基于多特征融合的东亚文种识别

王 刚[1] 靳彦青[2] 刘立柱[1] 储瑞来[3]
(解放军信息工程大学 郑州 450002)[1] (国家数字交换系统工程技术研究中心 郑州 450002)[2]
(南京理工大学 南京 210094)[3]

摘 要 针对目前基于统计特征和符号匹配的识别方法对字体较敏感的问题,提出一种基于多特征融合的东亚文种识别算法。该算法首先分析并提取高频形状特征、排版特征以及字符复杂度特征,然后采用模糊集贴近度准则进行识别。实验结果表明,该算法具有较高的识别准确率,并对不同字体具有较强的鲁棒性。

关键词 文种识别,多特征,字符复杂度特征,贴近度
中图法分类号 TP391 **文献标识码** A

East Asian Script Identification Based on Multi-feature

WANG Gang[1] JIN Yan-qing[2] LIU Li-zhu[1] CHU Rui-lai[3]
(PLA Information Engineering University, Zhengzhou 450002, China)[1]
(National Digital Switching System Engineering & Technology Research Center, Zhengzhou 450002, China)[2]
(Nanjing University of Science & Technology, Nanjing 210094, China)[3]

Abstract Script identification has important applications in the field of document image information retrieval. An east asiatic script identification approach was proposed based on multi-feature. Compared to traditional identification method based on statistical characteristics and symbols matching, the algorithm first analyzes and extracts the token shape matching features, layout features and character complexity features, and then uses closeness degree of fuzzy sets to identify. The experimental results show that the algorithm has higher recognition accuracy and strong robustness to different fonts.

Keywords Script identification, Multi-feature, Character complexity features, Closeness degree

1 引言

随着网络通信技术和信息处理技术的快速发展,社会信息化程度不断提高,文件、档案等文字材料通过图像数据采集设备生成的文档图像作为信息传递和存储的主要媒介,在信息处理和传输系统中应用越来越广泛。随着全球信息化步伐的加快,网络中的文档图像不再由单一语言文字组成。文档图像文字种类的自动识别是对以图像形式存储的各种文字,提取能用于计算机识别的特征,以实现不同文种的自动划分。在文档图像信息检索中,作为OCR(Optical Character Recognition)系统的前端处理技术,文种识别具有重要的应用价值。

目前,针对文种识别进行的研究可以分为基于统计特征、基于符号匹配和基于纹理特征3类。基于统计特征的算法主要有字符上凹面投影法[1]、字符形状编码法[2]以及灰度投影法[3]等。基于符号匹配的算法主要有基于特殊字符的模板匹配法[4]和基于聚类的模板匹配法[5]。基于统计特征和符号匹配的文种识别算法识别准确率较高,但对字体、噪声、倾斜等适应性较差。基于纹理特征的算法[6]主要有灰度级共生矩阵法[7]、Gabor滤波器法[8]、小波变换法[9]等,这类方法的缺点是计算量较大[10]。

鉴于以上文种识别算法中存在的一些问题,提出一种基于多特征融合的东亚文种识别算法。该算法首先分析并提取高频形状特征、排版特征以及字符复杂度特征,然后采用模糊集贴近度准则进行识别,在保证较高识别性能和较小计算量的前提下,对不同字体的文档图像具有较强的鲁棒性。

2 识别特征分析

本文中,东亚文种包括中文、日文和韩文,通过对各语言文档的分析对比,分别提取字符复杂度特征、版式特征和高频形状特征等进行研究。

2.1 字符复杂度特征

通过字符内垂直和水平方向黑游程的最大值来描述字符的复杂度。如图1所示,字母"g"在垂直方向的最大黑游程数为3,水平方向为2,因此用(3,2)表示其字符复杂度,即垂直

到稿日期:2012-03-22 返修日期:2012-06-16 本文受国家"863"计划基金项目(2009AA011202)资助。
王 刚(1981—),男,硕士,讲师,主要研究方向为信息处理、模式识别,E-mail:angwzhg@163.com;靳彦青(1983—),女,硕士,工程师,主要研究方向为无线通信、网络安全;刘立柱(1949—),男,教授,博士生导师,主要研究方向为传真信号和网络信号分析;储瑞来(1982—),男,硕士,工程师,主要研究方向为信号分析、网络通信。

方向复杂度为3，水平方向复杂度为2。

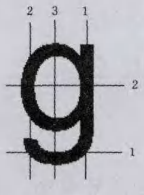


图1 字符复杂度的游程描述

根据定义，可以统计得到中文、日文和韩文字符垂直方向的复杂度分布，如图2所示。

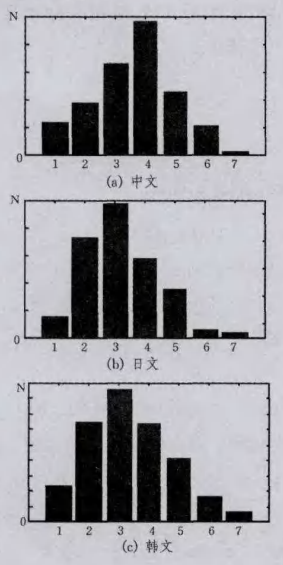


图2 垂直方向的复杂度分布(横坐标为复杂度，纵坐标为字符数，复杂度大于等于7的按照7计算)

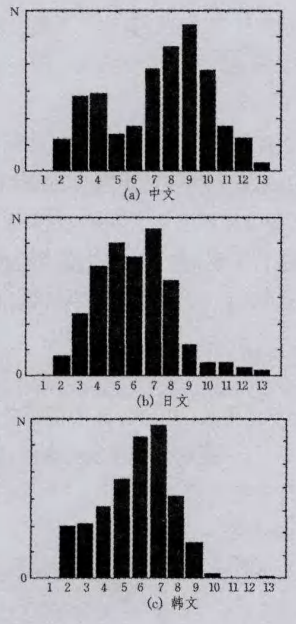


图3 东亚文种字符复杂度的分布(横坐标为复杂度，纵坐标为字符数，复杂度大于13的按照13计算)

对中文、日文和韩文水平和垂直方向复杂度之和的分布情况进行统计，结果如图3所示。从图中可以发现，中文文档中字符复杂度之和大于8的字符在整篇文档中占有很大的比例，约为40%左右，而日文和韩文中这样的字符则较少出现。可见，中文文档的字符复杂程度明显大于日文和韩文文档，因此字符复杂度特征可以作为中文文档的识别特征。

2.2 版式特征

通过对3种东亚语言文档的观察比较，发现韩文文档在版式特征上与中文和日文存在显著差异。中文和日文文档通常由单字直接构成语句，而韩文文档则首先由单字构成长度不等的词语，再由若干词语组成语句。排版时，中文和日文语句间的距离大于字符间的距离，而韩文则是词语之间的距离大于词语内部各字符之间的距离。因此，对东亚文种的文本行按照中文和日文的版式特征，以整句为单位进行分割，用矩形框标记语句分割的结果，如图4(a)、(c)和(e)所示，由于在这种规则下韩文中所标记的是词语而非整句，因此其矩形框内的字符数相对中文和日文要少得多。进一步统计3种语言语句的长度分布，即矩形框内字符个数的分布情况，结果如图4(b)、(d)和(f)所示。可见中文和日文中大量的语句长度大于10个字符，而韩文的语句长度(实际上是词语长度)则集中分布于3字符至6字符之间。因此该特征可作为韩文的识别特征。

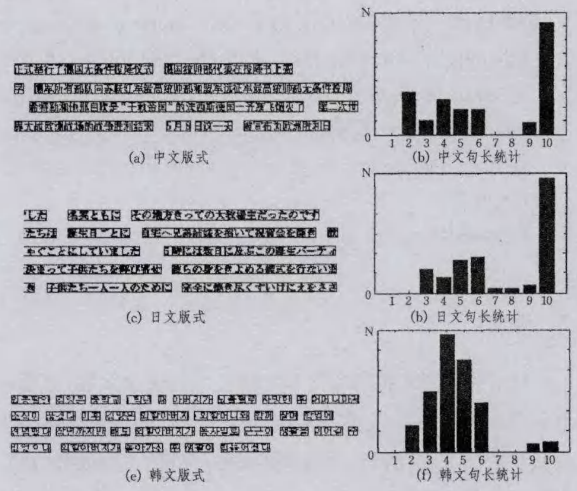


图4 东亚文种版式特征(横坐标为语句长度，长度大于10的以10计，纵坐标为语句的数量)

2.3 高频形状特征

类似于英语中“the”的使用频率很高，东亚语言中同样有部分字符的使用十分频繁，比如汉字中的“的”字，日文中比较常见的“の”，而韩文中“ㅇ”(字符的一部分)的出现频率也很高。依据在本语言中出现频率很高，而在其它语言中极少出现的原则，选择“の”和“ㅇ”分别作为日文和韩文的特征形状，如图5所示。

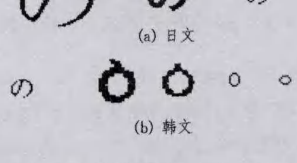


图5 形状特征在不同字体下的差异

对于不同的字体和字号，它们的形状会发生一定的变化，韩文中“○”在某些较大的字体中上方会增加一个突出的点，在较小的字体中其宽高比也存在一定的差异，而日文特征形状的差别较小，仅是下方的开口大小发生了一点改变。

3 文种识别

3.1 识别特征描述

1)字符复杂度特征的描述

以水平和垂直方向字符复杂度之和大于 8 的字符在文档中所占的比例作为字符复杂度特征的描述。根据训练样本计算垂直与水平方向字符复杂度之和大于 8 的语句在文档中所占的比例 S，在每种语言内计算 S 的均值并将其作为该语言的字符复杂度特征参考值。

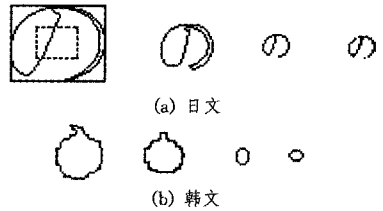
2)版式特征描述

统计各训练样本的句长分布，计算长度大于 10 的句子在文档中所占的比例 L，在每种语言内计算 L 的均值并将其作为该语言的版式特征参考值。

3)形状特征描述

对于形状特征，首先标记文档的连通域，提取各连通域的轮廓，通过轮廓特征间接对其进行描述，其轮廓如图 6 所示。通过圆形度、空心度、垂直和水平方向的边沿数以及近似对称性来描述这两种形状，具体定义如下：

- 圆形度 C：用宽高比和轮廓矩形边框的四角像素分布共同约束；
- 空心度 E：如图 6(a)中虚线框内的像素与总像素之比，定义为空心度；
- 垂直方向的边沿数 Ve：垂直扫描线上黑游程数的最大值；
- 水平方向的边沿数 He：水平扫描线上黑游程数的最大值；
- 轴对称性 S：矩形框内左半部分与右半部分的像素数之比。



(a) 日文

(b) 韩文

图 6 形状特征轮廓

如果某字符连通域满足 C、E、Ve 和 He 的约束条件，就认为是“の”。如果满足上述所有的约束条件，就认为是“○”，其中 Ve 和 He 对日文和韩文的约束条件是不同的。对于训练样本文档，根据上述形状特征描述方法查找匹配形状，统计这两种特征形状的分布情况，计算其在各语言中出现频率的平均值，将其出现频率的均值所构成的二维特征向量作为该语言的高频形状特征识别参考向量。

3.2 判决方法

3.2.1 线性判别分析 LDA(Linear Discriminant Analysis)算法

LDA 算法是广泛应用的降维和分类算法，其目的就是使变换空间中不同类别尽量远离，而类内元素则尽量集中，即类间方差与类内方差之比最大化。对于由 N 个训练样本 $\vec{x}_i$ ($1\leqslant i\leqslant N$)所构成的样本集合，假设集合中共含有 C 类数据，类别 j 中的样本数为 N_j ($1\leqslant j\leqslant C$)，显然有 $N=\sum^{C}N_j$，定义 $\vec{X}$ 为总体样本均值，$\vec{\mu}_i$ ($1\leqslant i\leqslant C$)为第 i 类样本均值，S_B 为类间散布矩阵，S_W 为类内散布矩阵。希望找到变换矩阵 w，使得式(1)取最大值，这样的变换矩阵应满足式(7)，即最终转化为方阵 $S_W^{-1}S_B$ 的本征向量求解问题，变换矩阵 w 由较大的本征值所对应的本征向量构成。

$$J(w)=\frac{w^{\mathrm{T}}S_Bw}{w^{\mathrm{T}}S_ww} \tag{1}$$

$$S_B=\frac{1}{C}\sum_{i=1}^{C}(\vec{\mu}_i-\vec{X})(\vec{\mu}_i-\vec{X})^{\mathrm{T}} \tag{2}$$

$$S_{W_j}=\frac{1}{N_j}\sum_{i=1}^{N_j}(\vec{x}_i-\vec{\mu}_j)(\vec{x}_i-\vec{\mu}_j)^{\mathrm{T}} \tag{3}$$

$$S_W=\frac{1}{N}\sum_{j=1}^{C}N_jS_{W_j} \tag{4}$$

$$\vec{X}=\frac{1}{N}\sum_{i=1}^{N}\vec{x}_i \tag{5}$$

$$\vec{\mu}_j=\frac{1}{N_j}\sum_{i=1}^{N_j}\vec{x}_i \tag{6}$$

$$S_Bw^*=\lambda S_Ww^* \tag{7}$$

3.2.2 模糊集贴近度准则[11]

设论域为 X，B 为待分类对象，A，A_1，A_2，…，A_n 和 B 为论域中的模糊集，定义：

$$N(A,B)=\frac{1}{2}[A\Delta B+(1-A\nabla B)] \tag{8}$$

$$A\Delta B=\frac{1}{n}\sum_{i=1}^{n}[\mu_A(x_i)\wedge\mu_B(x_i)] \tag{9}$$

$$A\nabla B=\frac{1}{n}\sum_{i=1}^{n}[\mu_A(x_i)\vee\mu_B(x_i)] \tag{10}$$

$$N(A_j,B)=\bigvee_i N(A_i,B),\ i=1,2,\cdots,n \tag{11}$$

根据 LDA 算法获得变换矩阵后，将训练数据投影至变换空间，计算各语言样本向量在变换空间中的均值，将其作为该语言的识别参考向量。获得待识别文档的特征向量，利用变换矩阵将其投影至变换空间，根据贴近度准则确定其类别归属。

4 实验结果和分析

从自建的文档图像数据库测试集中取 3 种语言(中文、韩文和日文)的文档图像来验证这种算法的性能。分别使用基于符号匹配的算法和本文算法进行两组实验，以验证算法在不同条件下的识别效果。

4.1 实验 1

实验 1 中采用的训练集和测试集具有相同或相似的字体，训练集中每种语言样本图像各 200 幅，测试集中每种语言样本图像各 100 幅。实验 1 的识别结果见表 1。

表 1 实验 1 识别结果

语言	本文算法	符号匹配法
中文	100	98
韩文	98	99
日文	98	98

4.2 实验 2

实验 2 中训练集和测试集具有不同字体，训练集中每种

语言样本图像各 200 幅，测试集中每种语言样本图像各 100 幅。实验 2 的识别结果见表 2。

表 2　实验 2 识别结果

语言	本文算法	符号匹配法
中文	98	91
韩文	91	69
日文	94	74

4.3　性能分析

基于符号匹配的算法需要将提取的特征与标准模板进行匹配，中文由于整体特征相对明显，受字体影响较小，而韩文、日文对字体的变化比较敏感。本文算法对于特征的描述不需要对字符进行归一化处理，因此对于字体的变化具有较强的适应能力，在目前引入的字体范围内取得了良好的识别效果。

结束语　针对常用的文种识别方法中存在的一些问题，提出一种基于多特征融合的东亚文种识别算法，其通过尺寸无关的高频特征形状匹配，同时在识别过程中结合了版式特征和字符复杂度特征的综合分析，并采用模糊集贴近度准则进行识别，在保证较高识别性能和较小计算量的前提下，对于字体的变化具有良好的适应能力。

参 考 文 献

[1] Pal U, Chaudhuri B B. Identification of different script lines from multi-script document[J]. Image and Vision computing, 2002, 20 (13/14): 945-954

[2] Spitz A L. Determination of the Script and Language Content of Document Images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(3): 235-245

[3] Elgammal A M, Ismail M A. Techniques for language Identification for hybrid Arabic-English Document Images[C]// Proceedings of Sixth International Conference on Document Analysis and Recognition. Seattle, 2001: 1100-1104

[4] Nakayama T, Spitz A L. European Language Determination from Image[C]// Proceeding of the International Conference on Document Analysis and Recognition. Tsukuba, 1993: 159-162

[5] Hochberg J, Kelly P, Thomas T. Automatic Script Identification From Image Using Cluster-Based Templates[J]. IEEE Transations on Pattern Analysis and Machine Intelligence, 1997, 19 (2): 176-181

[6] 顾立娟，平西建，程娟，等. 一种具有旋转鲁棒性的文本图像文种识别方法[J]. 中国图象图形学报，2010，15(6)：879-886

[7] Brush A, Bolse W W, Sridharan S. Texture for Script Identification[J]. IEEE Transations on Pattern Analysis and Machine Intelligence, 2005, 27(11): 1720-1732

[8] Tan T. Rotation Invariant Texture Features and Their Use in Automatic Script Identification[J]. IEEE Transations on Pattern Analysis and Machine Intelligence, 1998, 20(7): 751-756

[9] 曾理，唐远炎，陈廷槐. 基于多尺度小波纹理分析的文字种类自动识别[J]. 计算机学报，2000，23(7)：699-704

[10] 朱华光，平西建，程娟. 基于二元树复数小波变换的文种自动识别[J]. 数据采集与处理，2008，23(6)：766-771

[11] 王刚，靳彦青，刘立柱. 基于模糊隶属度特征和贴近度的徽标识别[J]. 计算机科学，2009，36(1)：184-193

(上接第 265 页)

洁；而推论 1 则是在定理 1 的基础上推导出来的，经过数值仿真实验验证了这两个结论的有效性和可行性。同时，定理 2 证明了本文结论优于文献[11]的结论。

参 考 文 献

[1] Chua L O, Yang L. Cellular neural networks theory[J]. IEEE Transaction on Circuits and Systems, 1988, 35(10): 1257-1272

[2] Chua L O, Yang L. Cellular neural networks applications[J]. IEEE Transaction on Circuits and Systems, 1988, 35(10): 1273-1290

[3] Ke Yun-quan, Miao Chun-fang. Existence analysis of stationary solutions for RTD-balsed cellular neural networks[J]. International Journal of Bifurcation and Chaos, 2010, 20(7): 2123-2136

[4] Huang Zu-da, Peng Le-qun, Xu Min. Anti-periodic solutions for high-order cellular neural networks with time-varying delays [J]. Electronic Journal of Differential Equations, 2010, 59: 1-9

[5] Ban Jung-chao, Cang Chih-hung. On the monotonicity of entropy for multilayer cellular neural networks[J]. International Journal of Bifurcation and Chaos, 2009, 19(11): 3657-3670

[6] Peng Jun, Zhang Du, Liao Xiao-feng. A Digital image encryption algorithm based on hyper-chaotic cellular neural network[J]. Fundamenta Informaticae, 2009, 90: 269-282

[7] Chen Li-ping, Wu Ran-chao. Exponential stability of stochastic fuzzy cellular neural networks with distributed delays[J]. International Journal of Bifurcation and Chaos, 2009, 19(10): 3387-3395

[8] Liu Jin-zhu, Min Le-quan. Robust designs for templates of directional extraction cellular neural network with applications[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(1): 87-100

[9] 廖晓峰，李学明，周尚波. 基于 LMI 方法的时滞细胞神经网络稳定性分析[J]. 计算机学报，2004，27(3)：377-381

[10] 贺勤斌. 一类 CNN 细胞神经网络的稳定性[J]. 科学技术与工程，2008，8(24)：3765-3769

[11] 廖晓昕. 细胞神经网络的数学理论(I) [J]. 中国科学(A 辑)，1994，24(9)：902-910

[12] YU Wen-wu. A LMI-based approach to global asymptotic stability of neural networks with time varying delays[J]. Nonlinear Dynamics, 2007, 48: 165-174

[13] 王曦，欧阳城添，张小红，等. CNN 的全局渐近稳定性分析与改进[J]. 计算机应用与软件，2009，26(4)：96-99

[14] 贾伟凤，曾囡莉，廖晓昕. 关于 Hopfield 型神经网络稳定性的注记[J]. 华中科技大学学报：自然科学版，2003，31(8)：74-76

[15] 黄立宏，李雪梅. 细胞神经网络动力学[M]. 北京：科学出版社，2007