

基于小波分析及 KNN 的民族文字分类方法

郭 海, 赵晶莹, 苏 飞

(大连民族学院 计算机科学与工程学院, 辽宁 大连 116600)

摘 要: 提出一种基于小波分析的少数民族文字分类识别方法. 该方法采用多辨识小波分解, 从而获得小波能量和小波能量比例分布的特征描述, 结合少数民族文字文本图片的纹理特征, 选择加权 KNN 分类器. 实验证明: 该识别方法对藏文、西双版纳傣文、纳西象形文、维吾尔文、德宏傣文和彝文 6 种常用的少数民族文字及汉字、英语共 8 种文字的分类测试达到 96% 的识别效果.

关键词: 少数民族文字; 语种识别; 小波分析; K 近邻

中图分类号: TP391.43

文献标识码: A

文章编号: 1000-7180(2010)02-0107-04

A Chinese Minority Script Recognition Method Based on Wavelet Feature and KNN

GUO Hai, ZHAO Jing-ying, SU Fei

(School of Computer Science and Engineering, Dalian Nationalities University, Dalian 116600, China)

Abstract: The method of recognizing the kinds of Chinese minority scripts based on wavelet analysis and K-Nearest Neighbour (KNN) is presented which adopts wavelet decomposition that obtains feature descriptor of wavelet energy and wavelet energy distribution proportion. Combined with the texture feature of Chinese minority scripts, radially classification in Feature-Weighted K-Nearest Neighbour (FWKNN). Among Chinese, English and Chinese minority scripts such as Tibetan, Tai Lue, Naxi Pictographs, Uighur, Tai Le, Yi, the experimental results show the recognition rate is up to 96%.

Key words: Chinese minority script; script recognition; wavelet analysis; K-Nearest Neighbour (KNN)

1 引言

随着我国计算机技术的发展, 少数民族信息处理已经逐渐成熟起来^[1-5]. 字体识别是判断文本图像中字体的类型, 是计算机自动文档分析和处理中重要的研究内容之一. 现有的少数民族 OCR 系统主要面向于“识字”的层面, 语种识别远没有引起人们应有的重视, 关于少数民族文字语种识别的研究还处于一片空白. 文中基于小波分析及 FWKNN 提出了一种少数民族文字语种识别方法.

2 基于小波分析的特征提取

在数字图像处理中, 常常对小波变换进行二进

制离散化, 运用离散二进制小波可以对图像进行多通道多分辨率分析, 利用小波提取的图像高频信息可获得图像的纹理特征^[6-7].

少数民族文本文字识别主要选择小波能量分布特征和小波能量比例分布特征两种特征向量, 在此首先要介绍一下图像的平均能量、小波能量分布特征、小波能量比例分布特征三个定义.

2.1 图像的平均能量

一个尺寸为 $N \times N$ 的图像的平均能量定义为:

$$f = \sum_{m=1}^N \sum_{n=1}^N \frac{f^2(m, n)}{N^2}, \quad m, n = 1, \dots, N \quad (1)$$

在进行小波分解前, 需要采用式(2) 将图片能量归一化处理.

$$f(m, n) \leftarrow \frac{f(m, n)}{(\text{energy}f)^{1/2}},$$

$$m, n = 1, \dots, N \quad (2)$$

2.2 小波能量分布特征

对图像进行频域分解后可以得到四个细节子图,即 LL、LH、HL 和 HH. 子带 LL 为低频成分,是近似分量,集中了原始图像的大部分信息. LH、HL、HH 为高频成分,代表了原始图像的细节信息. 对每次分解得到的 LL 还可以再次进行小波分解,以此类推,图 1 给出了两层小波分解示意图.

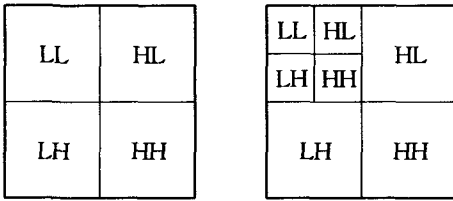


图 1 两层小波分解示意图

细节子图第 k 阶小波分解的平均能量分布定义为式(3)~(6),其中 $ELH^{(k)}$ 、 $EHH^{(k)}$ 、 $EHL^{(k)}$ 分别是图像的多尺度小波特征,称为小波能量分布特征, F_d 称为小波能量分布特征向量.

$$ELH^{(k)} = \sum_{m=(N/2^k)+1}^{N/2^{k-1}} \sum_{n=1}^{N/2^k} \frac{(LH^{(k)}(m, n))^2}{(N/2^k)^2}$$

$$k = 1, \dots, 16N \quad (3)$$

$$EHH^{(k)} = \sum_{m=(N/2^k)+1}^{N/2^{k-1}} \sum_{n=(N/2^k)+1}^{N/2^{k-1}} \frac{(HH^{(k)}(m, n))^2}{(N/2^k)^2}$$

$$k = 1, \dots, 16N \quad (4)$$

$$EHL^{(k)} = \sum_{m=1}^{N/2^k} \sum_{n=(N/2^k)+1}^{N/2^{k-1}} \frac{(HL^{(k)}(m, n))^2}{(N/2^k)^2}$$

$$k = 1, \dots, 16N \quad (5)$$

$$F_d = \begin{bmatrix} EHL^k \\ EHH^k \\ ELH^k \end{bmatrix}, k = 1, \dots, 16N \quad (6)$$

2.3 能量分布比例

小波能量分布特征包括 $ELH^{(k)}$ 、 $EHH^{(k)}$ 、 $EHL^{(k)}$ 三个主要部分,现在通过式(7)~(9),定义了 $EPHL^{(k)}$ 、 $EPHH^{(k)}$ 、 $EPLH^{(k)}$ 三个能量分布比重,分别表示子图 HL、HH 和 LH 在第 k 阶小波分解的能量分布比重.这三个特征构成式(10)中 F_{dp} ,小波能量分布特征.

$$EPLH^{(k)} = \frac{ELH^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}}$$

$$k = 1, \dots, 16N \quad (7)$$

$$EPHH^{(k)} = \frac{EHH^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}}$$

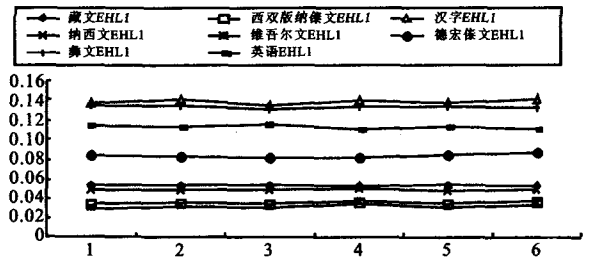
$$k = 1, \dots, 16N \quad (8)$$

$$EPHL^{(k)} = \frac{EHL^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}}$$

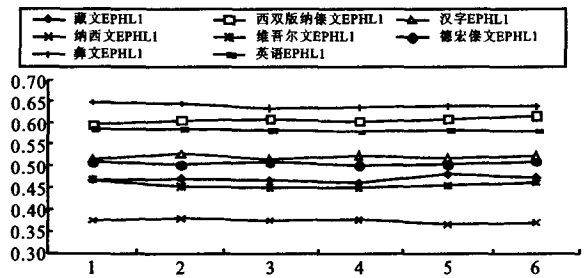
$$k = 1, \dots, 16N \quad (9)$$

$$F_{dp} = \begin{Bmatrix} EPHL^{(k)} \\ EPHH^{(k)} \\ EPLH^{(k)} \end{Bmatrix}, k = 1, \dots, 16N \quad (10)$$

通过对藏文、西双版纳傣文、纳西象形文、维吾尔文、德宏傣文、彝文 6 种少数民族文字及英语和汉字进行小波分解及能量提取,分别求出每种文字的多阶小波能量特征及能量分布比重特征.小波分解中的 k 值选择很重要,如果 k 值选择过低会导致分类效果不佳,如果 k 值过高,会导致特征维数过高,影响识别速度.根据实验研究,发现 $k = 2$,也就是当进行两层分解的时提取的 F_d 和 F_{dp} ,特征就达到了正确分类的要求.图 2 显示的是一阶分解时八种文字水平方向的能量特征曲线.



(a) 一阶分解时水平方向的能量特征曲线



(b) 一阶分解时水平方向的小波能量分布特征曲线

图 2 能量特征曲线

3 KNN 分类器的设计

3.1 标准 KNN 分类器

KNN 算法是一种基于实例的分类算法.首先,对于一个测试样本,计算它与训练样本集中每个样本的相似度,依相似度找出 k 个最相似的训练样本.然后在此基础上给每一个样本类打分,分值是 k 个训练文档中属于该类的样本与测试样本之间的文档相似度之和.对这 k 个样本所属类的分值统计完毕之后,即按分值进行排序^[8].为了分类合理,应当

选定一个阈值,可以认为测试样本属于超过阈值的所有类.具体的算法步骤为:

- (1) 根据特征项集合重新描述训练样本向量;
- (2) 在新样本到达后,根据特征词分词新样本,确定新样本的向量表示;
- (3) 在训练样本中选出与新样本最相似的 k 个样本,计算公式为:

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^M W_{ik} * W_{jk}}{\sqrt{\sum_{k=1}^M W_{ik}^2} \sqrt{\sum_{k=1}^M W_{jk}^2}} \quad (11)$$

式中, d_i 为测试文本的特征向量, d_j 为第 j 类的中心向量, M 为特征向量的维数, W_k 为向量的第 k 维. k 值的确定一般先采用一个初始值,然后根据实验测试的结果调整 k 值,一般初值定为几百到几千;

- (4) 在测试样本的 k 个近邻中,依次计算每类的权重;

- (5) 比较类的权重,将样本分到权重最大的那个类别中.

3.2 基于特征加权的改进算法

KNN 方法是建立在 VSM 模型上的,其样本距离的测度使用欧式距离或余弦距离,各维权值相同,也就是以为各维对于分类的贡献是相同的,这是不符合实际情况的,同等的权重使得特征向量之间距离或夹角余弦的计算不够准确,进而影响分类精度.针对这一不足,提出了基于神经网络和 CHI 的改进 KNN 方法,应用 SOM 神经网络进行 VSM 模型各维权重的计算.该方法首先运用 CHI 概率统计方法进行初步特征提取和模式聚合,其特征权重的计算原理为:如果某一维在各个类别中取值基本相同,那么此维对于样本分类的贡献率就相对较低,如果在各个类别中取值有较大的差异,那么就具有较强的文本分类能力,而方差正好是反应变量分布均匀状态的主要指标.该方法有效地提高了文本分类的精度.提出了利用 SVM 来确定特征的权重,即基于 SVM 特征加权算法.该方法利用 SVM 可以定量确定样本的每个特征与分类的相关度——由分类函数的权重向量给出:

$$W = (w_1, w_2, \dots, w_m)$$

$$= \sum_{i=1}^m y_i \alpha_i x_i, \quad \alpha_i \geq 0 \quad (12)$$

其中, α_i 为每个样本对应的 Lagrange 乘子.特征权重确定后,就可以修改样本之间的距离函数以便更好地反映实际问题.

4 实验设置及结果

4.1 实验设置

为了验证本文方法的有效性,选取了藏文、西双版纳傣文、纳西象形文、维吾尔文、德宏傣文、彝文 6 种常用的少数民族文字及汉字、英语共 8 种文字进行分类测试.由于语种识别与文本内容具有无关性,文中编写一个样本自动生成程序,如图 3 所示,采用随机函数随机生成每种文字的图片.每种文字的文本图片生成 200 张,分别从以上 8 种图片种各取 100 张构成样本集 A,库 A 有图片 8 种共 800 张.把剩余的图片作为样本集 B,样本集 B 有图片 8 种共 800 张.

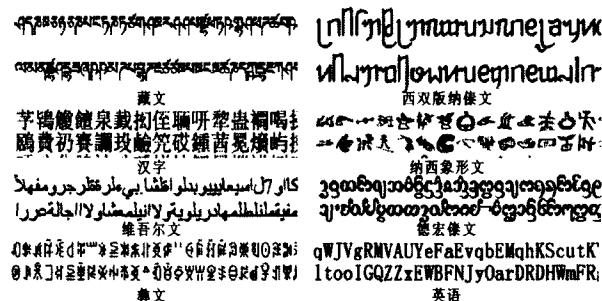


图3 8种文字的文本图片

4.2 实验结果

将样本集 A 作为训练集,样本集 B 作为测试集,经实验测试发现当 $K = 15$ 时分类效果最好,当 $K = 15$ 时候的分类结果如表 1 所示.

通过实验发现汉字、英语及西双版纳傣文 3 种样本采用文中方法时分类效果可以达到 100%,但是藏文、纳西象形文、维吾尔文、德宏傣文、彝文有一些误分类,其中藏文有 8 张被误判为纳西文、纳西文有 2 个样本被误判断为藏文,彝文有 5 个被误判断为中文、3 个判断为藏文,德宏傣文有两个样本被误判断为维吾尔文、维吾尔文有 10 个被误判为藏文.实验表明,文中的方法对于少数民族文字种类的识别效果较好,平均识别精度已经达到了 96%.

表1 八种文字分类数据

K 值	藏文	西双版纳傣文	汉字	纳西文	维吾尔文	德宏傣文	彝文	英语	平均精度
K = 15	92%	100%	100%	98%	90%	98%	92%	100%	96.25%

5 结束语

少数民族文字种类识别是一项非常有意义但也是非常困难的工作.文中对这个问题进行了研究,对每种文字的文本图像进行小波分解,在小波图像上提取小波能量特征及小波能量比例特征,用改进的 KNN 分类器对文字种类进行识别.实验表明,文中方法能够有效识别少数民族文字种类信息,平均识别率达到 96%.

参考文献:

- [1] 王维兰,丁晓青,陈力,等.印刷体现代藏文识别研究[J].计算机工程,2003,29(3):37-38.
- [2] 王华,丁晓青,哈力木拉提.多字体多字号印刷维吾尔文字符识别[J].清华大学学报:自然科学版,2004,44(7):946-949.
- [3] 李振宏,高光来,侯宏旭,等.印刷体蒙古文文字识别的研究[J].内蒙古大学学报:自然科学版,2003,34(4):

454-457.

- [4] 郭海,车文刚,聂娟,等.纳西象形文 WEFT 植入技术[J].计算机工程,2005,31(17):203-204.
- [5] 郭海,赵晶莹.纳西象形文字信息处理平台的开发[J].微计算机信息,2006(22):221-223.
- [6] 唐远炎,王玲.小波分析与文本文字识别[M].北京:科学出版社,2004.
- [7] 李小航,刘渊,刘元珍.基于小波多尺度分析的网络流量组合预测方法研究[J].微电子学与计算机,2008,25(1):130-133.
- [8] 贾可亮,樊孝忠,许进忠.基于 KNN 的汉语问句分类[J].微电子学与计算机,2008,25(1):156-158.

作者简介:

郭海男,(1979-),硕士,讲师.研究方向为少数民族信息处理、模式识别.

赵晶莹女,硕士,讲师.研究方向为少数民族信息处理、模式识别.

(上接第 106 页)

参考文献:

- [1] 陈广存.铁路信号概论[M].北京:中国铁道出版社,1995:2-3.
- [2] 当代中国铁路信号编辑委员会.当代中国铁路信号(1996-2000)[M].北京:中国铁道出版社,2002:70-104.
- [3] 邹少文,郭进,杨扬.微机化站间自动闭塞系统[J].西南交通大学学报,2003,38(4):418-422.
- [4] 袁崇义.Petri 网原理[M].北京:电子工业出版社,1997:7-8.
- [5] 魏敏.基于动态故障检测的铁路信号发码器及其 ASIC 实现[J].微电子学与计算机,2003,20(10):1-3.
- [6] 付军,张汝群,谢玉琼.计轴技术的发展及容错型计轴设

备的应用[J].铁道通信信号,2003(12):28-30.

- [7] 杜军威,徐中伟.联锁系统形式化模型的安全性评估[J].微电子学与计算机,2007,24(8):14-18.
- [8] 郭荣佐,王霖.嵌入式系统原理[M].北京:北京航空航天大学出版社,2008:228-310.

作者简介:

郭荣佐男,(1973-),副教授,讲师.研究方向为嵌入式系统研究、嵌入式信号技术及实现等.

郭进男,(1960-),教授,博士生导师.研究方向为铁路信号控制理论与应用、计算机控制系统等.

王霖男,(1970-),讲师.研究方向为计算机网络与信息安全.