第 35 卷 第 8 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.8 Computer Engineering April 2009
·安全技术· 文章编号：1000—3428(2009)08—0186—03 文献标识码：A 中图分类号：TP309

高效的验证码识别技术与验证码分类思想

文晓阳，高 能，夏鲁宁，荆继武
(中国科学院研究生院信息安全国家重点实验室，北京 100049)

摘 要：验证码图片是论坛类网站用以阻止自动化程序恶意行为的重要人机区分技术，其设计和实用安全性直接涉及到互联网的正常使用。为研究国内验证码实用安全性，设计实现一种验证码识别算法模型，对国内的论坛验证码类型进行了实验和分析。实验结果表明，实用中的验证码识别率通常在 50%以上，某些甚至达到 100%，难以起到对自动化程序的阻碍作用。从实用角度将验证码分为 4 类，发现最佳的一类基本是空白的，表明验证码实用技术还应做较大改进。

关键词：验证码；自动识别；Internet 安全

Efficient CAPTCHA Recognition Technology and CAPTCHA Classification Idea

WEN Xiao-yang, GAO Neng, XIA Lu-ning, JING Ji-wu
(State Key Laboratory of Information Security, Graduate University of Chinese Academy of Sciences, Beijing 100049)

【Abstract】CAPTCHA image is the major technique used by forums to tell computers and humans apart, so that its security is of importance in internet usage. To learn about domestic CAPTCHA images, this paper proposes and implements a model of CAPTCHA image recognition algorithm, and tests almost all image types used in domestic forums. Experiment results indicate that a large part of CAPTCHA images have a recognition rate above 50%, and some with 100%, both of which can hardly block automatic machines. In terms of its practicality, this paper further classifies them into four categories, only to find the optimal category is blank, which means CAPTCHA image technology needs great enhancements.

【Key words】CAPTCHA; automatic recognition; Internet security

1 概述

1.1 验证码使用现状

验证码是一种进行人机区分的方法，由于技术简单，易实施，传输数据小，因此被各网站特别是论坛性质的网站广泛使用来防止自动化程序(如论坛自动灌水机)进行大批量的恶意行为。此类自动化程序较为流行，所以，在国内排名前 100 名的论坛[1]中有超过 60%的论坛在注册、登录或发帖部分采用验证码技术[2]。验证码的另一个主要应用场所是电子邮件类网站，用来辅助预防和阻止垃圾邮件群发。总体来说，国外在验证码技术的应用场所与应用程度和国内基本相似。

1.2 国内外研究现状

验证码在互联网中的使用已经非常普遍，验证码实际也成为了网站和网民交互的一个特别的方面。鉴于这 2 类原因，国内外都有学者对验证码的设计和识别进行研究。一种基于外部轮廓特征的数字验证码识别方法[3]主要在单个数字图像的处理分析上进行了相应研究，对单个数字图像外部轮廓上、下、左、右 4 个方向进行特征计算来进行识别。此外，一种基于加权模板匹配算法的形变数字验证码识别系统，运用统计学方法，也可以达到较好效果[4]。文献[5]对基于验证码破解的 HTTP 攻击原理与防范进行了一定研究。卡耐基梅隆大学也有研究小组对验证码进行研究，包括验证码的设计和验证码的识别 2 个方面。

1.3 验证码图片特点

验证码图片具有诸多特点，使得验证码识别方法与一般图像例如照片、视频截图等的识别方法有所不同。识别算法设计应注意到验证码有 2 个方面的特点：即图片的格式方面和图片的内容方面。

在图片格式方面，验证码图片分辨率都较低，图高一般在 20 像素左右，图宽一般在 50 像素~100 像素之间。该特点有时会造成相邻字符间距过窄从而难以切分。此外分辨率较小意味着待识别字符本身信息量不大，对匹配算法造成影响。

在图片内容方面，验证码通常被加入各种干扰因素。主要包括：各种背景干扰，噪点像素，字体形变和累叠，字符位置随机及个数不定，反色等情况。这些干扰正是为了降低自动识别算法的识别效率，但也应注意到干扰因素对正常用户的网络使用体验的影响，这 2 方面都将在后文详细阐述。

2 自动识别算法模型

2.1 基本框架

验证码识别算法流程可分为明显的 3 大阶段：前期预处理阶段，中期切分阶段以及后期分析阶段。其中，后期分析阶段包括模板库建立和识别 2 种模式。图 1 展示了包括 3 大阶段的验证码识别算法，同时可看到在后期分析阶段中包括模板库建立和识别计算 2 个相对独立的处理方式。

前期预处理阶段在尽可能完整地保留每个字符体的像素信息下，尽可能多地去除图片中的大量干扰像素信息。本质是一个去噪过程，可以在彩色图片阶段进行彩色去噪，在二值转换阶段进行灰度去噪，在黑白图片阶段进行黑白去噪。

基金项目：信息产业部电子发展基金资助项目“互联网内容发布与管理系统”

作者简介：文晓阳(1983－)，男，硕士，主研方向：信息安全；高 能，讲师；夏鲁宁，博士后；荆继武，教授、博士生导师

收稿日期：2008-10-20 **E-mail：**wenxiaoyang@lois.cn

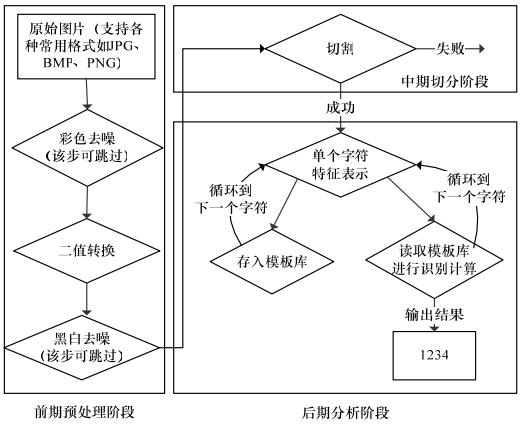

图 1　识别算法流程

本文采用的彩色去噪算法可看作一种滤波计算。当处理像素点(*X*,*Y*)时，计算该像素点及其周围的 8 个像素点(如该点在边缘或者顶角，此个数分别为 5 个和 3 个)的平均 *R*, *G*, *B* 值。然后分别计算这 9 个点到平均值的欧氏距离。再从算的 9 个欧氏距离值中选出最小的。例如该最小距离所对应的像素点坐标为(*X'*,*Y'*)，那么用(*X'*,*Y'*)的 *RGB* 值替代(*X*,*Y*)的 *RGB* 值，此时像素点(*X*,*Y*)完成彩色去噪运算，跳到下一个像素点重复上述步骤，直到遍历验证码整图的所有像素点。

二值转换是算法必经的一步，同时也可以利用二值转换进行去噪。二值转换是指把彩色图片转换成黑白二值图片，即像素点的表示从 *RGB* 三色表示转变为单比特的 0 或 1 二值表示。首先利用公式 $Grey = 0.3\times red + 0.59\times green + 0.11\times blue$ 算出每个像素的灰度值 *Grey*，然后设定或计算一个灰度阈值 *GreyThreshold*。当某点的灰度值小于灰度阈值时，将该点设为 1，表示黑色；反之则设为 0，表示白色。灰度阈值的设定非常重要，假设噪点的灰度值在 150~250 之间，而字符主体像素点的灰度值在 50~100 之间，则易知将灰度阈值设为 100~150 之间，二值化后噪点即可去除(噪点像素点被设为 0，白色)。

黑白去噪对彩色和二值去噪后残留的噪点进行处理。当处理点(*X*,*Y*)时，若该点像素值为 0，则跳过；为 1 时，进行去噪分析，统计该点周围 8 个点(如该点在边缘或顶角，此分别为 5 个和 3 个)中白色点的个数为 *M*，然后设定黑白去噪阈值 *N*，当 *M* 不小于 *N* 时，则认为(*X*,*Y*)为噪点，对其像素值赋值为 0 完成去噪。例如某黑色点周围有 8 个点，设 *N* 为 8，则统计后所得 *M* 为 0 到 7 时不处理，*M* 为 8 时认定其为噪点(易知该点为一个孤立黑点，常见的噪点形式)，则给(*X*,*Y*)赋值 0。

中期切分阶段对前期预处理后的图片进行切割处理，定位和分离出整幅图片中每个字符主体部分。假设某图片长为 50，设为 *X* 轴，高为 10，设为 *Y* 轴。则 *X* 轴每个坐标点对应该点 *Y* 轴上的 10 个点。对 *X* 轴每个坐标点计算其 *Y* 轴对应 10 点中黑点个数，一般会得到如图 2 的情况，波谷波峰波谷的一次转换则对应于一个字符。可设定一个切割阈值 *E*，图 2 情况可设 *E* 为 2，黑点数大于 *E* 的 *X* 轴坐标标记为含字符区域，显然可得 5 个区域，对应于图中 5 个字符。根据这些区域边界定位切割即可，见图 3 效果示例。*Y* 轴上利用相同的处理方式可进一步得到紧密包裹住单个字符主体的更小的图片区域。

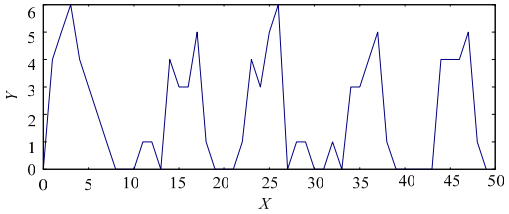

图 2　*X* 轴坐标点各自对应黑点数统计示例

图 3　*X* 轴切割效果示例

得到的比特串将被存入模板库的相应位置。如图 4 所示，若处理的是论坛 U 的验证码，则把该比特串写入:/Modes/U/3.txt。Modes 文件夹为模板库的集合，文件夹 U 就是论坛 U 的验证码模板库，存放着若干记录着特征比特串的 txt 文件。

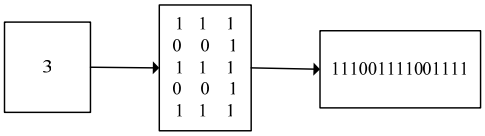

图 4　3×5 矩阵转比特串示例

在后期分析阶段，首先需对切分阶段分离出来的单个字符数据进行特征表示，此后的计算步骤以特征作为运算基本单元。选用比特串作为特征表示方式，主要因为前文提到验证码低分辨率特点(本身信息简单)及识别计算的高速要求。

切分后覆盖每个字符的区域大小不同。如图 3 中的字符 3 长宽可能为 5 和 6 而字符 8 可能为 6 和 8，则需进行标准化运算。采用双线性插值算法[6]把切分得到的字符区域插值到固定的长宽。另有研究表明，长宽皆为 26 是一个最佳的比例[7]。标准化计算后会得到 26×26 的数据结构，每个元素为 0 或者 1，把矩阵元素按序看成比特串(图 4)，即为字符的特征表示。

模板库的大小直接影响到识别速度，识别时间和样本数量负相关，样本量并非越多越好。因此，选择合适的样本量非常重要。可找到一个值 *T_mode*，它是使得识别率达到或接近最高时的所需最少建库样本图片的个数。大量实验发现 *T_mode* 一般在 50 左右。样本量大于该值后，识别率增长一个单位需要更多新增样本的投入，从而造成识别时间较大降低。对自动化程序来说，识别率增长带来的优势不及识别速度减慢带来的劣势。图 5 给出某常见验证码(例：5167 2081)的建库样本数和识别率(识别 135 幅验证码)之间的关系，其他验证码也具有明显的类似关系。

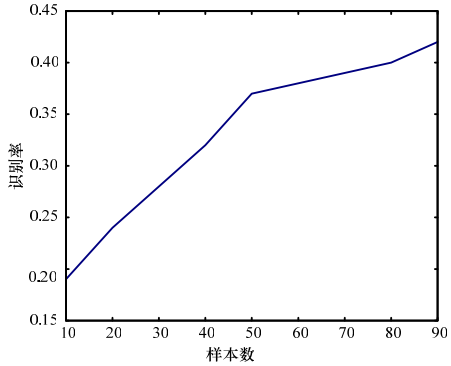

图 5　建库样本数和识别率的关系示例

模板库建成后，利用 KNN(K-Nearest Neighbor)算法[8]进行模板库匹配识别。对待识别验证码的运算过程，在前期预处理及中期切分阶段和对样本图片建库时的完全一致，待识别图片被提取出特征串后利用模板库数据进行 KNN 计算，KNN 中的距离采用汉明距离，字符特征串样本在 *K* 个近邻中最多的为识别结果。

K 值对识别率也有影响。识别较简单验证码时，*K* 取 1 或更大值对识别率无明显影响；识别较为复杂验证码时，*K* 较大时会得到更高的识别率。此外，由于识别算法耗时主要在于遍历模板库，因此与 *T_mode* 不同的是 *K* 对识别速度基本无影响(图例同上，*K* 取 1, 5 和 9，识别 135 幅图片皆为 18.5 s 左右)，可见调节 *K* 值来获取最佳识别率是很好的方法。*K* 并非越大越好，其他条件固定时它会有一个最佳取值。图 6 给出 *K* 值和识别率之间的关系(图例同上，识别 135 幅验证码，样本数为 50 幅)，最佳值应取 5。

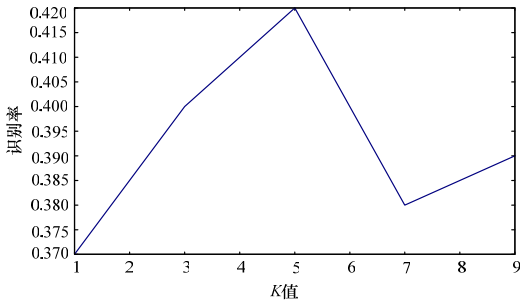

图 6　KNN 取值和识别率的关系示例

2.2　算法优化

在对某种验证码进行建库和识别时，可针对其特点和弱点在算法中的一些环节上进行优化，从而达到最佳识别效果。可能的优化处理大致如下：

背景色去除：某些图片具有特定的背景色彩，且背景色像素点具有相同的 RGB 组合，如这些背景色干扰了预处理，可首先去掉背景色点。背景色点的数量一般是图中最多的，因此统计图中最多的一种 RGB 组合，然后把它们改为白色。

切割优化控制：切割失败指自动切分出的字符数不等于实际字符数，通常是 2 种情况，噪点聚合块被当作字符或者 2 个字符紧靠成一个字符。对前者，可计算切分出来的每个矩形区域的面积，而最小的面积通常为噪点块，将其去掉。对后者，也可通过计算每个矩形区域的面积(或计算区域所占 *X* 轴宽度)，而最大的通常为合二为一的字符体，在该区域进行一次均分。

灰度阈值精细调节：很多验证码图片看似背景复杂噪点众多，但其主体字符灰度值固定在某狭小区间，且与其他部分灰度值不重合，这是重大设计缺陷。只需对二值转换的灰度阈值进行扩展，改为灰度区间，该区间的像素点设为黑色，其余为白，且保证该区间刚好覆盖住主体字符的灰度值，则可一步去掉各种噪点。

其他优化方法还有诸如多次去噪、切割位置定制、模板库样本去重等。

3　识别效果分析

3.1　实验结果

对国内排名前 100 名论坛中所有的验证码(具体有 62 种)进行了识别，目前能有效识别 51 种(识别率 30%以上认为有效)，其中，识别率达到 100%的有 20 种，90%级别的有 9 种，80%级别的有 6 种，70%级别的有 5 种，60%级别的有 4 种，50%级别的有 4 种，30%~50%级别的有 3 种，如图 7 所示。

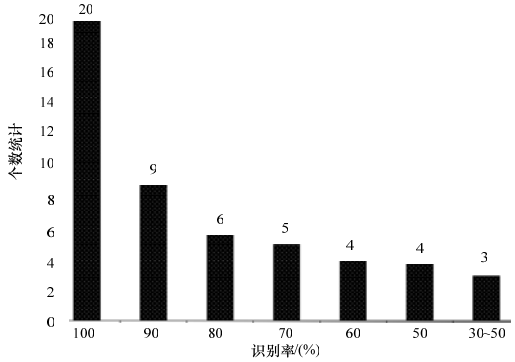

图 7　识别率统计

在识别速度方面，一般来说在 KNN 的 *K* 取值为 1，模板库由 50 幅样本组成的情况下，识别每幅图片在 0.15 s 左右，这等于每分钟可识别 300 幅~400 幅验证码图片。

最后，其他验证码识别方法[3-5]只能处理字符端正、位置固定的验证码，或仅考虑数字为内容的验证码。本文的识别技术能处理复杂的验证码图片(如英文字母混杂，字符位置随机、噪点繁杂、字符形态多样等)，且保持很高识别率。

3.2　验证码分类

文献[2]中简单按照图片内容复杂程度来对验证码分类，这种分类法随着认识的深入显得并不合适，难以对验证码设计提供正确的指导思想。更好的分类方法应基于验证码技术的初衷和实用性，验证码应是在不影响用户正常网络使用体验的前提下尽可能地阻碍自动化程序的攻击。因此按照表 1 进行分类更能体现验证码的优秀程度，且从 IV 类型到 I 类型依次递减。

表 1　验证码分类

类型	样图
I 人难，极易	
II 人易，极易	
III 人难，极难	
IV 人易，极难	暂无特别合适候选者

I 类型是最差的验证码设计，由于各种设计缺陷造成易被自动识别但人眼却较难分辨，造成用户体验不佳。IV 类型为人眼极易辨识而机器难以识别，是最佳的验证码设计，应为验证码设计的指导，国内使用中的验证码暂无此类型。II 类型没有意义，仅会对用户正常使用造成不便。III 类型能起到阻碍自动化程序的作用，本文算法不能有效识别的多为 III 型，给正常使用造成极大不便[9]，适用性值得探讨。国内多数验证码属于 I 和 II 型，皆可被本文的算法准确快速识别。

4　结束语

本文对应用广泛的验证码图片深入研究，设计了高效的验证码自动识别算法，分析了算法中影响识别率的重要参数以及算法优化。利用该算法对百大论坛使用中的验证码进行识别实验，能识别较大部分的验证码，识别率在 30%~100%，这意味着完全无法阻碍自动化识别软件的攻击，它们的存在仅会降低用户的上网体验。而算法难以识别的小部分验证码，基本属于人眼也较难辨识的类型，这也远非优秀的验证码设计。可见国内论坛实用中的验证码存在重大问题，应尽快改进。验证码设计者在设计前必须了解各种识别算法，利用普遍存在的不足，结合文中的 IV 类型作为指导思想进行设计。

(下转第 191 页)