文章编号:1001-5132（2007）04-0429-05

认证码字符识别方法的研究

张淑雅，赵一鸣，赵晓宇，李均利

（宁波大学 数字技术与应用软件研究所，浙江 宁波 315211）

摘要：从加强网站安全性的角度，研究认证码字符识别的技术要点. 首先，提出预处理算法去除图像中的干扰因素，分割出字符；其次，采用分类器技术对预处理后得到的字符进行识别. 实验指出有效的预处理算法和选择合适分类器对于快速准确识别认证码起着至关重要的作用，证实了目前一些认证码在保障网站安全性的同时，尚存在一些漏洞，需要继续完善.

关键词：认证码；字符识别；图像预处理；分类器

中图分类号：TP391.43　　文献标识码：A

随着人工智能的发展，模式识别技术得到了人们的广泛重视，它所研究的理论与方法被成功地应用于许多科学和技术领域. 作为其中的一个分支，数字与字符识别是近年来研究的热点，如车辆牌照、手写体数字识别等，在诸多需要对信息进行自动处理的领域都有着极大的理论意义与实用价值.

认证码是现在很多网站通行的方式，在每次访问页面时随机生成，它的识别也受到了许多人的关注. 由于认证码生成程序的不同，认证码图像各种各样，安全级别也不相同，有些生成程序甚至还存在着不少漏洞. 对认证码识别技术的研究，可以及时发现和改善认证码生成程序的漏洞，在加强网站安全性，防止恶意程序的攻击方面有着重要意义.

本文研究了认证码字符自动识别的技术要点. 首先，针对字符背景、噪声和边框等不同的干扰因素，提出了认证码图像的预处理算法. 其次，采用了目前较为广泛使用的几种分类器对预处理后得到的字符进行识别，并对实验结果进行了比较.

1　实验流程与图像来源

图 1 给出了本实验认证码字符识别总流程.

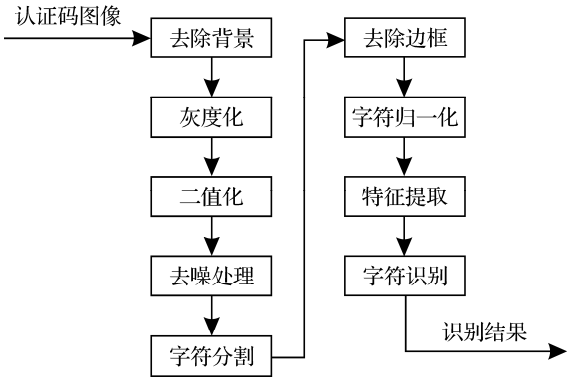

图 1　认证码识别总流程

鉴于某一特定用户一般都是针对某一具体网站进行攻击和认证码图像的多种多样，本文主要选取了一种有代表性的认证码图像类型来说明问题，这些认证码图像来源于我们的 BBS 网站新用户注册页面，图像由一段 PHP 代码随机生成，该段代码核心取自 SMTH_BBS 系统的认证码生成程序

收稿日期：2007-03-12.　　宁波大学学报（理工版）网址：http://3xb.nbu.edu.cn
基金项目：国家自然科学基金（60302012）；浙江省自然科学基金（Y106505）.
作者简介：张淑雅（1982－），女，浙江湖州人，在读硕士研究生，主要研究方向：图形图像与虚拟现实技术. E-mail: zhang_shu_ya@yahoo.com.cn

(SMTH_BBS 是水木清华 BBS 开发组依据 GPL 协议发布的 BBS 系统源代码). 这里，认证码生成程序能生成不同的背景信息，每幅图像含 4 个字符，分别由数字 1～9 和英文字母 A～Z(除 D、O、U、V 这些人们视觉上易混淆的字符外的 31 个字符)组成，字符周围可能有噪声和边框干扰.

2 图像预处理

2.1 去除背景

认证码图像中一般都加入了背景信息. 经观察，图像的背景种类有限. 我们采用统计学方法得到背景图像，即收集一系列背景相同的图像，对这些图像同一位置的像素值进行统计，在该位置取大多数图像都对应的像素值，从而得到完整的背景图像. 将待识别图像与背景图像比较，背景部分置为白色，其余部分保持不变，达到去除背景的目的.

2.2 灰度化

设 g 为某点处灰度值，在 RGB 颜色空间下，图像的灰度化公式为:

$$g = 0.299R + 0.587G + 0.114B. \quad (1)$$

2.3 二值化

由于在带框字符中，存在黑底浅色字符的情况，如图 2 所示，不能简单地通过设置阈值来达到二值化的目的，因此本文采用了Otsu方法[1]. 对于图像 $I(x, y)$，前景(目标)和背景的分割阈值记做 T，属于前景的像素点数占整幅图像的比例记为 ω_0，其平均灰度 μ_0；背景像素点数占整幅图像的比例为 ω_1，其平均灰度为 μ_1，图像的总平均灰度记为 μ，类间方差为：

$$g = \omega_0(\mu_0 - \mu)^2 + \omega_1(\mu_1 - \mu)^2, \quad (2)$$

当 g 达到最大值的阈值 T，即为所求.

2.4 去噪处理

噪声通常是一些孤立的黑色点，这里采用的算法是扫描整个图像，发现 1 个黑色点的时候，考察和该黑色点间接或直接相连接的黑色点的个数，如果大于一定的值，则说明该点不是噪声点，否则将它作为噪声点去掉.

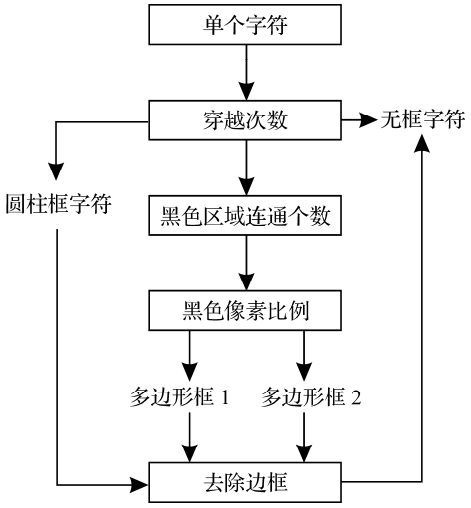

图 2 去除边框流程

2.5 字符分割

分割采用的方法是字符向水平方向投影，投影图的波谷代表字符的分解位置，通过分解位置可以把单个字符分割出来. 类似地，对单个字符在垂直方向上投影，可去除多余的空白背景部分.

2.6 去除边框

在我们所用的认证码带框图像中，分为多边形框和圆柱框 2 种，其中，多边形框分为黑框白底黑字和黑框黑底白字(分别记为框 1 和框 2)，圆柱框为黑框黑底白字.

首先对无框和带框字符进行区分，步骤如下：

(1) 计算穿越次数. 在字符水平或垂直某些位置，如中心处扫描，统计由白像素到黑像素的变化次数，穿越次数为 1 的必为无框字符，而圆柱框字符截至大约 1/3 高度处穿越次数就可达到 3 或 4，这步可区分出圆柱框字符和部分无框字符.

(2) 计算黑色像素的连通区域个数. 无框字符是白底黑字的图像，一般只有 1 个黑色连通区域，而框 1 的连通区域个数必大于 1，框 2 的连通区域个数可能为 1 个或多个，转到步骤(3).

(3) 对于黑色像素的连通区域个数为 1 的字符(无框字符和部分框 2 字符)和黑色像素的连通区域

个数大于 1 的字符(框 1 字符和部分框 2 字符)，分别计算黑色像素比例来区分，这是由于框 2 字符黑色像素大大多于其他字符. 其次，去除边框. 对于框 1，扫描字符图像，将遇到的第 1 个黑色连通区域置为白色即可；对于框 2，将遇到的第 1 个白色连通区域置为黑色，然后对图像反色即可；对于圆柱框，可重复采用类似步骤，这里不再赘述. 图 2 所示为去除边框流程图.

2.7　字符归一化

单个字符图像需要进行归一化处理以消除字体，字号等因素带来的字符在尺寸和位置上的变化，本文将字符图像归一化为 10×18 像素的图像.

3　基于不同分类器技术的字符识别

分类器技术是模式识别系统中的一个关键环节，本文主要采用了目前较为广泛使用的 K 近邻分类器、BP 神经网络分类器和支持向量机分类器来识别字符，并对实验结果作了一定比较.

3.1　K 近邻分类器(K Nearest Neighbor)

KNN法[2]的思路非常简单直观：如果一个样本在特征空间中的 K 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别. 具体可描述为在 N 个已知类别表示的样本中，找出未知向量 x 的 k 个近邻. 设这 N 个样本中，来自 ω_c 类的样本有 N_c 个，若 $k_1, k_2, \cdots, k_c$ 分别是 k 个近邻中属于 $\omega_1, \omega_2, \cdots, \omega_3$ 的样本数，则可定义判别函数为：

$$g_i(x) = k_i, i = 1, 2, \cdots, c. \quad (3)$$

决策规则：若 $g_j(x) = \max_i(k_i)$，则决策 $x \in w_j$. 此类决策在 k 趋向于无穷，样本数 N 趋向于无穷时，错误概率将趋向于贝叶斯错误率. 当 $K=1$ 时，K 近邻规则称为最近邻规则，这里不再赘述.

3.2　BP 神经网络分类器(Back-Propagation)

神经网络以其独特的优点在模式识别中得到了广泛的应用[3]. BP神经网络是一种有导师的学习算法，算法的基本思想是，学习过程由信号的正向传播与误差的反向传播 2 个过程组成. 传播中权值不断调整的过程也就是网络的学习训练过程. 此循环过程一直进行到网络输出的误差减少到可接受的程度，或进行到预先设定的学习次数为止.

对于隐含层神经元的数目，应根据情况选择，数目较少时，网络每次学习的时间较短，但可能因为学习时间不足导致网络无法记住全部学习样本的信息，使权值无法达到全局最小. 数目越多网络识别越精确，训练时间也越长，但取太多会造成网络存储容量过大，也会导致网络对未知输入的归纳能力下降，降低网络的抗噪能力，识别率急剧下降.

3.3　支持向量机分类器(Support Vector Machine)

SVM是建立在统计学习理论基础上的学习方法[4]. 基本思想是通过一个非线性映射，将输入数据映射到一个高维内积空间，并在这一高维特征空间中进行分类. 同时，通过使用核函数，使所有必要的计算都在输入空间中进行. 常用的核函数有：

线性核函数：$K(x, y) = x \cdot y$. (4)

多项式核函数：$K(x, y) = ((x \cdot y) + 1)^d$. (5)

径向基函数：$K(x, y) = \exp(-\sigma \| x - y \|^2)$. (6)

Sigmoid 函数：$K(x, y) = \tanh(k(x \cdot y) - \mu)$. (7)

本文在试验中利用多个 SVM 二分类组合的 one-against-others 算法，将多类识别问题转为二类识别问题来解决.

4　实验与分析

4.1　预处理结果

有效的预处理结果直接影响了后期字符的识别. 图 3 给出了 1 幅原图像及处理后的实验结果，图 4 给出了不同边框类型的字符及处理后结果.

实验表明，对于待识别图像，有效的预处理基本上能正确完整地将单个字符分割出来. 部分边框本身或者在二值化后出现了断裂现象，故而在用穿越次数判断边框类型时，可能会产生一定误差.

对于预处理完毕后的字符图像，为提高识别速

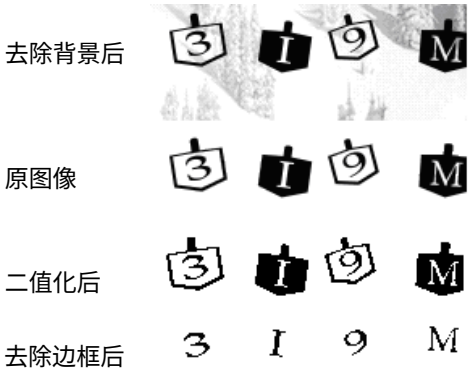

图 3 某图像及处理后结果

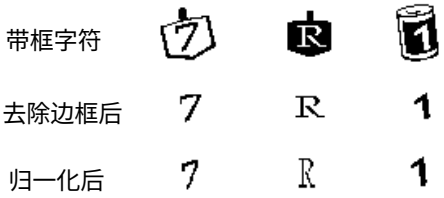

图 4 不同边框类型的字符及处理后结果

度，降低识别系统的复杂度，输入字符不再另外进行特征提取，将 10×18 的字符矩阵转换成 180 维的向量，作为每个分类器的输入.

4.2 K 近邻分类器识别结果

随机选取了一定数量的已知字符样本，其中每种字符样本的个数并不相同，另外再对每种字符各选取 20 个(共计 620 个)样本作为测试，表 1 给出了在 $K=1$ 时的识别结果.

表 1 KNN 识别结果

已知样本总数/个	620	1 550	3 100	6 200	7 435
识别时间/s	2.53	4.69	8.17	14.86	17.67
识别率/%	98.23	99.84	100.00	99.84	99.84

这里，识别时间为识别 620 个字符所用时间. 实验表明，在已知字符样本之间数量不平衡的情况下，KNN 方法仍可以取得较好的结果，KNN 方法可以较好地避免样本的不平衡问题. 随着已知样本数量的增加，识别率有所提高，因而该方法比较适用于样本容量比较大的类域的自动分类，而样本容量较小的类域采用这种算法比较容易产生误分. KNN 法不需要训练时间，但不足之处是计算量较大，因为对每一个待识别字符都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点.

4.3 BP 神经网络识别结果

对于 BP 神经网络的构建，我们采用与 SVM 相同的 one-against-the-others 算法对相同的训练集进行实验，测试集与上述 KNN 方法相同. 对于每个二分类器，我们获取了每个字符的 150 个正例，然后随机选取不同类别的字符 150 个作为反例.

BP 神经网络采取 3 层结构，输入层 180 个神经元，输出层 1 个神经元. 表 2 给出了在不同隐含层神经元个数下的识别结果.

表 2 BP 识别结果

神经元数/个	5	10	20	30	40
训练时间/s	2.91	4.14	6.42	8.66	10.74
识别时间/s	1.44	1.48	1.59	1.67	1.80
识别率/%	91.29	93.87	95.48	94.84	95.00

BP 网络使用广泛，已成功应用于不同的复杂而困难的问题，但是存在训练时间长、不易收敛等缺点. 目前，研究人员已经提出了很多改进方法，使用时可根据实际情况来选择.

4.4 SVM 识别结果

实验比较了采用不同核函数的 SVM 算法识别结果(惩罚因子 $c=100$)，见表 3～5.

表 3 线性核函数识别结果

平均 SV 数/个	45
训练时间/s	0.97
识别时间/s	0.31
识别率/%	99.19

表 4 多项式核函数识别结果

d	1	2	3	4	5
平均 SV 数/个	44	58	70	80	90
训练时间/s	1.62	1.52	1.61	1.57	1.47
识别时间/s	0.86	1.24	1.42	1.53	1.77
识别率/%	99.19	99.84	99.84	99.68	99.52

表 5 径向基函数识别结果

σ	0.01	0.02	0.05	0.1	0.2	0.3
平均 SV 数/个	65	85	191	219	223	223
训练时间/s	1.67	1.72	2.25	2.09	2.32	2.40
识别时间/s	1.47	1.60	3.89	4.48	4.32	4.92
识别率/%	99.84	100	100	100	99.68	96.29

实验表明，SVM 法对小样本情况下的自动分类有着较好的分类结果. 选用不同的核函数和参数，识别率会略有不同，需要根据实验来选择.

此外，SVM 算法最终转换为凸二次优化问题，得到的解是全局最优解，BP 网络识别方案得到的解由于可能存在局部最优解，因此 SVM 识别方案解决了 BP 网络方法中无法避免的局部极值问题.

从实验结果可看出，不同分类器各有优缺点，可根据实际情况来选择合适的分类器.

5 总结与讨论

认证码是一项相对简单而实用的技术，为保证登录网站用户的合法性提供了有力的保障. 本文通过研究认证码图像的识别，分析了字符识别的技术要点，取得了较好的实验结果. 实验结果指出有效的图像预处理算法. 选择合适分类器和参数对于快速准确识别认证码字符十分重要，也证实了目前一些认证码生成程序尚存在一些漏洞，需要改善加强.

参考文献：

[1] Otsu N. A threshold selection method from gray-level histogram[J]. IEEE Trans, 1979, 9:62-66.

[2] Cover T, Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[3] 韩力群. 人工神经网络理论、设计及应用[M]. 北京: 化学工业出版社, 2002.

[4] Theodoridis S, Koutroumbas K. Pattern Recognition[M]. 2nd Edition. Atlanta: Elsevier Science Press, 2003.

An Approach to Recognition of Authentication Code

ZHANG Shu-ya, ZHAO Yi-ming, ZHAO Xiao-yu, LI Jun-li

(Institute of DSP and Software Technology, Ningbo University, Ningbo 315211, China)

Abstract: Recognition of authentication code is investigated from the perspective of strengthening website security. A preprocessing algorithm is first developed to remove the disturbance factors in the image and obtain the characters, followed by using some classifiers for recognition purposes. The experimental results suggest that the developed algorithm and selected classifiers are the most critical part in the recognition process. Some loopholes are still identified in the authentication codes in the website protection phase, and hence further investigation into this issue is needed.

Key words: authentication code; character recognition; image preprocessing; classifier

CLC number: TP391. 43　　**Document code:** A

（责任编辑　史小丽）