DOI: 10.13382/j.jemi.2016.11.020

东巴经典古籍象形文字智能识别研究*

吴国新[1] 丁春艳[2] 徐小力[1] 王 宁[1]
(1. 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; 2. 中央民族大学 北京 100081)

摘 要:东巴象形文字被国际学界认为是当今世界上唯一还在使用的象形文字,用象形文字书写的经典,称为东巴经。东巴象形文字多具图画特性,结构复杂形式多样且笔划各异。针对东巴文特有的结构进行了深入的研究,主要讨论了东巴象形文的特征提取和文字识别。特征提取是文字识别中很重要的环节,文字识别中特征提取的方法有很多,但由于东巴文字的字型有很多种特点,提出了适合东巴文识别的最优特征提取方法:特征点法、投影法;识别方法:高阶神经网络法。通过实验对该方法进行了验证,结果表明该方法的可行性。

关键词:特征提取;文字识别;东巴象形文字;神经网络
中图分类号:G202;H257 **文献标识码**:A **国家标准学科分类代码**:520.6040

Intelligent recognition on Dongba manuscripts hieroglyphs

Wu Guoxin[1] Ding Chunyan[2] Xu Xiaoli[1] Wang Ning[1]
(1. Key Laboratory of Modern Measurement & Control Technology Ministry of Education, Beijing Information Science & Technology University, Beijing 100192, China;
2. Minzu University of China, Beijing 100081, China)

Abstract: Dongba hieroglyphs by international scholars believe is only used in the world today with hieroglyphics, pictographic writing classic, known as dongba script. It has the picture characteristics and the complex structure of different forms & strokes. According to the structure characteristic of Dongba we do some in-depth study. Feature extraction and character recognition of Dongba Pictograph are mainly discussed. Feature extraction is a very important link in the character recognition. there are many methods in feature extraction, but due to the many kinds of characteristics of Dongba font proposed the optimal feature extraction method for Dongba identification: the characteristic point method, projection method; identification method: high order neural network method. The method has been verified by experimental. The results show that the method is feasible.

Keywords: Feature extraction; character recognition; Dongba hieroglyphs; neural network

1 引 言

由我国纳西族祖先在远古创造的一种独特的象形文字书写而成,该象形文是当今公认的世界上唯一还在使用的象形文字,是一份很有学术价值的文化瑰宝[1-3]。我国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”,急待进行高效高质量抢救与传承。用象形文字书写的经典,称为东巴经,其宝贵资料记载着人类文化史中典型社区文化发展变迁,积淀积累着中华远古文化源流变迁[4-8]。东巴象形文字现在有 1 800 多个,是一种属于图画记事和表意文字中间发展阶段的原始象

收稿日期: 2016-09 Received Date: 2016-09
*基金项目:国家社科基金重大项目(12&ZD234)、现代测控技术教育部重点实验室开放项目(KF20161123201,KF20161123202)资助

形文字符号系统,其形态结构复杂,它是介乎于图画文字和表意文字之间的一种文字符号。若想对东巴文字进行识别,首先要对文字进行特征提取。所提取的有效特征不仅要能够体现文字的特征,更重要的是要能够体现与其他文字的区别。而东巴象形文字多具图画性特点,结构复杂,形式多样笔划各异,这就为特征提取工作增加了一定的难度,但是也使特征提取显得尤为重要[9-10]。

在过去的几十年里,科研人员大量的科学研究已经形成了对各种字符进行特征提取和识别的方法。其中包括统计方法、结构方法以及神经网络方法等[11-15]。本文旨在吸取其他文字特征提取方法的基础上,针对纳西东巴文字的特点对其进行深入研究,采用特征点提取和投影法相结合的方法对东巴文字进行特征提取,采用高阶神经网络的方法对其进行文字识别。

2　东巴象形文的识别原理

东巴经书古籍特征获取研究主要涉及:东巴经书图像颜色统一化处理,将三维数值转化为一维数值;改进中值滤波方法,用自适应中值滤波方法进行滤波,去除东巴经书图像噪声信号,保护图像的边缘信息与细节;提取东巴经书图像的灰度直方图并选取最佳阈值对经书进行图像笔画特征突出化,再经过比较分析,选取 Canny 边缘检测器对东巴经书进行边缘检测。

单纯采用模板匹配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,但是对于形态结构复杂、轮廓不够清晰的东巴象形文字来说,模板匹配就不能进行正确匹配。单纯采用网络反馈法对所有的东巴文字进行分类识别时,识别率高,但是涉及到迭代计算的问题,识别过程中计算量大,用时较长。所以根据实验分析:先用模板匹配法对简单、易识别的字符进行粗分类,然后采用网络反馈法对一些结构复杂且识别困难的字符进行识别,这样虽然相对于单纯使用网络反馈法进行识别用时长,但是识别正确率会比使用模板匹配法高。

具体东巴象形文字的识别过程主要有以下步骤:首先将采集到的东巴经文用 300 dpi 的数字化扫描仪进行扫描存储,其次将扫描后的图片经过图像预处理,再利用算法对东巴文图像符号进行分析,提取文字特征,最后对文字特征通过识别技术进而达到识别东巴文的目的。东巴文识别技术包括图像预处理和文字识别两个阶段。其识别原理如图 1 所示。

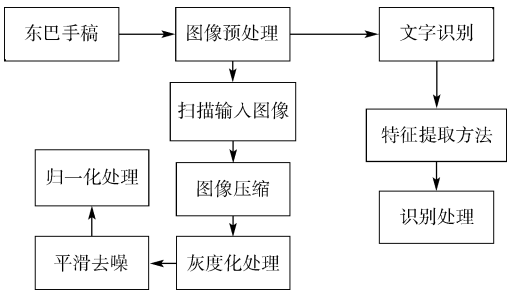

图 1　识别过程

Fig. 1　the Process of Recognition

3　东巴文字的特征提取

东巴文字的特征提取:采用特征点提取和投影法相结合的方法对东巴文字进行特征提取。

3.1　特征点法

根据东巴文字的特点,可以用特征点来描述东巴文字的特征,特征点主要是指东巴字的端点 D、交点 J、歧点 Q 及折点 Z。交点是线条相交的点,其中有 3 条线相交的点叫三叉点,4 条线相交的点叫四叉点。折点是在书写方向上其文字笔画显著变换的点。用文字的端点 D、折点 Z、歧点 Q 和交点 J 来表示东巴文字的特征。记特征向量 $C=(D,Z,Q,J)$。这些点就是东巴文字的特征点。

令 C 为东巴字特征点表达式,C_k 为东巴字特征点,$K=1,2,\cdots,k$,S_k 是特征点的类型,(X_k,Y_k) 是特征点在东巴字点阵图中的坐标,$\{M_k\}$ 是特征点其他属性的集合。则:

$$C=\{C_k\},\ k=1,2,\cdots,k \tag{1}$$

$$C_k=(S_k,(X_k,Y_k),\{M_k\}) \tag{2}$$

特征点的判别方法如下:

设当前点为 M、M_0、M_1、M_2、M_3、M_4、M_5、M_6、M_7 为 8 个邻接点,$d=\sum_{i=1}^{7}|m_i-m_{j+1}|$。

则特征点的判别条件为 1)若 $d=2$,则 m 为端点;2)若 $d=6$,则 m 为三叉点;3)若 $d=8$,则 m 为四叉点;设 i,j 为 m 的两邻点方向,若 $d=4$ 且

$|i+j|=2$ 则 m 为折点。

通过用东巴字特征点 C_k 作为特征，可以有效的对线条清晰明确的东巴字进行特征提取，操作简单方便。

3.2　投影法

将笔画结构复杂，用特征点方法不容易对东巴文字进行特征提取时，就采用投影法进行。具体方法如下。

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：

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

$$u(y)=\sum_{x=0}^{m-1}f(x,y)\quad 0\leqslant x,y\leqslant m-1 \tag{4}$$

式中：$f(x,y)$ 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 $f(x,y)$ 的投影值作为文字的特征：将步骤 1）中的每个网格再分成 $N_1\times N_2$ 个子区域，N_1 为水平方向的网格数，N_2 为竖直方向上的网格数，且 $N_1<10$，$N_2<10$，则根据步骤 2）中的投影公式可以得到二维图形 $f(x,y)$ 的投影值为：

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

$$u(y)=\sum_{x=1}^{N_1}f(x,y)\quad 1\leqslant y\leqslant N_2 \tag{6}$$

式中：$f(x_i,y_i)$ 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 $f(x_i,y_i)$ 值为 1，否则 $f(x_i,y_i)$ 的值为 0。

4　字符识别

根据提取的东巴象形文字特征，结合相似法和网络反馈法对东巴象形文字进行识别，包括以下两种方法。

1）针对于结构笔画简单，形态结构各异且很容易辨识的东巴象形文字采用相似法进行识别：将提取的东巴象形文字特征通过相似法度量待识别东巴文字样本与现有模板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：

$$T=\frac{\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[M(i,j)-\overline{M}][N(i,j)-\overline{N}]}{\sqrt{\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[M(i,j)-\overline{M}]^2\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}[N(i,j)-\overline{N}]^2}} \tag{7}$$

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，$\overline{M}$ 为模板东巴文字特征向量的平均值，$\overline{M}=\frac{1}{st}\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}M(i,j)$，$\overline{N}$ 为待识别东巴文字特征向量的平均值，$\overline{N}=\frac{1}{st}\sum_{i=0}^{s-1}\sum_{j=0}^{t-1}N(i,j)$。

2）对于结构笔画复杂，形态结构类似的东巴象形文字采用网络反馈方法进行识别，步骤如下。

（1）将待识别的东巴象形文字作为训练集，从训练集中得到一个训练样本的特征向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$，权系数向量为 $\boldsymbol{W}=(w_1,w_2,\cdots,w_n)$，令期望输出是 $\boldsymbol{D}=[D_1,D_2,\cdots,D_m]$。

（2）将训练样本进行分类，采用如下公式完成对训练样本的分类：

$$Z=f\Big(\sum_{i=0}^{n}w_ix_i+\sum_{1<i<j<n}w_{i,j}x_iy_j+\cdots w_{1,2,\cdots,n}x_1x_2\cdots x_n\Big) \tag{8}$$

式中：w_i 为权系数，Z 为样本输出。

（3）通过如图 2 所示的高阶神经网络结构在 $\boldsymbol{X}$ 和外加一个常量作为输入送至神经网络结构中进行学习，组成乘积层结构单元，乘积层单元节点求和得到求和层。乘积层与输入层之间的权值均为恒定常量 1，因此乘积层可以由输入层唯一确定。乘积层与求和层之间的权值矩阵，需要通过网络训练得到。

网络的训练过程采用梯度下降算法的在线学习方式，选用 S 函数作为激活函数，令 $t=0$，最大

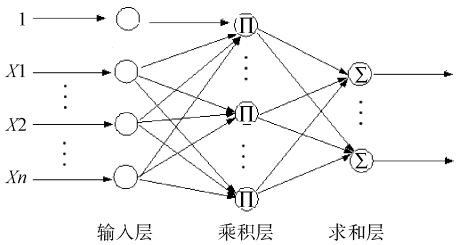

图 2　高阶神经网络结构

Fig. 2　Higher order neural network structure

迭代次数设为 10 000 000 次,训练误差为 0.000 1,学习速率为 0.8 对网络进行训练。在 $\boldsymbol{X}$ 输入下通过网络反馈得出实际输出 y_z 如下:

$$y_z = f\left(\sum_{s=1}^{n_{K-2}} w_{sz}^{k=K-1}\cdots f\left(\sum_{j=1}^{n_1} w_{jp}^{k=2} f\left(\sum_{i=1}^{n} w_{ij}^{k=1} x_i\right)\right)\right)$$

$$z = 1,\cdots,m \tag{9}$$

式中:$f(\cdot)$ 是 Sigmoid 函数,w_{ij}^k 表示第 $k-1$ 层的节点 i 连接到第 K 层的节点 j 的权值,

Sigmoid 函数具体表达式为:

$$f(*) = \frac{1}{1+e^{-*}} \tag{10}$$

(4)从求和层开始调整权值,采用以下公式进行修正权值实现权值更新:

$$w_{ij}^k(t+1) = w_{ij}^k(t) + \Delta w_{ij}^k(t)$$

$$j = 1,\cdots,n_k,\ i = 1,\cdots,n_{k-1} \tag{2}$$

式中:$\Delta w_{ij}^k(t)$ 是权值修正项,$\Delta w_{ij}^k(t) = \alpha\delta_j^k x_i^{k-1}$;其中 α 为学习步长。

将各求和层的权值通过公式进行更新,直接将 $i = 1,\cdots,n_{k-1}$ 分别带入式中进行新权值的计算。更新完权值后将训练样本代入式(9)中重新计算输出,并计算其与期望输出的误差,直到该误差小于预先设定的阈值,则判定模板东巴文字与待识东巴文字样本相似程度最高;否则置 $t = t+1$,并返回步骤(3)重新识别。

将经过训练完毕最终得到的权值矩阵保存,用于步骤(2)中分类器进行文字识别。

5　实验结果

根据以上的理论分析对典型东巴象形文字进行了实验验证研究。图 3 是根据上述特征提取理论处理步骤对东巴象形文中的文字进行特征提取。

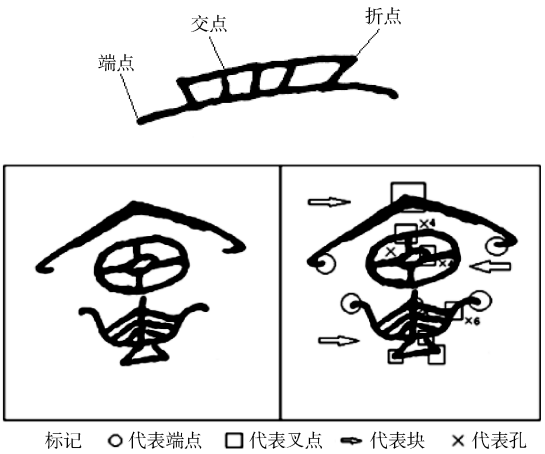

图 3　对东巴象形文中的文字进行特征提取

Fig. 3　Feature extraction of Dongba hieroglyphs

通过以上识别方法进行实际的测试:我们选用待识别象形文字为 1 500 个,分成两组进行识别,一组为文字结构比较简单,如  等共 675 个,另一组结构相对复杂,如  等共 825 个,经过识别得到如下实验数据(见表 1)。

表 1　实验结果统计表

Table 1　Table of experimental results

实验组	实验字数	识别字数	识别率/(%)	20 个东巴文用时/s
一组	675	672	99.56	0.69
二组	825	771	93.45	0.81

通过以上实验数据得出:当识别第一组东巴文字时,识别率几乎达到 100%,故此识别方法能快速有效对东巴文进行识别。当识别第二组时,识别率相对第一组来说有所降低,经分析其中包括以下原因,东巴文字由不同老东巴祭司手写完成,对于结构复杂文字,因为每个人的书写方式与笔画方式不同,每个字都可能会造成一定误差,导致识别率下降。

6　结　论

本文提出了符合东巴文字的一个特征提取方法,通过特征点提取方法使文字进行特征提取时简单、方便,但是对有些复杂的象形文字进行特征提取时,需要采用投影法,投影法抗干扰能力强,易于

实现。通过对大量文字识别技术的分析,总结了一套适合东巴象形文字的识别方法,通过此方法对东巴文字进行识别时,识别率比较高,用时短,使东巴文的文字识别达到很好的效果。

参考文献

[1] 王建明, 王树斌, 陈仕品. 基于数字技术的非物质文化遗产保护策略研究[J]. 软件导刊,2011, 10(8): 49-51.
WANG J M, WANG SH B, CHEN SH P. Research on the Digital protection strategies of intangible cultural heritage [J]. Software Guide, 2011, 10(8): 49-51.

[2] 郑丽萍. 活着的象形文字——纳西东巴图画文字[J]. 艺术与设计(理论), 2009(12): 311-313.
ZHENG L P. The living hieroglyphs, the picture and hieroglyphs of Naxi Dongba [J]. Art and Design, 2009(12): 311-313.

[3] 王树勋, 娥满. 取法自然:纳西东巴文美学风格分析[J]. 云南农业大学学报:社会科学, 2014, 8(6): 54-58, 63.
WANG SH X,E M. An analysis of aesthetic style of the Naxi Dongba language with inspiration from nature [J]. Journal of Yunnan Agricultural University: Social Science, 2014, 8(6): 54-58, 63.

[4] 郭品, 诸昆雄. 丽江古城:一部活的历史[J]. 云南档案, 2004 (2): 40-42.
GUO P, ZHU K X. The old town of Lijiang: A living history[J]. Yunnan Archives, 2004 (2): 40-42.

[5] 张志宏. 国家非物质文化遗产保护与传承依托研究:以纳西东巴画为例[J]. 大众文艺:非遗研究, 2013(22): 9-10.
ZHANG ZH H. Study on the protection and inheritance of national intangible cultural heritage:The Naxi Dongba painting [J]. Popular literature: Intangible Cultural Heritage, 2013(22): 9-10.

[6] 仝艳锋. 民族文献遗产隐性信息特征探讨[J]. 内蒙古社会科学:汉文版,2014, 35 (1): 138-143.
TONG Y F. Research on the characteristics of the hidden information of the national heritage[J]. Inner Mongolia Social Sciences: Chinese, 2014, 35 (1): 138-143.

[7] 胡莹. 档案学视野下的东巴古籍文献遗产保护研究[J]. 档案学通讯, 2015 (2): 65-67.
HU Y. Dongba literature heritage from the perspective of the protection of the archives [J]. Archives Science Bulletin, 2015 (2): 65-67.

[8] 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 Dongba manuscripts [J]. Journal of Beijing Information Science & Technology University, 2015,30(5): 40-43.

[9] 陈强, 田杰, 黄海宁, 等. 基于统计和纹理特征的 SAS 图像 SVM 分割研究[J]. 仪器仪表学报, 2013, 34(6): 1413-1420.
CHEN Q, TIAN J, HUANG H N, et al. Study on SAS image segmentation using SVM based on statistical and texture features [J]. Chinese Journal of Scientific Instrument, 2013, 34 (6): 1413-1420.

[10] 代雷, 吴迪, 张健. 基于 OpenCV 视觉库的 ESPI 图像增强技术研究[J]. 电子测量与仪器学报, 2013, 27(10): 975-979.
DAI L, WU D, ZHANG J. Study on ESPI image enhancement technology based on OpenCV [J]. Journal of Electronic Measurement and Instrument, 2013, 27(10): 975-979.

[11] 闫连山, 熊如刚, 李晓银, 等. 基于 DSP 的直线特征快速提取算法研究[J]. 电子测量技术, 2013, 36(5): 68-71.
YAN L SH, XIONG R G, LI X Y, et al. Research on rapid linear feature extraction algorithm based on DSP [J]. Electronic Measurement Technology, 2013, 36 (5): 68-71.

[12] 冯夫健, 张乾, 林鑫, 等. 字符特征提取方法[J]. 软件导刊, 2012, 11(1): 18-19.
FENG F J, ZHANG Q, LIN X, et al. Method for character feature extraction [J]. Software Guide, 2012, 11(1): 18-19.

[13] 王树东, 何明. LabVIEW 在数据采集系统中的应用研究[J]. 国外电子测量技术, 2014, 33(6): 103-106.
WANG SH D, HE M. Data acquisition system applied research based on LabVIEW [J]. Foreign Electronic Measurement Technology, 2014, 33(6): 103-106.

[14] LV W T, YU Q Z, YU W X. Water extraction in SAR images using GLCM and support vector machine[C]. 2010 IEEE 10th International Conference on Signal Processing (ICSP), 2010: 740-743.

[15] AMIRMAZLAGHANI M, AMINDAVAR H. Two novel bayesian multiscale approaches for speckle suppression in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(7): 2980-2993.

作者简介

吴国新,1977 年出生,2000 年于北京机械工业学院获得学士学位,2003 年于北京机械工业学院获得硕士学位,2011 年于北京理工大学获得博士学位,现为北京信息科技大学副研究员,主要研究方向为信息化技术与机电装备运行状态检测技术。

E-mail:wgx1977@ bistu. edu. cn. com

Wu Guoxin was born in 1977, and received his B. Sc. degree in 2000 from Beijing Insititute of Machinery Industry, received his M. Sc. degree in 2003 from Beijing Insititute of Machinery Industry, received his Ph. D. degree in 2011 from Beijing Institute of Technology, now he is Associate Research in Beijing Information Science and Technology University. His main research interests include Information technology and operation state detection technology of mechanical and electrical equipment.