脱机印刷体彝族文字识别系统的原理与实现

朱宗晓[1],吴显礼[1,2]
(1. 中南民族大学 计算机科学学院 少数民族语言文字信息处理实验室,湖北 武汉 430074;
2. 中科院自动化所,北京 100086)

摘 要:脱机印刷体彝文文字识别系统包括字符分割、特征提取、特征压缩以及字典匹配四个主要模块,该系统利用总结出的彝文字符合并和反合并规则提高了字符分割准确率,采用1024维周边方向贡献度作为彝文字符统计特征,对彝文中存在的大量相似字符具有良好的区分能力。系统还采用基于KL变换的特征压缩算法和三级字典快速匹配算法,最终实现了一个基于Windows平台的脱机印刷体彝文识别平台,该平台对样本的一次识别率在99.4%以上。实验结果表明这些方法是可行的和高效的。

关键词:彝文识别;字符分割;周边方向贡献度;特征压缩;字典匹配
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2012)02-0085-04

Principles and Implementation of an Off-Line Printed Yi Character Recognition System

ZHU Zong-xiao[1], WU Xian-li[1,2]
(1. Information Processing Lab for Minority Language, College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China;
2. Institute of Automation, Chinese Academy of Science, Beijing 100086, China)

Abstract: The off-line printed Yi character recognition system consists of four main modules, including character segmentation, feature extraction, feature compressing and dictionary matching. In this system, Yi characters' merger rules and anti-merger rules are summarized to increase the accuracy of character segmentation; 1024 dimensional features of peripheral direction contribution are used as statiscal characteristics of Yi character with good discrimination between a large numbers of Yi similar characters. Feature compressing algorithm based on Karhunen-Loève transformation and three-level dictionary -matching algorithm is also adopted to realize an off-line printed Yi character recognition system based on Windows. Its first recognition rates of samples are above 99.4%, which proved that all arithmetic used in this system are feasible and efficient.

Key words: Yi character recognition; character segmentation; peripheral direction contribution; feature compressing; dictionary-matching

0 引 言

随着数字计算机深入人们的日常生活,人类希望机器也能模仿人类阅读书籍的能力,于是光学字符识别技术(Optical Character Recognition: OCR)应运而生[1]。OCR技术自20世纪80年代进入汉字识别领域已有多年,成果卓著,并涌现出如汉王科技、清华文通这样在国内拥有自主知识产权,占据绝大部分市场份额的高科技企业。在少数民族文字识别方面,如清华大学与西北民族大学合作研究的藏文识别[2]、新疆大学的维吾尔文[3]、哈萨克文字符识别[4]、内蒙古大学的蒙文识别研究[5]等都取得了较大进展。而另一种古老而重要的文字——彝文的OCR研究尚属空白。1980年,四川凉山彝族聚居区开始推行规范彝文。随后,规范彝文在文化、教育、政令、娱乐等众多领域得到广泛的推广和使用,多种书籍报刊如《凉山日报》、小学彝文课本等采用规范彝文印刷,对彝族人民生产、生活产生了重大影响。彝文OCR研究的空白制约了信息技术在该少数民族地区的普及和应用。为此,研制了一个脱机印刷体彝族文字识别系统,该系统包括字符分割、特征提取、特征压缩以及字典匹配四个主要模块,为规范彝文的数字化提供了一种快捷便利的方式,对彝族地区的经济发展和文化传承有着积极的作用。

收稿日期:2011-07-13;**修回日期**:2011-10-20
基金项目:国家自然科学基金面上项目(60975021);中南民族大学中央高校基本科研业务费专项资金项目(ZZY10007);中南民族大学自然科学基金项目(YZY07001)
作者简介:朱宗晓(1978-),男,湖北武汉人,讲师,博士研究生,CCF会员,主要研究方向为模式识别、图像处理、少数民族文字信息处理;吴显礼,教授,主要研究方向为模式识别、图像处理、少数民族文字信息处理。

1 研究流程和研究重点

根据对已有的各种识别系统进行分析研究，脱机印刷体彝族文字识别系统的研究流程如图1所示：

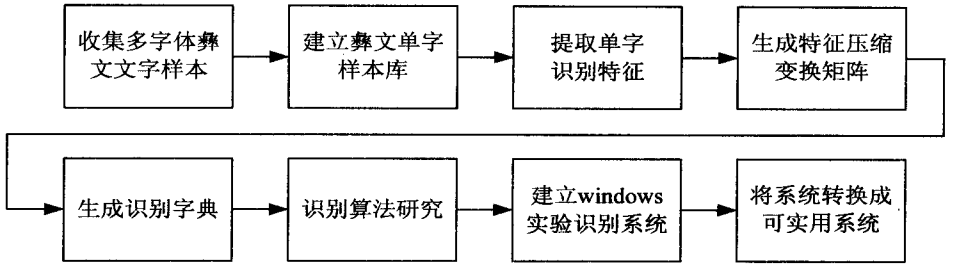

图1 脱机印刷体彝族文字识别系统研究流程

在这个研究流程中，需要研究的重点算法包括：

1）彝文字符分割算法。

待处理图像中通常包含多个彝文文字，通过字符分割可以将每个彝文字符单独的提取出来。字符分割是彝文文字识别中的一个重要环节，切分是否准确对文字特征提取和分类识别将产生巨大影响，从而直接影响最后的识别结果。

2）存在大量相似字的特征提取算法。

在总共1165个规范彝文中，有345个规范彝字含有次高调字[6]，规范彝字与其对应的次高调字通常只有一笔的差别，这导致印刷体彝文中存在许多相似字，这样对描述字符的特征提出了较高的要求——它必须使得两个相似字的类内差距尽可能小，而类间差距尽可能大。这一要求将被用来作为特征选择的标准。

3）利用压缩变换矩阵，生成识别字典。

利用K-L变换对其高维进行压缩，降低特征向量维数。

4）利用识别字典和压缩变换矩阵，建立基于单一特征的粗分类、细分类和识别的三级匹配识别算法。

2 主要算法

2.1 字符分割

如图2所示，字符分割可分为字初分和字合并两个步骤。字初分首对彝文文字进行行切分，即通过逐行搜索求出彝文文字的行上界和行下界。记录行上界和行下界的差值，即彝文文字的高度。统计彝文文字的行上界和行下界的个数。接着对彝文文字进行字切分，将每一行上界和行下界的区域组成一个块，并记录该块的高度。将彝文文字样本分成多个固定块。在每一个块中，求出彝文文字向X轴方向投影，从而确定列上界和列下界。记录列上界和列下界的差值，即彝文文字的宽度。最后根据行上界和列上界，彝文文字的高度和宽度确定彝文文字的切分区域。

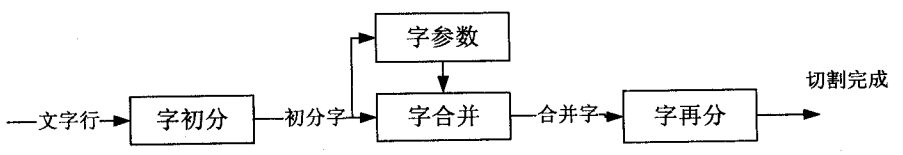

图2 字符分割框图

受到彝文字符结构的影响，有部分单个字符的组成部件被分割成了字符。这时通过观察分析，可以总结出一些字合并的规则，但单纯的使用字合并规则，会将一些标点符号错误地合并进来，此时要再总结出一些将错误合并的字符再分解开来的反合并规则。字合并规则和反合并规则需要利用字和字间的参数来描述。主要的参数包括当前初分字所在行的宽和高、当前初分字所在行中心线的高度、当前初分字及前后初分字的宽、高、间距等等。利用这些参数，可以比较准确地描述各种情况下字符合并与反合并规则，包括两字合并、三字合并、数字反合并、右括号与后面标点符号反合并等规则。这些规则来源于大量彝文字符和标点符号分割实践，对一般彝文印刷体的书报杂志上的文本分割具有较好效果，是本系统能达到较高识别率的重要保障。

2.2 特征提取

在彝文识别中的一个特点是存在大量相似字，特征的选取应使得两类相似字的类内差距尽可能小，而类间差距尽可能大。近几十年来专家学者提出很多不同的特征提取算法，大致可以分为两大类：基于统计的特征和基于结构的特征。统计特征主要是对文字点阵像素进行变换，找出其中能反应整个文字特征的点，即找出该文字的特征点。结构特征更加注重于用笔画或文字部件空间上的构成关系来反应文字的信息。常用的特征提取算法有粗网格特征（rough grid feature）[7]、整体特征Z（overall features of Z）[8]、弹性网格特征（elastic mesh feature）[9,10]、外围特征（peripheral feature）、全局笔画密度特征（G-DCD）、局部笔画密度特征（L-DCD）[11]、以及周边方向贡献度特征（peripheral direction contribution）[12]。通过在上述特征间反复比较相似字之间类内差距和类间差距的比值大小，系统最终采用基于周边方向贡献度1024维特征作为描述彝文字符的统计特征。

2.2.1 基于八方向笔画方向量的周边方向贡献度

如图3所示，假定P为彝文字符上的一个黑像素，分别求该点沿八个量化方向到黑像素边缘的距离，即统计八个方向中每一个方

向由 P 点起始连续黑像素点的个数。由公式(1)求得八维向量 $[d_1,d_2,d_3,\cdots,d_8]$。

$$d_i=\frac{l_i}{\sqrt{\sum_{i=1}^{8}l_i^{\,2}}}\quad(i=1,2,\cdots,8)\qquad(1)$$

其中,i 表示向量方向,l_i 表示第 i 个方向连续黑像素的个数。

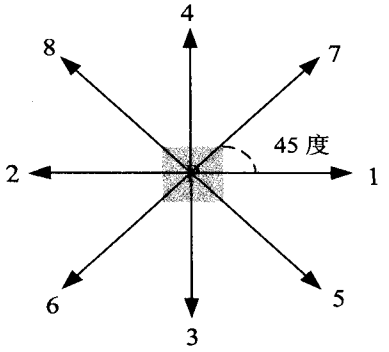

图 3　方向贡献度示意图

将其中互为相反方向的方向量 1 和 5,2 和 6,3 和 7,4 和 8 两两进行合并,实现方向贡献度的压缩,降低彝文字符特征的维数,公式如下:

$$d_1=\frac{l_1+l_2}{\sqrt{(l_1+l_2)^2+(l_3+l_4)^2+(l_6+l_7)^2+(l_5+l_8)^2}}\qquad(2)$$

$$d_2=\frac{l_3+l_4}{\sqrt{(l_1+l_2)^2+(l_3+l_4)^2+(l_6+l_7)^2+(l_5+l_8)^2}}\qquad(3)$$

$$d_3=\frac{l_6+l_7}{\sqrt{(l_1+l_2)^2+(l_3+l_4)^2+(l_6+l_7)^2+(l_5+l_8)^2}}\qquad(4)$$

$$d_4=\frac{l_5+l_8}{\sqrt{(l_1+l_2)^2+(l_3+l_4)^2+(l_6+l_7)^2+(l_5+l_8)^2}}\qquad(5)$$

按照图 3 所示的八个方向搜索整个彝文字符,并在每个搜索方向上求取沿该方向的四层边界点。如图 4 所示,以水平方向为例,对彝文字符从左到右搜索,然后根据该搜索线与彝文字符交点的个数,把第一个交点叫做第一层边界点,第二个交点叫做第二层边界点,依次求出第三层和第四层边界点,每一层边界点即为所求的特征点。依次求出每个特征点的八方向笔画方向量,应用公式(2)到(5)求出每个特征点压缩的四种笔画方向量,将所有的笔画方向量进行组合形成彝文字符的周边方向贡献度特征。

2.2.2　周边方向贡献度特征提取算法的实现

根据周边方法贡献度算法设计思想,该特征提取算法从八个方向四层深度搜索彝文字符的边界点,这样能较好的搜索到全包含结构彝文字符的内部结构信息,算法的具体实现步骤如下:

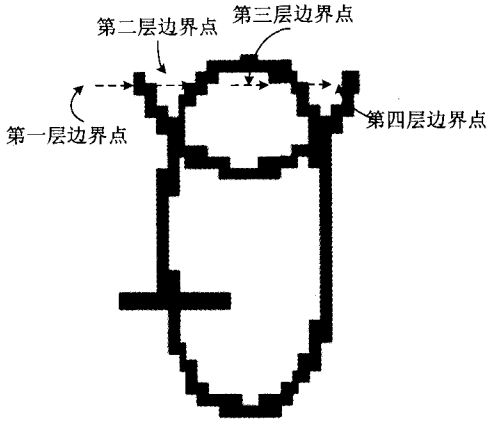

图 4　周边向贡献度示意图

步骤一:按照图 3 所示的八个方向平行将 64×64 的二值点阵分成八个区域。

步骤二:依次按照八个方向逐行平行搜索彝文字符的四层边界点。

步骤三:按照公式(1)依次求出每个边界点的八种笔画方向量,将同一区域内的同一层边界点的八种笔画方向量相加。应用公式(2)~(5)求出每个边界点的四种压缩的笔画方向量。

步骤四:将全部边界点的笔画方向量进行组合,则可以得到 8 区域×8 搜索方向×4 层深度×4 维 = 1024 维周边方向贡献度特征。

周边方向贡献度从多方向多层深度搜索彝文字符的边界点。该算法能较为完整的求出彝文字符有效的特征点,适合应用于彝文字符的特征提取上。

2.3　特征压缩和特征字典的生成

在分类识别之前要对高维的彝文字符特征进行压缩,特征维数过大会影响系统识别速度。本系统特征提取后得到 1024 维周边方向贡献度特征,对其采用 K-L 变换进行特征压缩,使特征向量维数分降低到 128 维。特征压缩后用压缩的 128 维特征建立特征字典,特征字典中保存每个彝文字符样本标准模板类。分类识别就是将待识别彝文字符的特征向量与特征字典中的每个标准模板类特征向量进行匹配。

2.4　基于单一特征的三级距离字典匹配算法

文中针对彝文存在大量相似字符的特点设计了一种三级距离分类器,该分类器在第一级分类中,通过求样本字符和标准模板类中字符的距离,将求得的距离从小到大排序,并选择前 N 个距离最小的候选字符进入第二级距离分类器。第二级距离分类的方式与第一级分类方式一样,直到进入第三级分类获得识别结果。其中一级距离分类器采用 Manhattan 距离[13],二级距离和三级距离分类器采用误差均衡距离[14]。Manhattan 距离算法简单,判别速度快,并且累积分类率较高,

所以用 Manhattan 距离分类器作为粗分类是合理的。而在存在大量相似字符的彝文中存在一些不稳定的部分,直接采用欧式距离分类器,会因为这些不稳定的特征使分类出现错误,造成最终识别结果不理想。因此文中对于彝文字符采用误差均衡距离作为细分类,它通过引入均衡系数,使得每一维特征向量造成的平均误差均等,使得彝文字符特征中稳定的部分得到突出,不稳定的部分被抑制。在时间的花费上误差均衡距离和欧式距离相差不大。但误差均衡距离在识别性能上有明显提高。

3 实验结果

采用基于周边方向贡献度的特征提取算法再经过 K-L 变化后得到 128 维的压缩特征,此特征与特征字典中的标准特征进行匹配,计算其三级距离,取加权距离最小的十个匹配结果作为识别字符候选结果。在"文字识别实验平台"中使用一键生成字典功能,即对分割出来的样本文件全体,一键完成其特征提取、特征压缩、特征字典生成、样本识别等功能。得到参与生成字典的十种彝文字体共计 720 个样本文件,一次识别率为 99.21%,前十次识别率为 99.95%。其中特征提取耗时 1009 秒,平均每秒钟提取 832 个字符特征,此速度可以保证在"族语通"(彝文版)文字识别软件中,每秒识别的总字符数(含标点符号及英文数字)稳定在 400 以上,足以满足一般应用的要求。但必须指出,这一环节在图 5 一键生成字典的各个环节中,耗时时间最长,是影响最终识别速度的主要因素。

4 结束语

综合运用前文介绍的各种方法,文中最终建立了在图 1 研究流程中提到的 Windows 实验识别系统——文字识别实验平台,该平台包含字符分割、特征提取、特征压缩以及字典匹配四个模块,能快速完成特征比较、字典建立、字典验证等功能。提高了文字识别工作中建立、验证识别字典的工作效率。在文字识别实验平台的基础上,文中将其成果转化为可投入实际应用的文字识别软件——"族语通"。在接下来的研究工作中,我们将进一步挖掘不同特征用于文字分类的显著性度量标准,并考虑增加一定的结构特征辅助,从而进一步提高对手写文本识别的准确率。

Pnt文件路径为: E:\PNTDATA\YIPNT\常用字体　　字典文件存放路径为: E:\PNTDATA\YIPNT\常用字体

720 个PNT文件参与生成字典 一次总识别率: 99.215186 %　前 10次总识别率: 99.946352 %　平均每文件总耗时 3.812500秒

Pnt文件校验耗时	3	秒	平均每文件耗时0.004167	秒	
Pnt特征提取耗时	1009	秒	平均每文件耗时1.401389	秒	
生成KL转换矩阵耗时	351	秒	平均每文件耗时0.487500	秒	
进行KL变换耗时	831	秒	平均每文件耗时1.154167	秒	
生成加权字典耗时	20	秒	平均每文件耗时0.027778	秒	
进行PKL文件识别耗时	526	秒	平均每文件耗时0.730556	秒	

参与生成字典字符总数: 838800 一次识别出错字符总数: 6583　识别失败字符总数: 450　总耗时2745 秒

图 5 文字识别实验平台中一键生成字典统计报表

参考文献:

[1] Ayadevan R. Offline Recognition of Devanagari Script: A Survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2011, 41(6): 1-15.

[2] 普次仁. 多种印刷字体藏文字符的特征提取方法研究[J]. 西藏大学学报(自然科学版), 2008, 23(1): 25-28.

[3] 伊力哈木·亚尔买买提, 哈力旦·A. 基于 Matlab GUI 的维吾尔文字符识别系统的设计[J]. 计算机技术与发展, 2010, 20(11): 92-94.

[4] 达吾勒·阿布都哈依尔, 古丽拉·阿东别克. 基于 ANN 的哈萨克文手写文字识别系统的研究[J]. 计算机工程与应用, 2008, 44(1): 225-228.

[5] 魏宏喜, 高光来. 一种基于连通域的蒙古文文档图像版面分析方法[J]. 内蒙古大学学报(自然科学版), 2007(5): 586-590.

[6] 《彝语大词典》编撰委员会. 彝语大词典[M]. 成都: 四川民族出版社, 1997.

[7] 王玉雷, 李永忠, 王汝山. 粗网格在印刷体藏文特征提取中的应用[J]. 科学技术与工程, 2009(18): 5546-5548.

[8] 欧阳应华. 一种基于特征提取的脱机手写汉字识别技术[D]. 兰州: 兰州大学, 2007.

[9] 金连文, 高　学. 几种手写体汉字网格方向特征提取法的比较研究[J]. 计算机应用研究, 2004(11): 38-40.

[10] Jin Lianwen, Yin Junxun, Xue Gao, et al. Study of Several Directional Feature Extraction Methods with Local Elastic Meshing Technology for HCCR[C]//Proceedings of the Sixth Intel. Conference for Young Computer Scientist. Hanzhou, China: [s. n.], 2001: 232-236.

[11] Suen C Y, Mori S, Kim S H, et al. Analysis and Recognition of Asian Scripts-the State of the Art[C]//Proceedings of the 7th International Conference on Document Analysis and Rec-

(下转第92页)

R_n 为框架内规则的推理步长。

综合评估函数的计算模型如下描述：

(1) 框架重要度 R_i 作为静态属性，其权重可由 W_1 表示。

(2) 框架置信度 R_c 作为静态属性可由 W_2 表示。

(3) 推理步长 R_n 作为动态拟值属性可由 W_3 表示。

根据上述所述模型，提出综合评估函数计算公式

$$Q_r = \frac{W_2}{R_n}\sum_{0}^{R_n}(r_{x1} + r_{x2} + \ldots + r_{xn}) + R_i * W_1 + R_n * W_3 \quad (3)$$

这种综合评价模型考虑到在推理过程中的动态变化[10]，所以不仅可以较大程度上保证推理过程的顺利，同时也提高了推理的准确度，进而保证了工程领域决策和调整的精度。

另外，引入综合评价模型也可以作为以后逆推理过程的基础，通过决策和调整的精确程度不断更新和完成规则库，进而不断提高决策的准确度[11]。

5 实例分析

研究油井措施调整方案设计知识结构，基于工程决策知识支撑环境，建立各类异常井分析和诊断系统。从业务划分，异常井领域内分类包括产油、含水饱和度、有效厚度、砂岩厚度、射孔情况、沉积相、水淹和电测解释结果等，初始概念达 10000 多个，并且不同的油层油区存在不同的分析标准[12]。油井措施调整需要分析生产状态和包含井史等方面的静态生产数据(A2 数据库)分析判断异常井和诊断异常原因，并且根据异常原因制定措施方案。通过总结经验丰富的工作人员和专家的开发知识，建立工程决策知识支撑环境，驱动各个辅助支持系统的工作。

针对某油田实际情况，建立了动态标准的综合评估专家系统，通过动态标准产生机获取了规则 R1 ~ R3 规则的参数，同时完成了将事实转化为框架规则的过程。

设油田某油层油井满足 R1 的对应框架 FN1，则可以通过对框架规则的推理获取相应规则并进行封装和深度推理。

推理过程如下：通过分析判断得到某井区产油井产油比<30%，获取重要度和框架名称 FN1，经过推理，满足产油>0.4，推理到 A1 族，通过沉积相为 2，推理到 A21 族，满足射孔 = 1 获取措施 M，通过解释表得到措施为对该井进行补孔和压裂。

在推理过程中推理步长为 3，通过对应规则-置信度表获取 01，08，15 号规则的置信度并带入综合评估函数计算获得该措施的综合评估结果。

6 结束语

通过实际应用，该模型较好地完成了对油田异常井的诊断和措施方案调整的任务。不仅为油田地质工程提供了辅助决策，同时也提供了辅助决策的准确度。同时实际中对专家知识库的不断完善和更新，辅助决策的准确度也在不断提高。但是该模型由于设计较为复杂，所以依旧存在性能等问题，将在未来的工作中不断改进和优化。

参考文献：

[1] Wang Nengbin. Database System Tutorial [M]. Beijing: Publishing House of Electronics Industry, 2004.

[2] Josephc G, Gary D R. Expert system principles and programming[M]. 4th ed. [s. l.]: Thomson, 2005.

[3] Meng Xiaofeng, Zhou Longxiang, Wang Shan. State of the art and trends in database research [J]. Journal of Software, 2004, 15(12): 1822-1836.

[4] 杨 兴，朱大奇，桑庆兵. 专家系统研究现状与展望[J]. 计算机应用研究，2007，24(5)：2-5.

[5] 蔡自兴，约翰·德尔金，龚 涛. 高级专家系统：原理、设计及应用[M]. 北京：科学出版社，2005.

[6] 李昌春，左为恒. 专家系统与专家控制系统[J]. 重庆工业管理学院学报，1996(4)：35-37.

[7] 陈振华，余永权，张 瑞. 模糊模式识别的几种基本模型研究[J]. 计算机技术与发展，2010，20(9)：32-34.

[8] 毛海军. 基于 Agent 的宏观经济智能预测决策支持系统研究[D]. 大连：大连理工大学，2003.

[9] 王 刚，王 浩. 基于粒度的知识粗糙性研究[J]. 计算机技术与发展，2008，18(1)：66-68.

[10] 戴 钎，王力生. 基于故障树和规则匹配的故障诊断专家系统[J]. 计算机应用，2005(9)：2034-2036.

[11] 杨 静，张楠男，李 建，等. 决策树算法的研究与应用[J]. 计算机技术与发展，2010，20(2)：114-115.

[12] 盖宗源，程国建，王 莹. 模糊专家系统在钻井风险预测中的应用[J]. 计算机技术与发展，2009，19(1)：224-226.

(上接第 88 页)

ognition. Edinburgh, Scotland: [s. n.], 2003: 866-878.

[12] Hagita N, Naito S, Masuda I. Handprinted Kanji characters recognition based on pattern matching method[C]//Proc ICTP83. [s. l.]: [s. n.], 1983: 169-174.

[13] Perlibakas V. Distance measures for PCA-based face recognition[J]. Pattern Recognition Letters, 2004, 25(6): 711-724.

[14] 金连文，梁羽杰. 一种新的距离分类方法及其应用[J]. 计算机工程，1999，25(8)：30-32.