文章编号:1001 - 9081(2009)12 - 3360 - 03

基于小波分析及改进二次鉴别函数的民族文种识别

郭 海,赵晶莹
(大连民族学院 计算机科学与工程学院,辽宁 大连 116600)
(guohai@tom.com)

摘 要:为了能够对文档中的少数民族文字种类进行正确地识别分类,提出一种基于小波分析与改进的二次分类函数(MQDF)的少数民族文字种类识别方法。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,利用 MQDF 分类器对少数民族文种进行识别。构建藏文、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 种文字的样本库,采用该方法对少数民族的样本库进行了进行训练和测试。实验结果显示,该方法在多层小波分解的情况下,对于少数民族文种识别的精度好于传统的贝叶斯和 KNN。

关键词:中国少数民族文字;文种识别;小波分析;改进的二次分类函数
中图分类号: TP391.43 **文献标志码**:A

Chinese minority script identification method based on wavelet feature and MQDF

GUO Hai, ZHAO Jing-ying
(*School of Computer Science and Engineering, Dalian Nationalities University, Dalian Liaoning* 116600, *China*)

Abstract: In order to classify the type of the Chinese minority scripts, the method of identifying the kinds of Chinese minority scripts based on wavelet analysis and Modified Quadratic Discriminant Function (MQDF) was presented. Using wavelet energy and wavelet energy distribution proportion as features by wavelet multi-resolution transform, multivariate classifier in MQDF was constructed. A sample data set was built which contained six common Chinese minority scripts: Tibetan, Tai Lue, Naxi Pictographs, Uighur, Tai Le, Yi and Chinese and English in total, some samples were used for training, others were for testing, and the proportions of the training samples in dataset were variant. Obviously, the experimental result shows that, in multi-level decomposition, the method is better than the traditional Bayes and K-Nearest Neighbor (KNN) classification in recognition rate.

Key words: Chinese minority script; script identification; wavelet analysis; Modified Quadratic Discriminant Function (MQDF)

0 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,如清华大学与西北民族大学合作研究的藏文识别[1]、清华大学与新疆大学合作研究的维哈柯多语种识别[2]、内蒙古大学的蒙文识别研究[3]、纳西文信息处理[4-6]等都取得了较大进展。文字种类识别就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。现有的少数民族光字符识别(Optical Character Recognition, OCR)系统主要面向于"识字"的层面,文种识别远没有引起人们应有的重视。

文种识别研究目前在模式识别领域也是一个研究热点。最早采用的是模板匹配方法[7],现在比较流行的方法是采用纹理方法[8-9]。本文根据中国少数民族文字的特点,提出了一种基于小波分析及改进二次鉴别函数(Modified Quadratic Discriminant Functions, MQDF)的少数民族文字种类的识别方法。

1 多分辨率小波分析

近年来,随着小波理论的逐渐成熟,小波分析作为一种数学理论和方法在模式识别领域引起了越来越多的关注和重视,利用小波提取的图像频域信息可获得图像的纹理特征[10]。

对图像进行频域分解后可以得到四个区域,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 给出了三层小波分解示意图。

2 MQDF 分类器设计

2.1 MQDF 分类器

二次分类函数(Quadratic Discriminant Function, QDF)作为分类器已经广泛的应用到各种图像、语音分类上,但是其存在一些缺点:1)存储量过大;2)运算速度慢;3)参数估计(协方差矩阵的估计)易收到噪声影响,从而影响决策性能。针对 QDF 分类器的这些缺点提出了 MQDF 分类器。

其原理如下:设各个类别的先验概率相同,且各类样本均为高斯分布,那么根据贝叶斯准则可推导得到 QDF 为最优分类器。

收稿日期:2009 - 06 - 08;**修回日期**:2009 - 08 - 10。 **基金项目**:国家自然科学基金资助项目(60803096);国家民委项目(07DL07)。
作者简介:郭海(1979 -),男,黑龙江哈尔滨人,讲师,硕士,主要研究方向:少数民族信息处理、模式识别; 赵晶莹(1978 -),女,吉林伊通人,讲师,硕士,主要研究方向:少数民族信息处理、模式识别。

首先对协方差矩阵进行特征值分解及对角化:

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{\Lambda}_i \boldsymbol{B}_i^{\mathrm{T}} \tag{1}$$

那么 QDF 分类函数就被写成了:

$$g(x,\omega_i) = [\boldsymbol{B}_i^{\mathrm{T}}(x-\mu_i)]^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \boldsymbol{B}_i^{\mathrm{T}}(x-\mu_i) + \mathrm{lb}\,|\boldsymbol{\Lambda}_i| = \sum_{j=1}^{d} \frac{1}{\lambda_{ij}}[\boldsymbol{\beta}_{ij}^{\mathrm{T}}(x-\mu_i)]^2 + \sum_{j=1}^{d} \mathrm{lb}\,\lambda_{ij} \tag{2}$$

将 $d-k$ 个最小值的特征值用它们的平均值代替:

$$\delta_i = \frac{1}{d-k}\sum_{j=K+1}^{d} \lambda_{ij} \tag{3}$$

从而得到 MQDF 的距离判别函数的公式为:

$$g_i(x,\omega_i) = \sum_{j=1}^{k} \frac{1}{\lambda_{ij}}[\boldsymbol{\beta}_{ij}^{\mathrm{T}}(x-\mu_i)]^2 + \sum_{j=K+1}^{d} \frac{1}{\delta_i}[\boldsymbol{\beta}_{ij}^{\mathrm{T}}(x-\mu_i)]^2 + \sum_{j=1}^{k} \mathrm{lb}\,\lambda_{ij} + (d-k)\mathrm{lb}\,\delta_i = \sum_{j=1}^{k} \frac{1}{\lambda_{ij}}[\boldsymbol{\beta}_{ij}^{\mathrm{T}}(x-\mu_i)]^2 + \frac{1}{\delta_i} r_i(x) + \sum_{j=1}^{k} \mathrm{lb}\,\lambda_{ij} + (d-k)\mathrm{lb}\,\delta_i \tag{4}$$

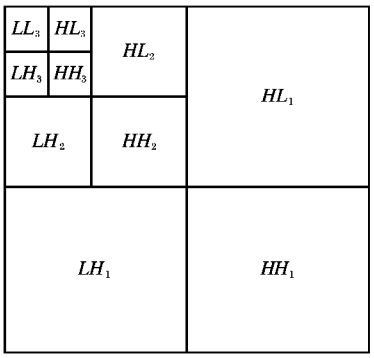


图 1　三层小波分解示意图

2.2　MQDF 的时间复杂度分析

MQDF 相对以前的 QDF 和的二次判别函数 QDF1 来说,MQDF 在时间复杂度上有一定的优势。

QDF 的一般形式如下:

$$g_0(x) = \sum_{i=1}^{n} \frac{1}{\lambda_i}\{\varphi_i^t(x-\mu_M)\}^2 + \mathrm{lb}\prod_{i=1}^{n}\lambda_i \tag{5}$$

一个提高的二次判别函数 MQDF1 形式如下:

$$g_1(x) = \sum_{i=1}^{n} \frac{1}{\lambda_i + h^2}\{\varphi_i^t(x-\mu_M)\}^2 + \mathrm{lb}\prod_{i=1}^{n}(\lambda_i + h^2) \tag{6}$$

再次改良的二次 MQDF 的形式为:

$$g_2(x) = \sum_{i=1}^{k} \frac{1}{\lambda_i}\{\varphi_i^t(x-\mu_M)\}^2 + \sum_{i=k+1}^{n} \frac{1}{h^2}\{\varphi_i^t(x-\mu_M)\}^2 + \mathrm{lb}\left(h^{2(n-k)}\prod_{i=1}^{k}\lambda_i\right) \tag{7}$$

通过以上三个公式可以看出:MQDF 的耗时是 QDF 和 QDF1 的 k/n。k 被认为是固定而不是随着 n 改变原因是不会轻易增加样本学习的数量,k 的值只会随着特征的数量增加,因此 MQDF 的时间复杂度是 $O(n)$。

3　实验设计及结果

3.1　预处理

不同的文字图像其文字的大小、字间距、行间距往往不相同,有些原始的文字块还可能包含许多空格(例如在大多数段落的结尾处),这些因素都给字体识别带来困难,为了消除这些非本质因素的影响,在进行字体识别之前进行预处理是有必要的。预处理的主要目的是:1)将文字块中的文字规范到预先指定的大小;2)将文字块中的行间距和字间距规范到预先指定的大小;3)填补文字块中可能的空格。字体识别方法是与内容无关的,用来填补空格的文字可以直接从原文字块中的任何非空格部分抽取。

3.2　特征提取

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,在此先先介绍三个定义[11]。

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:

$$f = \sum_{m=1}^{N}\sum_{n=1}^{N} \frac{f^2(m,n)}{N^2};\quad m,n = 1,\cdots,N \tag{8}$$

在进行小波分解前,需要采用式(9) 将图片能量归一化处理。

$$f(m,n) \leftarrow \frac{f(m,n)}{(energyf)^{\frac{1}{2}}};\quad m,n = 1,\cdots,N \tag{9}$$

2) 小波能量分布特征。

对图像进行频域分解后可以得到 4 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 给出了三层小波分解示意。

细节子图第 k 阶小波分解的平均能量分布定义为式(10) ~ (12),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,而式(13) 中的 $\boldsymbol{F}_d$ 称为小波能量分布特征向量。

$$ELH^{(k)} = \sum_{m=(N/2k)+1}^{N/2k-1}\sum_{n=1}^{N/2k} \frac{(LH^{(k)}(m,n))^2}{(N/2^k)^2};\quad k = 1,2,\cdots,\mathrm{lb}\,N \tag{10}$$

$$EHH^{(k)} = \sum_{m=(N/2k)+1}^{N/2k-1}\sum_{n==(N/2k)+1}^{N/2k-1} \frac{(HH^{(k)}(m,n))^2}{(N/2^k)^2};\quad k = 1,2,\cdots,\mathrm{lb}\,N \tag{11}$$

$$EHL^{(k)} = \sum_{m=1}^{N/2k}\sum_{n==(N/2k)+1}^{N/2k-1} \frac{(HL^{(k)}(m,n))^2}{(N/2^k)^2};\quad k = 1,2,\cdots,\mathrm{lb}\,N \tag{12}$$

$$\boldsymbol{F}_d = \begin{bmatrix} EHL^k \\ EHH^k \\ ELH^k \end{bmatrix};\ k = 1,2,\cdots,\mathrm{lb}\,N \tag{13}$$

3) 能量分布比例特征。

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(14) ~ (16),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(17) 中 F_{dp},小波能量分布比例特征。

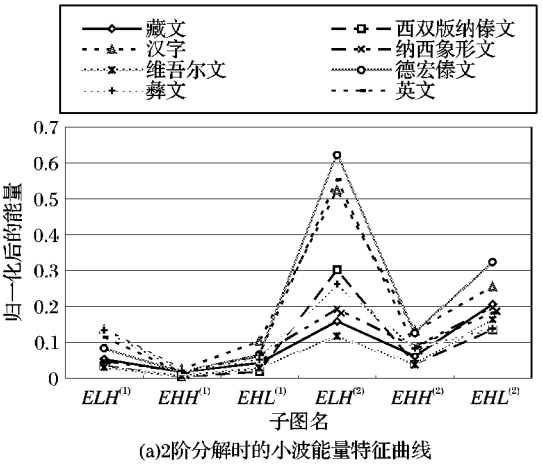
$$EPLH^{(k)} = \frac{ELH^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}};\quad k = 1,2,\cdots,\mathrm{lb}\,N \tag{14}$$

$$EPHH^{(k)} = \frac{EHH^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}};\quad k = 1,2,\cdots,\mathrm{lb}\,N \tag{15}$$

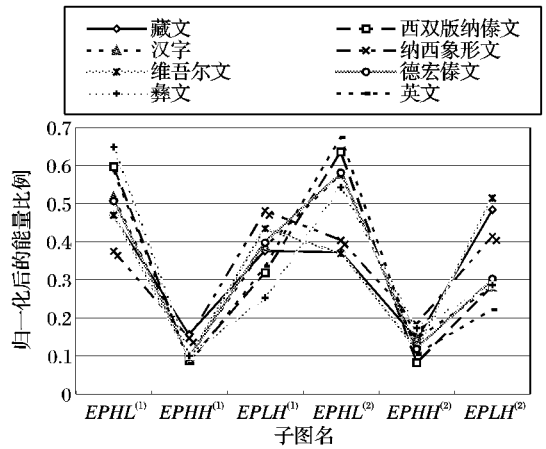
$$EPHL^{(k)} = \frac{EHL^{(k)}}{ELH^{(k)} + EHH^{(k)} + EHL^{(k)}};$$
$$k = 1,2,\cdots,\mathrm{lb}\ N \tag{16}$$

$$\boldsymbol{F}_{dp} = \begin{bmatrix} EPHL^{(k)} \\ EPHH^{(k)} \\ EPLH^{(k)} \end{bmatrix}; k = 1,2,\cdots,\mathrm{lb}\ N \tag{17}$$

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文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。图 2(a) 显示的是 2 阶分解时 8 种文字的能量特征曲线(每种文种选择了 1 张文档图片样本)；图 2(b) 显示的是 2 阶分解时 8 种文字小波能量比例分布特征曲线(每种文字选择了 1 张文档图片样本)。



(a)2阶分解时的小波能量特征曲线

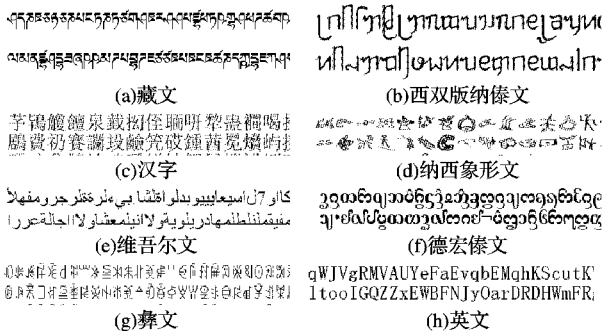


(b)2阶分解时的小波能量比例特征曲线

图 2 少数民族文种的特征曲线

3.3 样本库构建

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 种文字类型进行测试。由于语种识别与文本内容具有无关性，本文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种文档的图片，每张样本图片大小为 640 × 640 的灰度图像。为了更符合实际的需要，从藏文、西双版纳傣文、纳西象形文、维吾尔文、德宏傣文、彝文、汉字、英语等多种出版物上又扫描了一部分的样本，经过处理也加入到样本库中，最终构建了包含有 40000 个各种文种文档图片的样本库。



(a)藏文 (b)西双版纳傣文

(c)汉字 (d)纳西象形文

(e)维吾尔文 (f)德宏傣文

(g)彝文 (h)英文

图 3 8 种文字的文本图片

3.4 实验结果

为了验证本文方法的有效性，分别选取了总样本库的 2%、4%、6%、8%、10% 的样本作为训练集，然后分别用 Bayes、KNN 和 MQDF 三种分类器进行训练和测试，得到如表 1所示的平均识别精度。当小波分解尺度为 4 时得到了最高的分类精度。但是随着小波分解层数的增加，特征的维数也成几何数增加，训练和测试的时间也迅速增长。在实验中发现，分解到四尺度就可以满足中国少数民族文种分类的需要。如图 4 所示，在同样的特征及样本集情况下本文的方法明显优于 Bayes、KNN 两种分类方法。

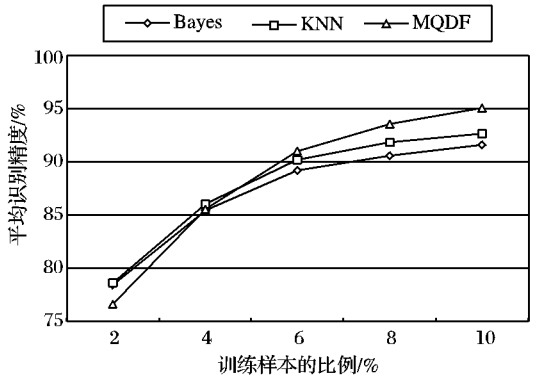


图 4 4 层小波分解时 3 种分类器的平均识别精度

表 1 不同层小波分解时三种分类器的平均识别精度 %

训练样本的比例	两层小波分解			三层小波分解			四层小波分解		
	Bayes	KNN	MQDF	Bayes	KNN	MQDF	Bayes	KNN	MQDF
2	68.86	70.08	71.96	78.04	79.6	79.67	78.44	78.64	76.61
4	76.88	78.50	84.91	84.44	85.67	85.73	85.44	86.03	85.56
6	82.27	83.90	88.9	89.40	89.85	90.16	89.21	90.22	91.03
8	85.47	87.28	90.65	90.84	91.30	93.02	90.6	91.86	93.58
10	86.85	88.24	91.81	91.81	92.40	94.60	91.62	92.67	95.11

4 结语

本文对少数民族文字种类识别进行了研究，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MQDF分类器对文字种类进行识

(下转第 3365 页)

$$Z_{pq} = \frac{p+1}{\pi} \iint_{x^2+y^2 \leq 1} f(x,y) V_{pq}^*(\rho,\theta) \mathrm{d}x\mathrm{d}y$$

对于数字图像,可得出 Zernike 矩的离散近似:

$$Z_{pq} = \frac{p+1}{(N-1)^2} \sum_{x=0}^{N-1} \sum_{j=0}^{N-1} R_{pq}(\rho_{xy}) \mathrm{e}^{-jq\theta_{xy}} f(x,y)$$

其中:

$$r_{xy} = \sqrt{(\frac{2}{N-1}x - 1)^2 + (\frac{2}{N-1}y - 1)^2}$$

且:

$$0 \leqslant r_{xy} \leqslant 1, \theta = \arctan\left(\frac{2y - N + 1}{2x - N + 1}\right)$$

Zernike 矩可以任意构造高阶矩,所以在特征描述方面,Zernike 矩远远优于其他矩特征。

以上使用 HHT 得到了汉字的瞬时频率特征,瞬时频率特征表现的是汉字的局部特征,汉字的位移、噪声、变形对识别效果的影响比较大。为了提高汉字的识别率我们使用 Zernike 矩取得汉字的全局特征、瞬时频率特征,采用串行特征融合技术融合作为最后的识别特征。

3　实验结果

本实验对1000类书写比较规范的手写体汉字进行识别,每类汉字有 50 个样本,40 个样本用于训练,10 个样本用于测试。

将测试样本归一化为64×64大小,对瞬时频率特征采样得到 128 维特征向量,取 11 阶(42 维)Zernike 矩特征向量,归一化后串行融合,采用欧氏距离分类器对汉字进行识别。识别结果如表 1 所示。

表 1　识别结果

方法	识别率/%
HHT	92.4
Zernike	90.9
HHT + Zernike	95.4

由识别结果可以看出,基于 HHT 的手写体汉字识别方法的识别率要高于使用 Zernike 进行汉字识别。而将这两种方法结合起来,会进一步提高手写体汉字识别率。

4　结语

本文提出了一种将 Hilbert-Huang 变换应用于脱机手写体汉字特征提取的方法,实验结果表明这种方法可以有效地提取出汉字的特征。

参考文献:

[1] MORI S, SUEN C Y, YAMAMOTO K. Historical review of OCR research and development [J]. Proceedings of the IEEE, 1992, 80(7): 1029 - 1058.

[2] 王先梅, 杨扬, 王宏. 基于 HMM 的分类器集成方法在脱机手写大写金融识别中的应用[J]. 计算机应用, 2005, 25(12): 6291 - 6293.

[3] 左泽华, 杨扬, 颉斌. 脱机手写汉字识别中 SVM 参数优选问题[J]. 计算机应用, 2006, 26(6): 27 - 31.

[4] HUANG N E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [EB/OL]. [2009 - 04 - 10]. http://keck.ucsf.edu/~schenk/Huang_etal98.pdf.

[5] 刘忠轩, 彭思龙. 方向 EMD 分解与其在纹理分割中的应用[J]. 中国科学: E 辑, 2005, 35(2): 113 - 123.

[6] 崔峰, 沈滨, 彭思龙. 基于 EMD 细化四元数谱的纹理分割[J]. 计算机应用, 2005, 25(3): 573 - 576.

[7] BULOW T H, SOMMER G. Hypercomplex signals — A novel extension of the analytic signal to the multidimensional case [J]. IEEE Transactions on Signal Processing, 2001, 49(11): 2844 - 2852.

[8] KHOTANZAD A, HONG H Y. Invariant image recognition by Zernike moments [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(5): 489 - 497.

(上接第 3362 页)

别。实验表明,本文的方法能够有效识别少数民族文字种类信息,在四尺度小波分解的情况下平均识别率达到 95.11%。本文仅仅是对少数民族文种分类进行了初步的研究,还有些问题没有完善,有待进一步研究:在特征提取时没有考虑降维的问题,在下一步工作中将研究如何使用 LDA 或者 PCA 降低样本的维度,提高识别的速度;在文种识别时候,仅仅对不同少数民族的文种进行了识别,而同种文字的不同字体的识别有待在今后的工作中进一步的验证。

志谢:感谢袁留凯、杨茂盛和唐尧在实验过程中提供的协助工作。

参考文献:

[1] 王维兰, 丁晓青, 祁坤钰. 藏文识别中相似字丁的区分研究[J]. 中文信息学报, 2002, 16(4): 60 - 65.

[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] GUO HAI, ZHAO JING-YING. The design and realization of the Naxi pictographs information processing system [J]. WSEAS Transactions on Systems, 2009, 6(2): 302 - 311.

[5] 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.

[6] GUO HAI, ZHAO JING-YING, LIU YONG-KUI, *et al.* Naxi pictographs information processing based on Web embedding fonts technology [J]. Journal of Computational Information Systems, 2009, 5(1): 495 - 501.

[7] HOCHBERG J, KELLY P, THOMAS T, *et al.* Automatic script identification from document images using cluster-based templates [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(2): 176 - 181.

[8] SUEN C Y, BERGLER S, NOBILE N, *et al.* Automatic identification of oriental and other scripts in image documents [J]. International Journal of Computer Processing of Oriental Languages, 2005, 18(2): 77 - 94.

[9] PATI P B, RAMAKRISHNAN A G. Word level multi-script identification [J]. Pattern Recognition Letters, 2008, 29(9): 1218 - 1229.

[10] 张振宇, 黄崇林, 谭恒松. 基于小波变换的图像识别算法[J]. 计算机应用, 2007, 27(12): 97 - 99.

[11] 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.