DOI: 10.13382/j.jemi.2017.01.022

基于拓扑特征和投影法的东巴象形文识别方法研究*

徐小力 蒋章雷 吴国新 王红军 王 宁

(北京信息科技大学 现代测控技术教育部重点实验室 北京 100192)

摘 要:东巴象形文是由云南丽江纳西族先民创造并使用的,被誉为"世界上唯一活着的象形文字"。在图形识别、内容释读以及形、音、义信息等方面,现有的英文、汉字等识别系统及翻译系统往往不能适用于东巴象形文字,提出一种先拓扑特征处理后投影法特征提取的分步骤信息处理方法,并采用模板匹配法进行文字识别。通过实验验证表明,基于象形文固有特征的提取,利用拓扑特征与投影法相结合的特征提取方法进行东巴象形文字识别,具有准确度更高的特点,是东巴象形文识别的一种有效方法。

关键词: 文字识别;模板匹配法;拓扑特征;投影法

中图分类号: TP39 文献标识码: A **国家标准学科分类代码**: 510.40

Identification method of Dongba pictograph based on topological characteristic and projection method

Xu Xiaoli Jiang Zhanglei Wu Guoxin Wang Hongjun Wang Ning

(Key Laboratory of Modern Measurement & Control Technology, Ministry of Education, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: Dongba pictograph has been known as "the only living pictograph in the world". In the aspects of image recognition, content interpretation, the current English and Chinese character recognition system often can not be applied to Dongba pictograph. Concerning the difficulties in the identification of Dongba pictograph, a new character recognition is proposed. Topological features processing and projection methodcompose thefeature extraction method, then, the character recognition method based on template matching is adopted. It is showed that the feature extraction method based on the intrinsic characteristic of the pictograph, and the Dongba character recognition method based on template matching, has high accuracy through the experiment.

Keywords: character recognition; template matching; topological characteristic; projection method

1 引 言

纳西族是居住在中国西南部丽江古城的少数民族,该民族用最古老的东巴象形文字写下了数万卷经典,其东巴象形文字是当今公认的世界上唯一还在使用的象形文字[1]。东巴象形文字的文字形态比苏美尔和巴比伦的楔形文字、古埃及的圣书文字,以及中美洲的玛雅文字和中国甲骨文都更原始。2003年以该象形文字书写的纳西族东巴经典古籍被联合国教科文组织列为"世界记忆遗产"。

东巴象形文是用图画来表意的一种文字,从起初描绘在木石上简单的图像标记,发展到能书写成经书,经历了漫长的发展时期[2]。国内外诸多研究学者通过对东巴象形文写成的东巴经典进行研究陆续发表了一些代表性研究论著[2-10]。东巴象形文是纳西文化特征之一,象形文字都是以图表示物体、以图表示景色、以图表示事物,其中某些不能表达的部分用音调加以区分。东巴象形文

收稿日期:2016-12 Received Date: 2016-12

*基金项目:国家社会科学基金重大项目(12&ZD234)、北京市教委科研计划项目(KM201611232020)资助

的形、音、义都极其复杂,往往一字多形、多音、多义,也有异形同义的。东巴经典文化的传承自古以来都是以东巴家族内部口传心授的形式进行的,同时由于能够释读东巴古籍的老东巴祭司大都已年逾古稀,因此,急待需要采用现代信息化手段实现东巴文的识别[11-12]。东巴象形文属于图画文字过渡到象形文字时期的一种特殊文字形态,具有结构复杂、类别繁多、关联性强等特点,文字识别是一个研究的难点。

2　基于拓扑特征和投影法相结合的象形文特征提取方法

国内外学者针对特征提取进行了一些研究工作[13-14]。对东巴文字的统计研究发现,东巴象形文字用各种弧线勾勒出要表达的具体事物,强调“象形”,所以需根据东巴象形文的结构特点提取其特征。基于统计特征来对东巴象形文字进行特征提取是很有效、方便且利于识别的。常用的统计特征包括:拓扑特征、特征点法、投影法和 K-L 变换法等[15]。通过对以上方法进行研究,结合东巴象形文字自身结构特点,本文采用双层特征提取模式:将拓扑特征法和投影法相结合来提取东巴古籍中东巴文字的特征。

2.1　拓扑特征法

通过对东巴象形文字形状的统计,结合文字的固有特征,将东巴文字想象成在图论中的平面无向图来提取东巴象形文字的拓扑特征:孔、端点、叉点、块。

1)孔:图论中的内孔数。

2)块:图论中的连通区域,即象形文字中被隔离的块数。

3)端点:图论中度数等于 1 的顶点。当东巴象形文字中像素点 P 的四周 8 个像素点中有且仅有 1 个黑像素时,即 P 为端点。

4)叉点:图论中度数大于 2 的顶点,其中包括三叉点、四叉点等。当一个顶点度数过高时,将 n 叉点分离成的$(n-2)$个三叉点。

图 1 所示为以 P 为中心像素点的 3×3 邻域,根据此邻域定义变量 N 为:

$$N = P_1 + P_2 + P_3 + P_4 + P_5 + P_6 + P_7 + P_8 \quad (1)$$

P_4	P_3	P_2
P_5	P	P_1
P_6	P_7	P_8

图 1　3×3 邻域

Fig. 1　3×3 neighborhood

判定条件:1)若 $N=1$、$P=1$,则该像素点对应象形文字的端点;2)$N=3$ 或 4、$P=1$,则该像素点对应为三叉点或四叉点。

2.2　投影法

通过拓扑特征不足以对全部东巴象形文字进行特征提取,这种情况下就需要采用投影法对其进行特征提取,将待识别的象形文字在水平方向和垂直方向进行投影,计算东巴象形文字投影图像中有多少黑色素点,并将其转化为象形文字的特征向量,具体步骤如下。

1)对记载有东巴文字图像的范围用假设的水平方向和垂直方向的网格进行划分。

2)将东巴文字图像分别向 X 轴和 Y 轴进行投影,投影公式为:

$$u(x) = \sum_{y=0}^{m-1} f(x,y) \quad (2)$$

$$u(y) = \sum_{x=0}^{m-1} f(x,y) \quad 0 \leqslant x,y \leqslant m-1 \quad (3)$$

式中:是东巴文字图像的二维图形表达式。$u(x)$ 和 $u(y)$ 分别是东巴文字图像的二维图形 $f(x,y)$ 在 X 轴和 Y 轴的投影函数,m 为垂直方向和水平方向上直线总条数,$m-1$ 为分别在垂直方向和竖直方向上的直线形成网格的总个数。

3)通过步骤 1)中的网格提取东巴文字图像的二维图形的投影值作为文字的特征:将步骤 1)中的每个网格再细分成 $N_1 \times N_2$ 子区域,N_1 为水平方向的网格数,N_2 为竖直方向上的网格数,N_1 和 N_2 的具体数值根据具体图像的大小而定,经研究分析,根据东巴文字图像的大小,取 $N_1<10$、$N_2<10$,最能有效进行特征提取。则根据式(2)、(3)可以得到的投影值为:

$$u(x) = \sum_{y=1}^{N_2} f(x,y) \quad 1 \leqslant x \leqslant N_1 \quad (4)$$

$$u(y) = \sum_{x=1}^{N_1} f(x,y) \quad 1 \leqslant y \leqslant N_2 \quad (5)$$

式中:$f(x_i,y_j)$ 为网格中的子区域,为二维图形 $f(x,y)$ 取各个特定值时的函数;在每一个子区域中,若黑象素占子区域象素的一半或多余一半,则 $f(x_i,y_j)$ 值为 1,否则 $f(x_i,y_j)$ 的值为 0,将 $u(x_1)$, $u(x_2),\cdots,u(x_{N1})$ 和 $u(y_1),u(y_2),\cdots,u(y_{N2})$ 的特征值分别计算出来,最后组成特征向量的形式。

3　基于模板匹配法的东巴象形文识别方法

对东巴象形文字进行特征提取后,待识别的每个象形文字都具有相应的数字特征,根据数字特征来对东巴象形文字进行字符识别。当前对文字进行识别的方法大致有两种:模板匹配和神经网络法[16]。模板匹配算法简

单、快速,不涉及复杂的迭代评估问题,本文采用模板匹配法对东巴文进行智能识别。

模板匹配法将待识别的文字与模板中的文字进行特征对比,通过相似法进行匹配,将东巴象形文字分到相关性较大的一个类别[17]。待识别象形文字进行模板匹配时,首先要建立东巴象形文的文字模板库。

3.1 文字模板库的建立

将现有《纳西象形文字谱》中和《方国瑜字典》中的东巴文字进行数字化特征提取,并用 TrueType 技术构建东巴文字模,将标准字模逐个导入 Font Creator Program 中作为东巴象形文的文字模板库,导入后的字模如图 2 所示。

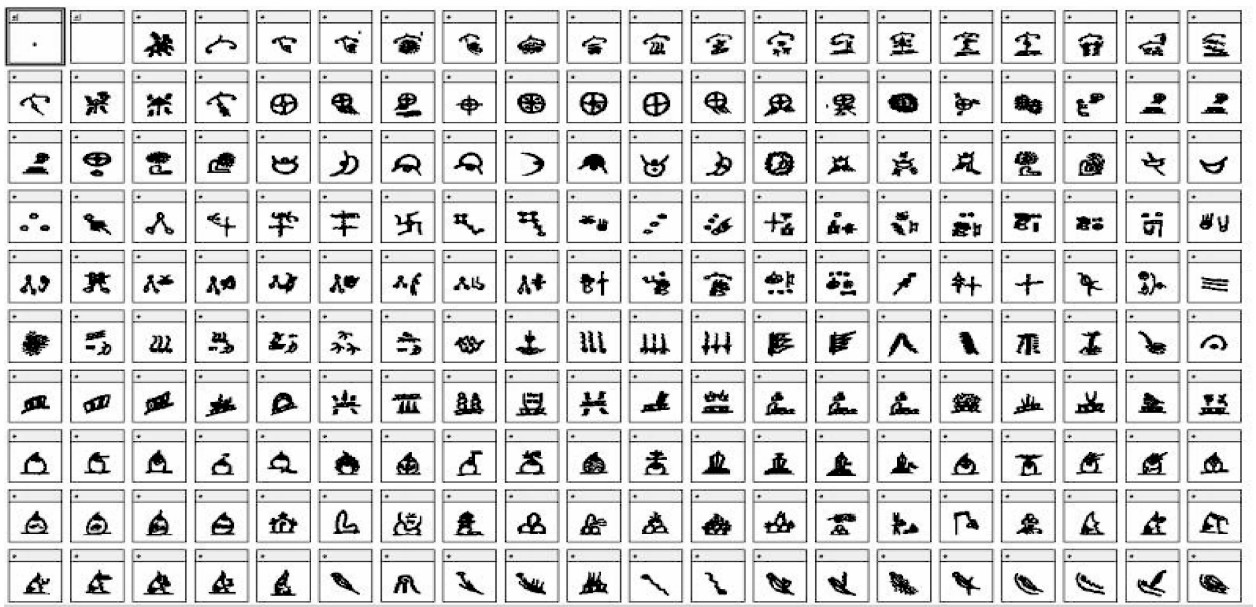


图 2 导入的东巴文字模

Fig. 2 The modes of Dongba characters imported

3.2 模板匹配

模板匹配的原理是选择图 2 中的东巴文字模作为模板,与东巴经典古籍中待识别的东巴象形文字进行特征比较,看与哪个模板匹配程度高。

1)粗分类

为了提高东巴象形文字的识别速度,首先根据上文所提取的拓扑结构特征采用最邻近法对东巴象形文进行粗分类,将待识别文字与文字模板库中具有相同拓扑特征结构的字模相匹配进行识别。

最邻近法:已知东巴象形文字样本集。

$$P_N = \{(x_1,\theta_1),(x_2,\theta_2),(x_N,\theta_N)\} \tag{6}$$

式中:x_i 是待识别东巴文字 i 的特征向量,θ_i 是它对应文字模板库中相似的的象形文字,设文字模板库中有 c 个类别,即 $\theta_i \in \{1,2,\cdots,c\}$。两个待识别样本间的距离为 $s(x_i,x_j) = \|x_i - x_j\|$。

对于待识别的东巴文样本 x,通过求出 P_N 中与之最小距离,即:

$$s(x,x') = \min_{j=1,\cdots,N} s(x,x_j) \tag{7}$$

式中:x'(对应的类别为 θ'),则待识别东巴文字 x 将和 θ' 类别匹配成功。

2)相似法

经过最邻近法进行匹配后的待识别东巴象形文字分类到一个较小的范围内,这样就需要一个精度更高算法来确定。根据上文的投影法提取的字符特征通过相似法度量待识别东巴文字样本与现有模板中东巴文字之间的相似性,得到两者的相关值,并采用遍历的搜索算法得到最大相关值,相关值为最大值时则判定模板东巴文字与待识东巴文字样本相似程度最高。其中,相关值为:

$$T = \frac{\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[M(i,j)-\bar{M}][N(i,j)-\bar{N}]}{\sqrt{\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[M(i,j)-\bar{M}]^2\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[N(i,j)-\bar{N}]^2}} \tag{8}$$

式中:$M(i,j)$ 为模板东巴文字图像的特征向量,$N(i,j)$ 为待识别东巴文字样本图像的特征向量,$\bar{M}$ 为模板东巴文字特征向量的平均值,$\bar{M} = \frac{1}{st}\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}M(i,j)$,$\bar{N}$ 为待识别东巴文字特征向量的平均值,$\bar{N} = \frac{1}{st}\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}N(i,j)$。

4　实验分析及结果

为了验证提出的基于拓扑特征法和投影法相结合的特征提取方法的有效性,利用东巴象形文字识别系统对该方法与拓扑特征特征提取方法进行分析。

4.1　东巴象形文字识别系统的建立

东巴象形文字识别主要由光电转换检测、图像预处理、东巴文字特征提取和东巴文字识别 4 个模块组成,如图 3 所示。

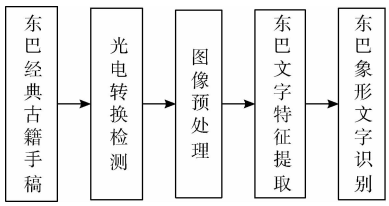


图 3　东巴象形文字识别系统

Fig. 3　The system of Dongba pictograph recognition

东巴经典古籍手稿:采集的东巴经书以及东巴古籍等。

光电转换检测主要功能:对经书上的文字进行光电转换后,在内部进行模数转换,将模拟信号转变成具有一定灰度的数字信号,便于后续图像的预处理和文字识别。本文将采集到的纸面上的东巴经文用 300 dpi 的数字化扫描仪进行扫描存储。

图像预处理:将采集到的东巴图像经过图像压缩(去除图片冗余度)、灰度化处理(将彩色图像转化为灰色图像)、平滑去噪(去除信号中的污点、空白等噪声)、归一化处理(文字的大小、位置和笔画粗细等进行规范化)等方法进行图像预处理。

东巴象形文字特征提取:按照东巴象形文字的特殊结构,选用拓扑特征和投影法分别对东巴象形文字进行特征提取,根据自身结构特点,先通过拓扑特征提取出东巴象形文字的端点、块数、孔数和叉点数,再结合投影法,提取东巴象形文字特征来描述文字。

东巴象形文字识别主要是利用模板匹配法进行识别,分两步:根据提取出的东巴文字的拓扑特征采用最近邻法对东巴象形文字进行模板匹配,将其进行粗分类缩小到一定范围内,然后通过更高精度的相似法来计算与东巴文字模板之间的相关性,来达到进一步的识别。

4.2　结果分析

拓扑特征法:如图 4(a)所示是一个代表“明天早晨”意思的东巴象形文字,图 4(b)所示说明了上述拓扑特征的含义。由图中的统计结果可以得出该东巴象形文中共有 5 个端点,22 个叉点(其中 15 个三叉点,4 个四叉点、3 个二叉点),3 个块,12 个孔。

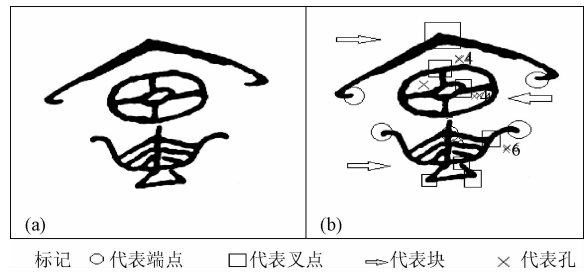


图 4　拓扑特征结构

Fig. 4　Schematic diagram of topological feature structure

对东巴象形文字进行拓扑特征提取后,仍有少部分东巴象形文字具有相同拓扑特征,对于这种文字进行投影法特征提取。采用投影法进行字符识别时抗干扰能力强、容易实现,由于东巴象形文字笔划简单,使得投影法对象形文字进行细分类识别达到很好的效果。

对以上识别方法进行基于模板匹配法的测试,选用待识别象形文字为 2 500 个,得到如表 1 所示实验数据。

表 1　实验数据对比

Table 1　Comparison of experimental data

特征提取方法	实验字数	识别字数	识别率/%
拓扑特征法	2 500	2 001	80
拓扑特征法和投影法相结合	2 500	2 110	84.4

由表 1 可知,采用拓扑特征法和投影法相结合的特诊提取方法对东巴象形文字进行识别时,能对大部分象形文字进行正确识别,具有更高的识别率。

5　结　论

先通过拓扑特征提取出东巴象形文字的端点、块数、孔数和叉点数,再结合投影法来对剩余的东巴文字进行特征提取,使所提取的特征可以准确的将该文字与其他文字区别开来。采用模板匹配法对东巴文进行智能识别,实验结果显示基于象形文固有特征的提取,利用拓扑特征法和投影法相结合的特征提取方法进行东巴象形文字识别,具有准确度更高的优点。

参考文献

[1]　西田龙雄. 活着的象形文字—纳西族的文化[M]. 东京:中公新书,1966.
TATSUO N. Living Pictographs—the Culture of the Naxi Nationality[M]. Japan: Chuko Shinsho, 1966.

[2]　杨福泉. 东巴教通论[M]. 北京:中华书局出版社,2012.

Yang Fuquan. A study of Dongba Religion[M]. Beijing: Chung Hwa Book Press,2012.

[3] 约瑟夫·洛克. 纳西文献研究[J]. 法兰西远东学报,1937,37(1).
ROCK J F. Naxi literature research[J]. Journal of the French Far East,1937,37(1).

[4] 李霖灿. 么些象形文字字典[M]. 台北:国立中央博物院,1944.
LI L C. Mexie Glyph Dictionary[M]. Taipei: National Central Museum,1944.

[5] 白庚胜. 纳西学丛书(共30卷)[M]. 北京:民族出版社,2008.
BAI G SH. Naxi Studies Series(30 volumes)[M]. Beijing: The Ethnic Publishing House,2008.

[6] 杨福泉. 杨福泉纳西学论集[M]. 北京:民族出版社,2009.
YANG F Q. Yang Fuquan's Essays on Naxi Learning[M]. Beijing: The Ethnic Publishing House,2009.

[7] 和志武. 纳西学论集[M]. 北京:民族出版社,2008.
HE ZH W. Naxi Studies Analects[M]. Beijing: The Ethnic Publishing House,2008.

[8] 和少英. 纳西族文化史[M]. 昆明:云南民族出版社,2001.
HE SH Y. The History of Naxi Culture[M]. Kunming: Yunnan National Press,2001.

[9] 方国瑜. 方国瑜纳西学论集[M]. 北京:民族出版社,2008.
FANG G Y. Fang Guoyu's Essays on Western Culture[M]. Beijing: The Ethnic Publishing House, 2008.

[10] 纳西东巴古籍译注全集编委会. 纳西东巴古籍译注全集(100卷)[M]. 昆明:云南人民出版社,1999.
The Collected Edition of NaxiDongba Ancient Books Annotation. The Collected Edition of NaxiDongba Ancient Books Annotation (100 volumes)[M]. Kunming: Yunnan People's Publishing House,1999.

[11] 王宁,徐小力,李志华,等. 东巴经典古籍释读数据库建立方法[J]. 北京信息科技大学学报:自然科学版,2015,30(5):40-43.
WANG N, XU X L, LI ZH H, et al. Interpretation database construction of Dongbamanuscripts[J]. Journal of Beijing Information Science & Technology University: Natural Science Edition,2015,30(5):40-43.

[12] 李志华,徐小力,王宁,等. 自适应中值滤波在东巴古籍图像去噪中的应用研究[J]. 北京信息科技大学学报:自然科学版,2015,30(5):36-39.
LI Zh H, XU X L, WANG N, et al. Application of adaptive median filtering to image denoising of Dongba manuscripts[J]. Journal of Beijing Information Science & Technology University: Natural Science Edition, 2015, 30(5):36-39.

[13] 李伟红,陈伟民,龚卫国. 一种人脸特征选择新方法的研究[J]. 电子测量与仪器学报,2006,20(2):16-20.
LI W H, CHEN W M, GONG W G. A novel feature selection method for face recognition[J]. Journal of Electronic Measurement and Instrument, 2006,20(2): 16-20.

[14] 张滢,齐美彬,周云,等. 基于特征提取和多示例学习的图像区域标注[J]. 电子测量与仪器学报,2014,28(8):909-914.
ZHANG Y, QI M B, ZHOU Y. Image region labeling on feature extraction and multiple-instance learning[J]. Journal of Electronic Measurement and Instrument, 2014,28(8):909-914.

[15] 段敬红,栾丹. 人民币号码自动识别方法研究[J]. 计算机工程与科学,2008,30(1):66-68.
DUAN J H, LUAN D. Research on an automatic number recognition method for RMB banknotes[J]. Computer Engineering & Science, 2008,30(1):66-68.

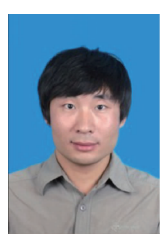
[16] 马瑾,陈立潮,张永梅. 轮廓跟踪与边沿检测的图像自动识别[J]. 中北大学学报:自然科学版,2006,27(5):431-435.
MA J, CHEN L CH, ZHANG Y M. A research on contour tracking and edge detection based on image automatic recognition[J]. Journal of North University of China: Natural Science Edition, 2006, 27(5):431-435.

[17] SEKITA M, TORAICHI K, MORI R, et al. Feature extraction of handwritten Japanese characters by spline functions for relaxation matching[J]. Pattern Recognition,1988,21(1):9-17.

作者简介

徐小力,1951年出生,博士、教授、博导,现代测控技术教育部重点实验室主任,研究方向为机电系统状态监测与控制。
E-mail: xuxiaoli@bistu.edu.cn

XuXiaoli was born in 1951, Ph. D. degree, supervisor of a Ph. D. student, he is a professor of Key Laboratory of Modern Measurement & Control Technology of Ministry of Education. His main research interests include monitoring and control of electromechanical system.



蒋章雷,1983年出生,博士、助理研究员,研究方向为机电装备运行状态监测检测技术。
E-mail: Jiang_Zhanglei@126.com

Jiang Zhanglei was born in 1983, Ph. D., his main research field is mechanical and electrical equipment operating status monitoring detection technology.