验证码的识别与改进

黄赛平, 许 明

(南京工程学院 计算机学院, 江苏 南京 211167)

[摘要] 针对因特网安全防范中验证码的普遍使用, 讨论了验证码的功能原理、常用的识别方法, 选取了部分网站干扰不同的验证码进行了分割、识别实验. 结果表明: 互联网上有相当多的验证码不能有效地保证系统的安全. 提出一种验证码的改进方法, 该方法程序简单, 产生的新验证码可以提高识别难度和降低破解的准确率, 具有一定的推广价值. 最后给出了验证码设计的建议.

[关键词] 网络安全, 验证码, 识别, 改进

[中图分类号] TP 393 [文献标识码] A [文章编号] 1672-1292(2009)02-0084-05

Recognition and Improvement of Identifying Code

Huang Saiping Xu Ming

(School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

Abstract In allusion to the wide use of identifying codes in internet security and identification authentication, the paper discusses the functional principles of identifying codes and commonly used ways of recognizing identifying codes. Segmentation and recognition experiments of identifying codes are made. The experimental results indicated that the existing identifying codes are not effective in ensuring the system security. An improved method is proposed in this paper. Its program is simple and effective, and the newly generated identifying codes can increase the difficulty of recognition and decrease the accuracy of decipherment. Thus, it has a good value of promotion. Some suggestions on identifying code designing are given in the end of this paper.

Key words network security; identifying code; recognition; improvement

随着互联网技术的发展和应用, 现代社会信息技术日新月异, 伴随而来的就是 Web系统的安全性问题. 为了确保用户提交的请求是在线进行的正常操作, 越来越多的网站采用了验证码技术, 以确保服务器系统的稳定和用户信息的安全[1].

本文讨论了验证码的功能原理及常用的识别方法, 选取有代表性的验证码做了识别实验, 结果表明: 对大多数验证码, 一旦分割问题解决, 破解验证码就变成了一个纯识别的问题, 验证码面临着严峻的安全隐患, 因此部分验证码不能有效地保证系统的安全, 本文提出对验证码进行改进, 以提高识别难度.

1 验证码

当我们注册一个邮件账户或登陆某些网站时, 不仅需要在登录页面上填写用户名和密码, 还经常会被要求输入一张图片中所显示的字符序列, 只有验证成功后才能使用某项功能, 这就是验证码技术[2].

1.1 验证码概念

验证码称为 CAPTCHA(Completely Automated Public Turing Test to Tell Computers and Humans Apart), 是全自动区分计算机和人类图灵测试的简称, 已由卡内基梅隆大学注册成商标[3]. CAPTCHA 的目的是区分计算机和人类的一种程序算法, 这种程序必须能生成并评价人类容易通过但计算机却通不过的测试[4]. 这个要求本身就是悖论, 因为这意味着一个 CAPTCHA必须能生成一个它自己不能通过的测试.

收稿日期: 2009-01-17.
基金项目: 南京工程学院科研基金(KXJ07040)资助项目.
通讯联系人: 黄赛平, 硕士, 讲师, 研究方向: 计算机图像处理. E-mail: huangsp@njit.edu.cn

验证码可分为文本验证码和图片验证码[5]. 在实际运用中文本验证码已被淘汰, 目前主要运用的是图片验证码, 如图 1所示. 图片验证码首先是由服务器随机产生字符序列, 然后与背景图像进行信息融合生成最终的验证码. 图片验证码的安全强度主要是基于图形识别的难度. 一方面, 在信息传输和页面显示中不存在可直接提取的验证码文本, 要进行图像到文本的程序转换, 必须通过图像识别; 另一方面, 针对图像识别技术, 一般在信息融合过程中添加了干扰信息, 同时进行图像混杂、扭曲或变形处理, 增加了图像识别的难度, 用户肉眼容易识别其中的验证码信息, 而攻击者编写的攻击程序, 因为难以识别图片上的字符串, 所以不能顺利地进行自动注册或登录.

图 1 验证码图片示例

Fig.1 A picture example of identifying code

1.2 验证码功能原理

服务器端将随机生成的验证码字符串保存在内存中 (一般是 Web系统中的 session对象), 然后将该字符串写入图片, 发送给浏览器端显示. 在浏览器端, 用户输入图片验证码上的字符串, 然后提交给服务器端, 服务器端将用户提交的字符串和服务器端保存在 session对象中的字符串进行比较, 判断是否一致, 若一致就继续执行下面的代码段, 从而完成后续操作, 否则就无法使用后继功能[6]. 由于验证码是随机产生的, 每次请求都会产生不同的字符串, 攻击者无法从浏览器端提取出验证码信息, 难以猜测其具体内容, 这样就实现了阻挡攻击的目的.

使用验证码可以防止对某些网站进行批量注册, 重复发帖; 防止他人使用广告软件发布大量的垃圾信息; 防止密码被暴力破解, 而且还可以防止对网站的恶意攻击.

2 验证码的识别

在过去的数十年, 研究者们提出了许许多多的识别方法, 这些方法大致可分为 3类: 基于模板匹配的方法[7], 基于字符结构的方法[8]和基于神经网络的方法[9]. 一般来说, 模板匹配法比较简单, 程序实现起来比较容易, 重要的就是把每个字符的模板做好. 为了提高识别率, 根据字符各自的结构特点作为识别特征, 自定义不同的研究算法, 这就是第二种识别方法, 该方法用途比较广泛, 识别速度比较快. 第三种识别方法即神经网络法, 是一种比较先进的方法. 如果训练时间足够长, 训练样本比较合适, 它的识别率相对前两种方法要更高.

2.1 实验情况

由于从网站上获得验证码图像一般是彩色的, 首先对彩色图像进行灰度化, 得到 256色灰度图; 然后对灰度图进行二值化, 去除干扰背景, 保留字符信息; 二值化后的图像上会存在很多噪声, 通过去除干扰噪声; 再把图像上的单个字符分割出来, 最后识别字符并输出结果.

本文选取国内热门网站上干扰不同的验证码做了部分实验, 实验前已经对验证码图片进行了预处理, 字符的识别, 采用图像识别中最简单最常用的识别方法 ——模版匹配法, 模板取为 5×8

实验一 (CSDN 验证码): 该验证码的特点是字体固定, 字符位置不定, 字符无变形, 有部分字符相连, 添加干扰点和干扰线, 字符背景颜色随机变化, 如图 2所示, 采用矩量保持法二值化, 连通去噪法去除背景, 效果如图 3所示.

采用模板匹配法对该验证码进行分割, 如图 4所示.

该类验证码出现的字符为 17个, 每个验证码图片上的字符为 5个, 每个字符按取 20个匹配模板计算, 作为训练样本的验证码至少为 340/5= 68 这里取 80个验证码作为训练样本, 最后得到的匹配模板为

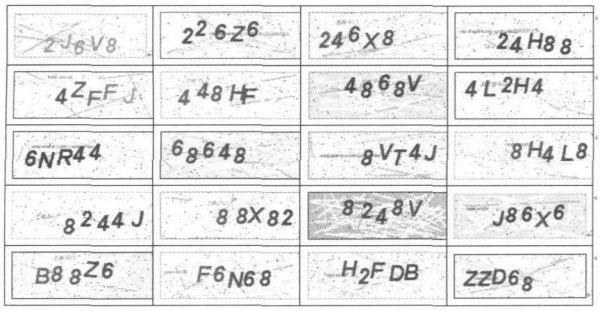

图 2　CSDN 验证码去背景前

Fig.2　CSDN identifying codes before removing background

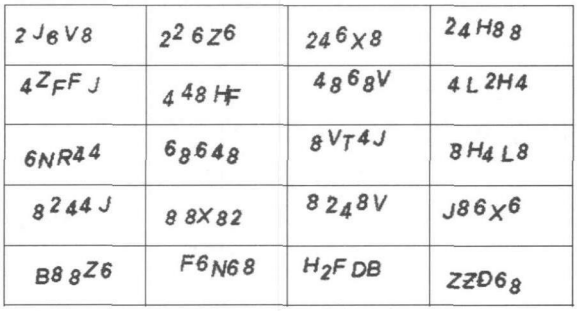

图 3　CSDN 验证码去背景后

Fig.3　CSDN identifying codes after removing background

51 个. 经过对 155 个验证码进行识别,正确率可达 98.7%.

实验二(动网论坛验证码):该验证码的特点是整个图片使用同一种色彩,噪点均匀,字体有一定变形,字符位置基本固定,如图 5 所示,因此采用 p 分位数法二值化,参数为 0.09,效果如图 6 所示.

采用上下边界投影法对验证码进行分割,如图 7 所示.

该类验证码出现的字符为 10 个,每个验证码图片上的字符

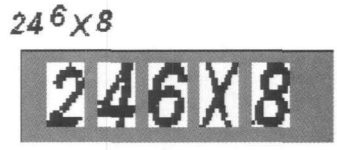

图 4　CSDN 验证码的分割

Fig.4　CSDN identifying code segmentation

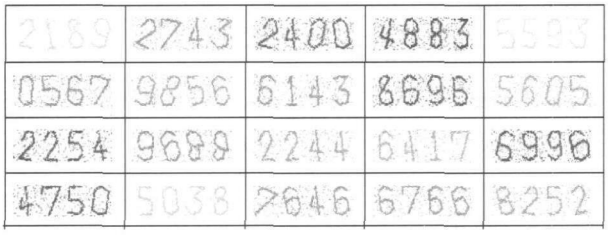

图 5　动网论坛验证码去背景前

Fig.5　DVBBS identifying codes before removing background

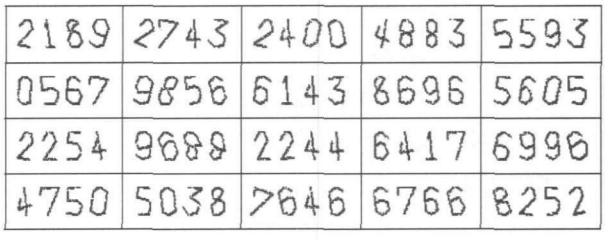

图 6　动网论坛验证码去背景后

Fig.6　DVBBS identifying codes after removing background

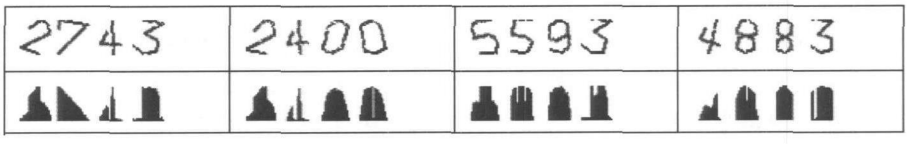

图 7　动网论坛验证码垂直方向上下边界投影

Fig.7　The vertical projections of DVBBS identifying codes

为 4 个,每个字符按取 20 个匹配模板计算,作为训练样本的验证码至少为 200/4 = 50. 这里取 70 个验证码作为训练样本,最后得到的匹配模板为 48 个. 经过对 100 个验证码进行识别,正确率可达 98.5%.

实验三(网易 188 验证码):该验证码的特点是字体固定,字符位置基本固定,字符背景为随机的彩色,如图 8 所示. 该验证码的直方图如图 9 所示. 观察该图发现,取中间靠左边的波谷位置能将背景较好的去除. 所以使用动态阈值法,查找该波谷的二值化位置. 效果如图 10 所示.

采用投影分割的方法对其进行分割. 选取 134 个该类验证码进行识别,正确率达 96.26%.

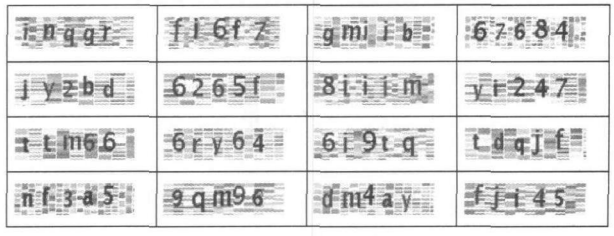

图 8　网易 188 验证码去背景前

Fig.8　Netease188 identifying codes before removing background

2.2　验证码识别总结

从上述实验的识别率及对许多种验证码的观察可知,多数验证码采用了标准印刷字体的数字和字母,字符间没有任何粘连,所以分割字符和抽取特征非常简单. 本文使用简单的分割、识别方法就能达到这么理想的识别率,所以有理由

相信,现在的验证码不具备安全性,一旦分割问题解决,破解它就变成了一个纯识别的问题.

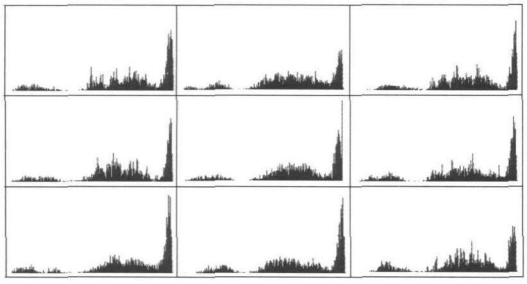

图 9 网易 188 验证码直方图

Fig.9 Histograms of Netease188 identifying codes

inqgr	fi6f7	gmiib	67684
jy2bd	6265f	8iiim	yi247
ttm66	6ry64	6i9tq	tdqjf
nf3a5	9qm96	dm4ay	fji45

图 10 网易 188 验证码去背景后

Fig.10 Netease188 identifying codes after removing background

3 验证码的改进与设计

经过对验证码特点和识别方法的分析,可以得出验证码的设计要把握两个方面:一方面是安全的健壮性,即机器难以识别;另一方面是用户的友好性,即人眼较易识别. 在验证码的识别过程中,识别的难点之一就是字符串的分割. 所以本文提出一种新的通过叠加正弦波的方式来生成变形验证码的方法.

3.1 变形验证码设计

新验证码的生成主要思路是将两个字符串简单叠加在一起,再加以扭曲变形,效果如图 11 所示. 这样得到的字符串难以分割,而用户只要正确输入两个字符串中的一个即可.

但是将两个字符串进行简单叠加,重叠少了容易分割而重叠多了用户友好性不太好,因此将两个字符串分别使用两个正弦波进行变形,如图 12 所示. 这两个正弦波的幅度、波长以及相位都不相同,并且各自的幅度、波长、相位是在一定范围内随机变化的. 因为叠加正弦波的过程类似于调制,为了防止攻击者计算出解调的参数,将参数设置成随机变化.

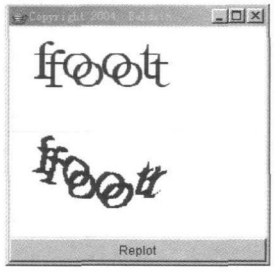

图 11 相同字符串的简单叠加

Fig.11 Simple superposition of the same string

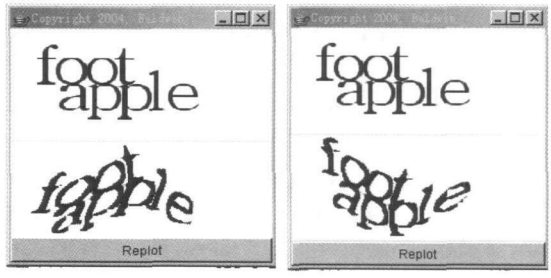

图 12 单方向加两个正弦波并平滑后的效果

Fig.12 Results with smoothing procedures following two one-way-sinusoidal procedures

图 13 给出了对验证码进行改进的程序流程图.

实验证明:通过叠加正弦波设计生成验证码的程序简短,方法简单,提高了破解难度. 对用户而言,尽管这种改进的验证码发生了比较严重的变形,但是人眼还是比较容易识别的,而对机器的识别提出了更高的要求.

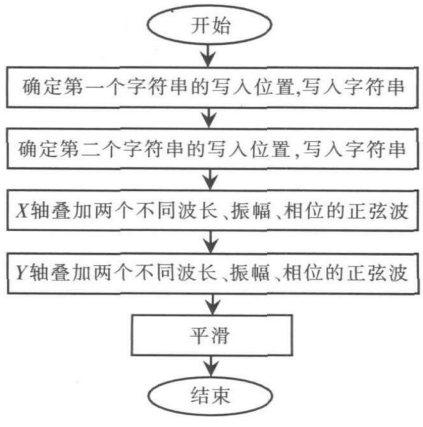

图 13 程序流程图

Fig.13 Program flow chart

3.2 验证码的改进建议

变形严重的验证码虽然提高了破解难度,但对用户的友好性做得不够,为了提高反识别能力,增加用户的友好性,对以后验证码的设计有一些建议:

(1) 在噪音的使用上,尽量把噪音设计得和字符一样,让用来混淆的前景和背景与字符不容易区分.

(2) 将字符的数量、位置、大小设计为随机变化.

(3) 尽量发挥人类擅长而 AI 算法不擅长的,使字符有一定程

度的扭曲、倾斜和粘连,而不是一味的添加复杂的噪音.

(4) 设计非字符型验证码,如图 14所示: 利用人类熟悉的十二生肖的图片来做验证码. 建立图像库,每张图片上都有一个确定的生肖,通过下拉框,让用户选择看到图片上的生肖名称来进行验证.

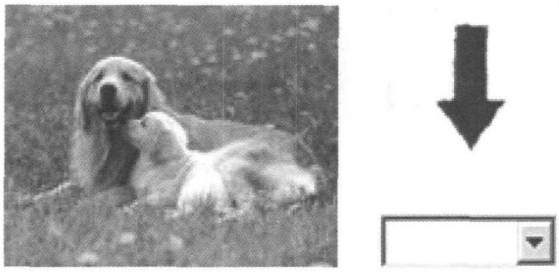

图 14　非字符验证码

Fig.14　Non-character identifying code

4 结　语

为加强互联网应用的安全性,验证码技术得到了广泛的应用,网站使用验证码后,使得用户登录时必须进行验证操作. 虽然登录麻烦了,但对于提高网站的安全性是很有必要的. 好的验证码算法在提高机器识别难度的同时不会增加人类识别的难度. 但是如何把验证码设计得既使人眼容易识别而又难以被破解,还有待于进一步研究.

[参考文献] (References)

[1] 吉治钢. 基于验证码破解的 HTTP攻击原理与防范[J]. 计算机工程, 2006, 32(20): 170-172.
Ji Zhigang. Principles and prevention of HTTP attacks based on identifying code recognizition[J]. Computer Engineering, 2006, 32(20): 170-172. (in Chinese)

[2] 苏磊,马良. 形状上下文在验证码识别中的应用[J]. 人工智能, 2007, 23(2): 252-254.
Su Lei, Ma Liang. Application of shape context in breaking visual CAPTCHA[J]. Artificial Intelligence, 2007, 23(2): 252-254. (in Chinese)

[3] 童情. Captcha[EB/OL]. (2008-06-06)[2009-01-01]. http://baike.baidu.com/view/538168.htm
Tong Qing. Captcha[EB/OL]. (2008-06-06)[2009-01-01]. http://baike.baidu.com/view/538168.htm. (in Chinese)

[4] Luis von Ahn, Manuel Blum, John Langford. dTelling humans and computers apart automatically[J]. Communications of the ACM, 2004, 47(2): 57-60.

[5] 陈珊. 基于 C#验证码的实现[J]. 网络安全技术与应用, 2007(5): 58-59.
Chen Shan. Realization of identifying code based on C#[J]. Network Security Technology and Application, 2007(5): 58-59. (in Chinese)

[6] 洪伟铭. 验证码的原理及实现方法[J]. 武汉科技学院学报, 2007, 20(4): 17-19.
Hong Weiming. Princile and realization of validation code[J]. Journal of Wuhan University of Science and Engineering, 2007, 20(4): 17-19. (in Chinese)

[7] 吴小艳,王维庆,杨春祥, 等. 基于模板匹配的数字图像识别算法[J]. 兵工自动化, 2005(6): 98-101.
Wu Xiaoyan, Wang Weiqing, Yang Chunxiang, et al. Recognition algorithm for digital image based on template match[J]. Ordnance Industry Automation, 2005(6): 98-101. (in Chinese)

[8] 贾婧,葛万成,陈康力, 等. 基于轮廓结构和统计特征的字符识别研究[J]. 沈阳师范大学学报: 自然科学版, 2006, 24(1): 43-46.
Jia Jing, Ge Wancheng, Chen Kangli, et al. Character recognition based on structural and statistical features[J]. Journal of Shenyang Normal University: Natural Science Edition, 2006, 24(1): 43-46. (in Chinese)

[9] 贾少锐,李丽宏,安庆宾, 等. BP神经网络算法在字符识别中的应用[J]. 科技情报开发与经济, 2007, 17(2): 167-169.
Jia Shaorui, Li Lihong, An Qingbin, et al. The application of BP neural network algorithm in the character recognition[J]. SciTech Information Development and Economy, 2007, 17(2): 167-169. (in Chinese)

[责任编辑: 刘　健]