2006, 42(17): 163 – 165. (in Chinese)

[41] 李玉静,杨杨,颉斌.基于矩和 Gabor 变换的手写体汉字识别方法[J],信息技术,2003,27(12):44 – 46.
Li yu-jing, Yang yang, Xie Bin. Handwritten Chinese character recognition based on moment and Gabor transformation[J]. Information Technology, 2003, 27(12): 44 – 46. (in Chinese)

[42] 高彦宇,杨杨.基于正交特征的手写体汉字识别方法[J].仪器仪表学报,2003,24(4S):446 – 447.
Gao Yanyu, Yang Yang. Handwritten Chinese character recognition based on orthogonal feature[J]. Chinese Journal of Scientific Instrument, 2003, 24(4S): 446 – 447. (in Chinese)

[43] 王先梅,杨扬,颉斌,等.基于 Krawtchouk 矩与 HMM 的脱机手写汉字识别技术[A]. The Sixth World Congress on Intelligent Control and Automation, 2006(WCICA 2006)[C]. Dalian, China, IEEE press, 2006. 10068 – 10072.
Xianmei Wang, Yang Yang, Bin Xie, et al. HMM-based offline handwritten Chinese characters recognition using Krawtchouk moments[A]. The Sixth World Congress on Intelligent Control and Automation, 2006(WCICA 2006)[C]. Dalian, China, IEEE Press, 2006. 10068 – 10072. (in Chinese)

[44] Xianmei Wang, Bin Xie, Yang Yang. Combining krawtchouk moments and HMMs for offline handwritten Chinese character recognition[A]. 2006 3rd International IEEE Conference Intelligent Systems (IS'06)[C]. London, United kingdom, IEEE Inc., 2006. 661 – 665.

[45] 范晓峰,施泽生.基于小波矩的新型图形识别算法[J].计算机工程与应用,2001,37(7):47 – 52.
Fan Xiaofeng, Shi Zesheng. A new method of image recognition based on wavelet moment[J]. Computer Engineering and Applications, 2001, 37(7): 47 – 52. (in Chinese)

[46] 王先梅,黄康,林子钰.Krawtchouk 矩在脱机手写汉字识别中的应用[J].广西师范大学学报(自然科学版),2006,24(4):227 – 230.
Wang Xian-mei, Huang Kang, Lin Zi-yu. Using Krawtchouk moments for off-line handwritten Chinese character recognition[J]. Journal of Guangxi Normal University (Natural Science Edition), 2006, 24(4): 227 – 230. (in Chinese)

[47] CHO-HUAK THE, ROLAND T. CHIN. On image analysis by the methods of moments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, 10(4): 496 – 513.

[48] 姜璐,章品正,舒华忠.矩在面部表情识别中的应用[J].东南大学学报(自然科学版),2004,34(4):557 – 560.
Jiang Lu, Zhang Pinzheng, Shu Huazhong. Moment application to human facial expression recognition[J]. Journal of Southeast University (Natural Science Edition), 2004, 34(4): 557 – 560. (in Chinese)

[49] 杨蕊红,潘泉,程咏梅.小波不变矩在图像识别中的应用研究[J].计算机应用研究,2005,22(11):239 – 243.
Yang Rui-hong, Prn Quan, Cheng Yong-mei. Application of invariant wavelet moment to image recognition[J]. Application Research of Computers, 2005, 22(11): 239 – 243. (in

Chinese)

[50] 居琰,汪同庆,彭建,等.特征融合用于手写体汉字识别研究[J].电子科技大学学报,2007,31(3):229 – 233.
Ju Yan, Wang Tongqing, Peng Jian, et al. Research on handwritten Chinese character recognition using feature fusion and modular RBF classifier[J]. Journal of UEST of China, 2007, 31(3):229 – 233. (in Chinese)

[51] 刘海龙,丁晓青.基于镜像学习和复合二次距离的手写汉字识别[J].清华大学学报(自然科学版),2006,46(7):1239 – 1242.
Liu Hailong, Ding Xiaoqing. Handwritten Chinese character recognition based on mirror image learning and the compound Mahalanobis function[J]. Journal of Tsinghua University (Sci & Tech), 2006, 46(7):1239 – 1242. (in Chinese)

[52] Cheng-Lin Liu, Hiromichi Fujisawa. Classification and learning methods for character recognition: advances and remaining problems[A]. Studies in computational intelligence: Machine Learning in Document Analysis and Recognition[C]. Springer Verlag, Berlin, Heidelberg, 2008. 139 – 161.

[53] 李美丽,杨杨,李岩.基于形态学变换的有限集手写体汉字识别[J].传感技术学报,2007,20(5):1184 – 1187.
Li Mei-li, Yang Yang, Li Yan. Small set handwritten Chinese character recognition based on mathematical morphology[J]. Chinese Journal of Sensors and Actuators, 2007, 20(5):1184 – 1187. (in Chinese)

[54] Yih-Ming Su, Jhing-Fa Wang. A novel stroke extraction method for Chinese characters using Gabor filters[J]. Pattern Recognition, 2003, 36(3):635 – 647.

[55] Xuewen Wang, Xiaoqing Ding, Changsong Liu. Gabor filters-based feature extraction for character recognition[J]. Pattern Recognition, 2005, 38(3):369 – 379.

[56] Weipeng Zhang, Yuan Yan Tang, Yun Xue. Handwritten character recognition using combined gradient and wavelet feature [A]. 2006 International Conference on Computational Intelligence and Security (ICCIAS)[C]. Guangzhou, China, IEEE Computer Society, 2006. 662 – 667.

[57] 付强,丁晓青,刘长松.用于手写汉字识别的级联 MQDF 分类器[J].清华大学学报(自然科学版),2008,48(10):1605 – 1608.
Fu Qiang, Ding Xiaoqing, Liu Changsong. Cascade MQDF classifier for handwritten character recognition[J]. Journal of Tsinghua University (Sci & Tech), 2008, 48(10): 1605 – 1608. (in Chinese)

[58] Hailong Liu, Xiaoqing Ding. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes[A]. 8th International conference on Document Analysis and Recognition[C]. Seoul, Republic of Korea, IEEE Computer Society, 2005. 19 – 25.

[59] 高学,金连文,尹俊勋.一种基于支持向量机的手写汉字识别方法[J].电子学报,2002,30(5):651 – 654.
Gao Xue, Jin Lian-wen, Yin Jun-xun. A new SVM-based handwritten Chinese character recognition method[J]. Acta Electronic Sinica, 2002, 30(5):651 – 654. (in Chinese)

[60] 王建平,张丽萍.脱机手写体汉字识别的支持向量机方法研究[J].计算机与数字工程,2008,36(4):146 – 150.
Wang Jianping, Zhang Liping. Research on method of off-line handwritten Chinese characters recognizing based on SVM [J]. Computer & Digital Engineering, 2008, 36(4): 146 – 150. (in Chinese)

[61] Fu Chang. Techniques for solving the large-scale classification problem in Chinese Handwriting Recognition[A]. Lecture Notes in Computer Science: Arabic and Chinese Handwriting Recognition[C]. College Park, MD, United states, Spinger Verlag, Heidelberg, Germany, 2008. 161 – 169.

[62] 高彦宇,杨杨,陈飞.基于融合特征和 LS-SVM 的脱机手写体汉字识别[J].北京科技大学学报,2005,27(4):509 – 512.
Gao Yanyu, Yang Yang, Chen Fei. Off-line handwritten Chinese character recognition based on fusion features and LS-SVM[J]. Journal of University of Science and Technology Beijing, 2005, 27(4):509 – 512. (in Chinese)

[63] 张芳,汪成军.基于支持向量机的手写体汉字的识别[J].计算机与数学工程,2006,34(1):65 – 68.
Zhang Fang, Wang Chenjun. Handwritten Chinese characters recognition based on support vector machine[J]. Computer & Digital Engineering, 2006, 34(1):65 – 68. (in Chinese)

[64] 官理,祖峰,唐文胜.快速的支持向量机多类分类研究[J].计算机工程与应用,2008,44(5):177 – 179.
GUAN Li, ZU Feng, TANG Wen-sheng. Research of fast multicalss SVM classification [J]. Computer Engineering and Applications, 2008, 44(5):177 – 179. (in Chinese)

[65] 刘冰.多类 SVM 分类算法的研究和改进[J].电脑知识与技术,2007(6):1590 – 1593.
Liu Bing. Research and improvement of classification methods for multi-class support vector machines[J]. Computer Knowledge and Technology, 2007(6):1590 – 1593. (in Chinese)

[66] Kok Seng Chua. Efficient computations for large least square support vector machine classifiers[J]. Pattern Recognition Letters, 2003, 24(1 – 3):75 – 80.

[67] 安建慧,宋柏.模拟退火算法在汉字图像识别中的应用与研究[J].计算机应用,2007,27(12):89 – 90.

[68] 黄戈祥,陈继荣.ART2 神经网络在手写体汉字识别中的应用[J].计算机仿真,2006,23(7):153 – 156.
Huang Ge-xiang, Chen Ji-yong. Application of ART2 neural netwrok to handwritten Chinese character recognition[J]. Computer Simulation, 2006, 23(7):153 – 156. (in Chinese)

[69] Da Lu, Qiwei Chen, Wei Pu, et al. Study on preclassification for handwritten Chinese character based on neural net and

fuzzy matching algorithm[A]. 2007 IEEE International Conference on Robotics and Biomimetics(ROBIO)[C]. Yalong Bay, Sanya, China, IEEE Computer Society, 2007. 1344 – 1349.

[70] Xue Gao. A training strategy of class-modular neural network classifier for handwritten Chinese character recognition[A]. Intelligent Computing in Signal Processing and Pattern Recognition(ICIC2006)[C]. Kunming, China, Springer, Berlin, Heidelberg 2006. 657 – 662.

[71] 赵巍,刘家锋,唐降龙.基于部件 HMM 级联的联机手写体汉字识别方法[J].哈尔滨工业大学学报,2004,36(5):570 – 573.
Zhao Wei, Liu Jia-feng, Tang Xiang-long. An on-line free handwritten Chinese character recognition method based on component cascaded HMMs[J]. Journal of Habin Institute of Technology, 2004, 36(5): 570 – 573. (in Chinese)

[72] 刘健,李会方,牛新伟.基于 MHMM 模型的手写体汉字识别算法[J].信息安全与通信保密,2007,(2):75 – 77.
Liu Jian, Li Huifang, Niu Xinwei. The algorithm of handwritten Chinese character recognition based on multiple HMM[J]. Information Security and Communications Privacy, 2007, (2): 75 – 77. (in Chinese)

[73] 高彦宇,杨杨.脱机手写体汉字识别研究综述[J].计算机工程与应用,2004,40(7):74 – 77.
Gao Yanyu, Yang Yang. A survey of off-line handwritten Chinese vharacter recognition[J]. Computer Engineering and Applications, 2004, 40(7): 74 – 77. (in Chinese)

[74] http://www.is.aist.go.jp/etlcdb/[OL], 2008 – 11 – 14.

[75] 郭军,蔺志青,张洪刚.一个新的脱机手写汉字数据库模型及其应用[J].电子学报,2000,28(5):115 – 116.
Guo Jun, Lin Zhi-qing, Zhang Hong-gang. A new database model of off-line handwritten Chinese character and its application[J]. Acta Eletronica Sinica, 2000, 28(5): 115 – 116. (in Chinese)

[76] HCL2000 手写汉字数据库系统[OL]. http://www.pris.net.cn/down2/Software.asp? id = 3. 2008 – 8 – 16.

[77] 李元祥,丁晓青,吴佑寿.一种基于字词结合的汉字识别上下文处理新方法[J].计算机研究与发展,2002,39(7),838 – 842.
Li Yuan-Xiang, Ding Xiao-Qing, Wu You-Shou. A novel method based on integrating characters with words for contextual processing of Chinese character recognition[J]. Journal of Computer Research and Development, 2002, 39(7), 838 – 842. (in Chinese)

[78] 董广宇,吕学强,王涛,等.基于 N-gram 语言模型的汉字识别后处理研究[J].微计算机信息(测控自动化),2009,25(4 – 1):276 – 278.
Dong Guang-yu, Lv Xue-qiang, Wang Tao, et al. Post-processing study of Chinese character recognition based on n-gram language model[J]. Control&Automation, 2009, 25(4 – 1): 276 – 278. (in Chinese)

[79] 龙 ▮,庄丽,朱小燕,等.手写中文地址识别后处理方法的研究[J].中文信息学报,2006,20(6):69 – 74.
Long Chong, Zhuang Li, Zhu Xiao-yan. A post-processing approach for handwritten Chinese address recognition[J]. Journal of Chinese Information Processing, 2006, 20(6): 69 – 74. (in Chinese)

[80] 袁毓林.基于统计的语言处理模型的局限性[J].语言文字应用,2004,17(2):99 – 108.
Yuan Yulin. The limitations of the statistically-based NLP models[J]. Applied Linguistics, 2004, 17(2): 99 – 108. (in Chinese)

[81] N-gram 模型[OL]. http://hi.baidu.com/bytechen/blog/item/94cf53def1d4ce5fcdbf1a40.html, 2009 – 5 – 29.

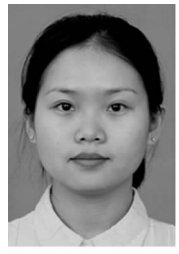
[82] 自然语言处理中理性主义与经验主义的优缺点[OL]. http://www.xiaolai.net/rshare/feed.php? channel = 126&y = 2009&d = 21&iid = 14537, 2009 – 5 – 31.

[83] 计算语言学和自然语言信息处理研究和应用综述[OL]. http://ling.cass.cn/yingyong/courses/nlpbase.htm, 2009 – 5 – 31.

[84] 李元祥,刘长松,丁晓青.一种利用校对信息的汉字识别自适应后处理方法[J].中文信息学报,2001,15(1):46 – 52.
Li Yuan-xiang, Liu Chang-song, DING Xiao-qing. An adaptive post-processing method using proofreading information for Chinese character recognition[J]. Journal of Chinese Information Processing, 2001, 15(1): 46 – 52. (in Chinese)

作者简介:

赵继印　男,1961 年出生于吉林省九台市.1993 年 9 月获得吉林工业大学通信与电子系统专业博士学位.现任吉林大学通信工程学院教授、博士生导师.主要从事智能信息处理与传输方面的教学和研究工作.
E-mail:zhaojiyin2000@163.com



郑蕊蕊　女,1982 年出生于河南省开封市.吉林大学通信工程学院博士学位研究生.主要从事图像处理与模式识别方面的研究工作.本文通信作者.
E-mail:zhengruirui@yahoo.cn