文章编号:1673-1549(2014)05-0029-04　　DOI:10.11863/j.suse.2014.05.08

基于 Gabor 特征与支持向量机的字符识别系统研究

文　广[1], 李　莺[2], 李其阳[2]

(1.攀枝花学院电气与信息工程学院, 四川　攀枝花　617000; 2.四川理工学院自动化与电子信息学院, 四川　自贡　643000)

摘　要:利用计算机进行字符自动识别与录入的技术对机器翻译、数据挖掘、人工智能等都有着重要的理论意义和实用价值,基于数字图像处理技术的字符识别是其中的一个重要发展方向。文章重点研究了字符特征提取和匹配识别这两个影响字符识别效果的因素,根据中文字符笔画的方向特点,选择了对图像方向特征敏感的 Gabor 变换作为特征提取方式,在获取字符的特征向量后,先利用最小距离分类器进行预分类,再利用最小距离分类中产生的候选样本集训练 SVM 分类器,识别时只需利用候选集分类器依次判决,降低了训练和识别工作量,同时提高了识别效率。实验表明,系统对网站导航字符平均识别率达 94% 以上,具有一定的理论意义和实用价值。

关键词:字符识别;Gabor 变换;支持向量机(SVM);特征提取

中图分类号:TP391　　**文献标志码**:A

引　言

随着互联网技术的不断发展,从 PC 端到移动端,人类交流信息的载体逐渐从传统的纸质演变到大大小小的屏幕上,而技术的不断发展已经引导人们有了更多的信息获取诉求,如根据文本获取语音信息等[1]。诉求的快速发展要求更加高效、快速地将人类数千年积累的各种信息(文献资料)等迅速准确地转换为各种“云端”上的数字流信息。但是,快速发展时代的特点就是有更加敏捷的需求反馈,不同的“大数据”提供商如何更加准确地实现用户定位和更加完善的用户体验成为了比重日益提高的诉求。

本文选取了较有代表性的一个诉求——各种门户网站如何更加准确地向用户提供有效的数据资源作为应用情景,利用 Gabor 变换和支持向量机分类器作为字符识别技术手段[2],实现网站导航页面的有效文字识别,为后续的数据挖掘提供了有效数据。

1　研究现况

IBM 公司、NCR 公司等都研制有自己的专用 OCR 软件,但仅能识别印刷体的数字、英文字母以及部分符号,符号也只能识别很少的经过特殊设计或指定的字符;富士通、日立、东芝等公司也相继有了自己的 OCR 产品,世界上第一个能够实现新建的自动分拣和手写邮政编码识别的系统就是由东芝公司研制的,两年后 NEC 公司研制了同样的系统,标志着简单的文字识别系统已经有了一定的发展。同时期的另外一件重要的事件就是建立了一些供字符识别技术研究的标准化字符数据库,为后续的发展提供了很好的字符样本,自动分拣率已经达到了 92% 左右,已经比较成熟地运用于邮政系统中了。

我国由于文字的特殊性以及计算技术发展的时间限制,字符识别稍显落后,但近几年也飞速发展,90 年后期手写体的字符识别也有了突飞猛进,脱机、联机的手

收稿日期:2014-06-03

基金项目:四川省教育厅项目(12ZB268)

作者简介:文 广(1973-),男,四川武胜人,副教授,博士,主要从事机械电子和材料成型方面的研究,(E-mail)wenguang1973@126.com

写汉字、数字识别已经有了相当高的识别率和准确率，汉王、方正、清华等公司也推出了不少商业产品，我国的大规模普查工作也是把 OCR 设备作为标准的计算机输入设备[3-5]，使各种社会统计数据更加快速和准确，同时使得我国的大规模数据录入技术成为国际领先水平。

近年来互联网技术的飞速发展又带来了一波新的发展，google 公司、汉王等都在投入财力物力，大力推动研发。

2 字符识别的过程

Gabor 变换理论因其方向敏感的特性，非常适合中文字符的笔画方向特征，所以是一种常用的字符特征提取方法。SVM 近些年来得到了很大的发展，同时有效避免了神经网络学习过程中局部训练达到最小值的缺点，是一种非常出色的模式识别手段，同时经过最小距离分类处理后的候选样本集能够提高 SVM 分类器的训练和识别效率，整个字符识别系统可靠性好、识别率高。

对含有字符的待识别对象经过一定的图像处理操作后提取到单个的字符图像，然后利用 Gabor 变换处理得到每一个字符图像的特征向量，利用与样本库的字符样本相对应的 Gabor 特征向量进行最小距离计算和排序，得到一定的候选字符样本集，然后利用这些字符集的 SVM 分类器进行一对一策略的分类识别，从而实现每个字符的识别。系统流程图如图 1 所示。

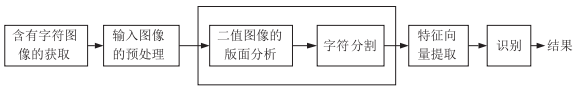


图 1 字符识别过程流程图

3 Gabor 特征提取与 SVM 分类器

3.1 字符的 Gabor 特征提取

Gabor 变换就是通过窗函数来实现将非平稳的信号分解成一系列的短时间内平稳的信号组合[6]，本文使用的是 2DGabor 滤波器，其具有优良的滤波器性能，并有着与生物视觉系统相近的特点，且具有易于调谐的方向和径向频率带宽以及易于调谐的中心频率，在空间和空间频率域同时达到了最佳分辨率，从而实现对字符不同方向的特征提取，数学表达式为[7-8]：

$$h(x,y,\theta,\sigma) = g(x',y')\exp[2\pi j(u_0x + v_0y)] \quad (1)$$

式中：

$$\begin{cases} x' = x\cos(\theta) + y\sin(\theta) \\ y' = -x\sin(\theta) + y\cos(\theta) \end{cases} \quad (2)$$

$g(x',y')$ 是二维的高斯函数，表达式为

$$g(x,y) = \frac{1}{2\pi\lambda\sigma^2}\exp\left[-\frac{1}{2}\left(\frac{(x/\lambda)^2}{\sigma^2} + \frac{y^2}{\sigma^2}\right)\right] \quad (3)$$

图 2 显示了二维 Gabor 基函数的偶分量、奇分量，分别表示字符图像的纹理特征和边缘特征。图 3 是不同的尺度、不同方向的 Gabor 函数图像。

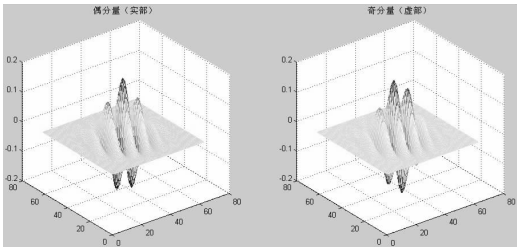
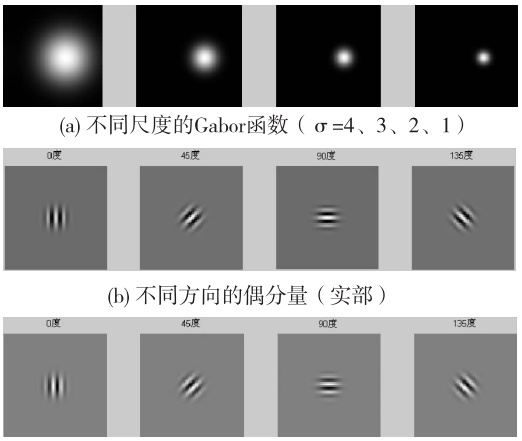


图 2 2D Gabor 函数的实部和虚部



(a) 不同尺度的Gabor函数（ σ=4、3、2、1）

(b) 不同方向的偶分量（实部）

(c) 不同方向的奇分量（虚部）

图 3 不同尺度和方向角的 Gabor 滤波器

图 4 是一个实例特征示例。

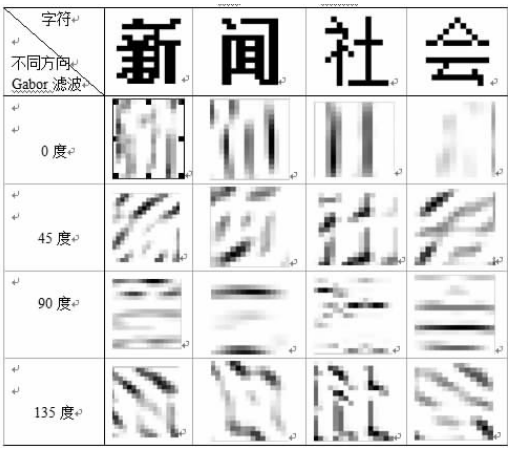


图 4 字符 Gabor 特征示例

3.2 支持向量机(SVM)实现分类器

为避免传统的学习方法中经验风险最小化准则(Empirical Risk Minimization)泛化能力差,本文选用统计决策理论中发展出来的 SVM(支持向量)[9]。在字符样本的特征空间中构造一个最优的超平面,使得不同特征向量之间的距离在字符识别应用中需要的支持向量机是多分类的支持向量机分类,选择采用的是通过将多个二分类器的组合实现,具体构造有 one - against - one(一对一)和 one - against - the rest(一对其余)两种方式[10]。前一种方式是对 n 类训练数据进行两两组合,构建 $c_n^2 = n(n-1)/2$ 个支持向量机,每一个支持向量机训练两类不同的数据,最后分类时采取“投票”方式决定分类结果。后一种方法是对 n 分类问题构建 n 个支持向量机,每个支持向量机分别区分本类数据和非本类数据。此分类器为每个类构造一个支持向量机,第 K 个支持向量机在第 K 类和其余的 $n-1$ 类之间构造一个分类超平面,最后结果输出离分界面距离 $W^T \cdot X + b$ 最大那个支持向量机决定。过程如图 5 所示。

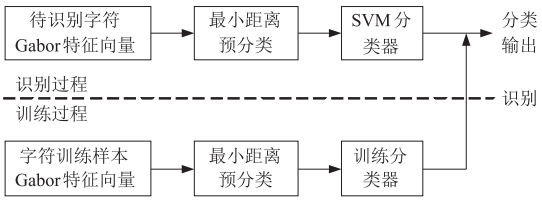


图 5　SVM 的训练和识别过程

本文选择径向基核函数参数 $\sigma^2 = 0.2$ 作为最佳的分类器,分类效果见表 1。

表 1　径向基核函数 $K(x,y) = \exp(-\|x-y\|^2/256\sigma^2)$

σ^2	0.1	0.2	0.5	1.0	1.2	2.0
平均支持向量数	36	33	18	18	17	17
平均识别时间/S	0.436	0.439	0.531	0.485	0.437	0.421
平均识别率/%	99.70	99.74	99.68	99.62	99.65	99.60

4　实　验

本文以新浪网的导航栏作为实验对象,运用 Gabor 变换和 SVM 分类器实现了对导航栏的字符识别,并与目前常用的模版匹配和欧式距离方式识别字符做了对比(表 2)。实验效果图如图 6 所示。

表 2　几种字符识别方法的平均识别率对比

本文方法	模版匹配方法	欧氏距离
113(94.1%)	91(75.9%)	95(79.16%)

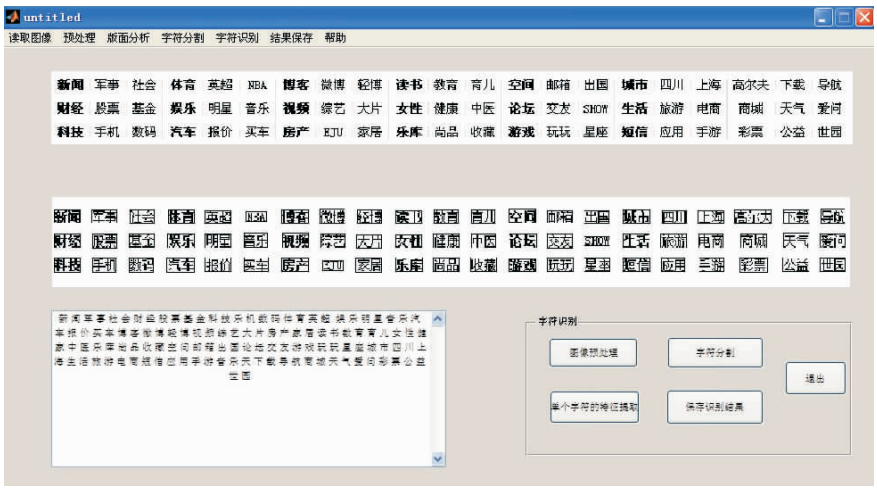


图 6　识别效果图

5　结束语

实验结果表明该方法能够有效地对网站的导航文字实现识别,识别率高达 94% 以上。由于本文的训练样本没有英文字符集,所以主要是针对中文的字符识别,可以看出对中文字符识别是非常成功的,识别结果可以用于后续的网页数据挖掘,分析各个不同门户网站的特点与优劣,基于 Gabor 特征和 SVM 的字符识别技术有较好的理论和实用价值。

参 考 文 献:

[1] 戴 维.基于 Gabor 特征与 SVM 的汉字识别系统的研究与实现[D].上海:上海交通大学,2008.

[2] 徐铭杰.基于支持向量机的字符识别系统的研究与实现[D].杭州:浙江工业大学,2007.

[3] 赵继印,郑蕊蕊,吴宝春,等.脱机手写体汉字识别综述[J].电子学报,2010,38(2):405-415.

[4] 涂岩恺,陈庆虎,黄 亮.手写汉字识别的伪二维弹性网格方法[J].华中科技大学学报:自然科学版,2010,38(11):37-40.

[5] 居 琰,汪同庆,彭 建,等.特征融合用于手写体汉字识别研究[J].电子科技大学学报,2007,31(3):229-233.

[6] 王 林.基于 Gabor 变换的木材表面缺陷识别方法的研究[D].哈尔滨:东北林业大学,2010.

[7] 康俊芳.基于 Gabor 变换的图像特征提取方法研究[D].昆明:云南大学,2010.

[8] 刘习文,薛家详.纸币图像 Gabor 特征提取与识别[J].光学与光电技术,2014,12(3):5-8.

[9] 陆 玉,张 华.基于改进的 BP 神经网络下的字符识别[J]. 韶关学院学报,2014,35(2):20-24.

[10] 张 博.基于 Trace 变换和支持向量机的车牌字符识别新方法[J].微型电脑应用,2014,30(2):9-13.

A Study of Character Recognition System Based on Gabor Feature and Support Vector Machine

WEN Guang[1], *LI Ying*[2], *LI Qiyang*[2]

(1. School of Electricity information engineering, Panzhihua University, Panzhihua 617000, China; 2. School of Automation and Electronic Information, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: The technique that recognizes and inputs characters automatically by using a computer has an important theoretical significance and practical value in many fields, such as machine translation, data mining and artificial intelligence. Character recognition using digital image processing is an important development. The article focuses on character feature extraction and matching identification which influence character recognition effect. According to the property of the Chinese character strokes, Gabor transform which is sensitive to image direction feature is applied to extract the feature. After getting character feature vector, minimum distance classifier presorts first, then candidate sample set produced by minimum distance classifier trains the SVM classifier. Using candidate sample set SVM classifier to judge in turn when recognizing, not only reduces the workload of training and recognition, but also improves recognition effect. Based on this system, the average recognition rate of web navigation characters achieves 94% and higher, the results indicate that this system has certain theoretical and practical value.

Key words: Character recognition; Gabor transform; Support vector machine; feature extraction