子 图 验 证 码①

Sub Image CAPTCHA

何培舟　温向明　郑　伟　（北京邮电大学 通信网络综合技术研究所 北京 100876）

摘　要：本文提出一种简单的方法用于区分人类用户和计算机程序，称之为子图验证码。在子图验证码中，采用中文随机特征码，通过设置字体、背景加噪、扭曲图像等步骤完成对子图验证码的预处理。之后，以子图形式把子图验证码呈现出来。考虑到计算机程序在识别中文、噪声、粘连字符、扭曲图像、分离图像等方面的缺陷，人类用户很容易被区分出来，子图验证码正是利用这一特点来区分人类用户和计算机程序。最后，子图验证码采用 C#语言并结合 ASP. NET 技术实现。

关键词：CAPTCHA　HIP　OCR　子图

1　引言

随着 Internet 技术的发展，Web 服务变得无处不在，人机交互模式大有取代人人交互模式的趋势。真人互动校对（Human Interactive Proof，HIP）是一组真人用户可以友好交互而计算机程序很难仿真的策略。其中，全自动区分计算机和人类的图灵测试（Completely Automated Public Turing Test to Tell Computers and Human Apart，CAPTCHA）项目[1]是 HIP 中最著名的一类。图灵[2]是第一个研究机器智能性的学者，他通过提供一种测试方法来评定机器是否会思考。CAPTCHA 也称验证码，目前已经出现多种形式，例如 Gimpy，Bongo，PIX，声音，Baffle Text，Pessimal Print 等[3]，这些验证码可以简单地分成三类：文字验证码、图像验证码和声音验证码。验证码应用广泛[3-5]：它可以用于阻止网页蜘蛛（web spiders）和蝇蛆（web bots）参与选举投票，可以阻止暴力攻击，可以阻止网页机器人在博客上添加广告，可以阻止机器人搜索引擎索引私人网页，可以阻止垃圾邮件制造者大量群发送垃圾邮件，可以对数字文档进行鉴定防止假冒，等等。

本文剩余部分的组织结构如下：第二部分介绍了验证码的前期研究，第三部分详细地描述了子图验证码的生成算法。第四部分分析了子图验证码的性能。最后，在第五部分对子图验证码进行了总结。

2　相关研究

Alta Vista[6]是第一个使用验证码技术来阻止滥用自动提交网址信息技术的网站，他们的首席科学家 Andrei Broder 和他的团队在 2001 年取得了该项技术的专利权。

Gimpy 方法由卡内基·梅隆大学提出，用于区分真人用户和计算机程序。Gimpy 验证码生成方法如下：选择特征词，腐化、扭曲、并显示在一个图片中。Yahoo！正是使用了 Gimpy 验证码的一个简单版本——EZ－Gimpy 来阻止广告制造者在聊天室内兜售广告和机器人自动注册免费邮件。[1]

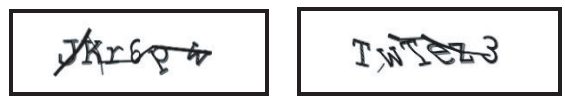

图 1　Yahoo！验证码样例

PIX[3]是另一个著名的验证码，如图 2 所示。PIX 有一个巨大的图片数据库，这些图片来源于日常生活。通过提供给用户一组图片，并要求用户给出这组图片的共同主题来达到区分真人用户和计算机程序的目的。其中，图 2 PIX 验证码的主题是山羊。PIX 的优点是图片只有理解以后才能给出正确的主题，计算机程

① 基金项目：国家自然科学基金（60743007）；北京市教育委员会共建项目专项资助（XK100130648）

序识别起来难度很大。缺点是需要巨大的空间来存储图片,而且图片库需要具有可扩展性,这就要求大量的费用开支。

图 2 PIX 验证码样例

Bongo 验证码[3],如图 3 所示,利用两个图片集,每个图片集都具有一类特性。图 3 中一个图片集显示为粗体,而另一个则是正常线条。系统会呈现给用户一张图片,要求用户指定这张图片属于哪一个图片集。由于可能的方案非常小,Bongo 验证码很容易被暴力猜想攻破。

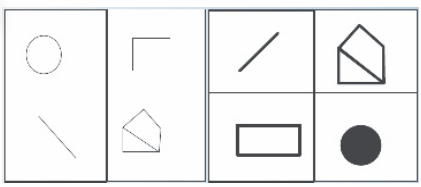

图 3 Bongo 验证码样例

声音验证码可以用于基于语音的服务,也可以作为图像验证码的一个补充提供给那些视觉比较弱的人。在文献[7]中作者提出一种在嘈杂环境中利用识别率合成方式生成声音验证码的方法。

随着验证码的出现,验证码识别技术的研究也在悄然兴起,利用字符切分和扭曲估计已经可以攻破一些相对简单的验证码[8,9]。为了对抗验证码识别技术,文献[10]提出 ScatterType 验证码,它使用分离的机器印刷字体作为特征,来阻止字符分隔攻击。文献[11]中提出画线验证码,这种方法的优点是只需要一个小键盘或者不需要使用键盘,使用日光笔或触摸屏显示。在这种方法中,大量的圆点被显示在屏幕上,其中一些点与其他点有着明显的差别,这种验证码的突破口在于把这些点连接起来,创建一个矩形或者菱形才能通过验证。文献[12]提出人脸识别验证码,它的基本思想是利用真人用户能很轻易地识别人脸而计算机程序却很难识别这样一个事实,通过扭曲人脸,从而达到区分真人用户和计算机程序的目的。文献[13]提出一种交换图像不相重叠区域的机制来区分人类用户和计算机程序,把图像中两个大小、形状完全相同的区域进行交换,人类用户可以很轻易地还原初始图像,计算机程序却很难。文献[14]提出一种联机拼图验证码,它的图片全部来自互联网,利用互联网的图片搜索引擎来更新图片并随机显示,用户根据问题点击图片,只有点击正确才能通过验证。

3 算法描述

本文提出一种区分人类用户和计算机程序的方法,称之为子图验证码(Sub – Image CAPTCHA,SI CAPTCHA)。它的基本思想是利用多图作为验证码,打破传统单图验证码的限制,把生成的验证码图像自动分割成 4 个子图,或者更多个子图,这些子图按固定顺序显示,相互之间都有空隙,由这些子图组合在一起共同构成验证码。子图验证码采用汉字作为随机特征码,之所以选用汉字,是因为汉字比数字和英文字符的字库更大,识别难度更高,更难被 OCR 软件攻破。当然,也可以采用数字、英文字符和汉字的任意组合作为随机特征码,这样生成的子图验证码更难被 OCR 软件识别。子图验证码的实现非常简单,下面以 4 子图中文验证码为例,对其实现过程进行介绍:

(1) 设置子图验证码的长度。验证码长度要适当,长度太短,容易被 OCR 软件攻破;长度太长,又会增加人类用户输入的时间,给人类用户造成麻烦。验证码一般由 4 至 7 个字符组成,最小长度为 4,最大长度为 7。本文默认验证长度为 6,即 6 个汉字字符。

(2) 设置子图验证码字体型号、风格和颜色。字体大小可以根据实际需要进行设置,但字体不能太小。如果字体太小会影响子图验证码的性能,增大 OCR 软件攻破难度的同时,也增大了真人用户识别的难度。字体型号和颜色可以预先设置,也可以随机生成。字体和颜色的多样性可以轻易增加 OCR 软件的识别难度,但对真人用户的影响较小。本例中字体类型默认为宋体,字号大小为 40 像素,字体风格为粗体,字体颜色随机生成,为每个汉字随机生成一种颜色。

（3）设置子图验证码的背景颜色。背景颜色一方面可以突出子图验证码，降低人类用户的识别难度；另一方面能够增大 OCR 软体识别难度，降低被攻破的概率。本例中默认的背景颜色为红珊瑚颜色。

（4）添加随机噪声。添加噪声的目的是模糊验证码内容，增大 OCR 软件识别难度。噪声密度越大，识别难度越大。噪声可以是噪声点，也可以是噪声线。本例中采用噪声点，颜色采用黑色，大小采用像素块，输出个数为 72 个，输出位置随机选择。

（5）进行扭曲处理。人类用户可以很轻易地识别扭曲过的图像，而计算机程序却很难。进行扭曲处理的目的，是在不增加人类用户识别难度的情况下，增加 OCR 软件的识别难度，从而更好地区分人类用户和计算机程序。本例中采用公式 2 正弦曲线来扭曲图像，扭曲后的随机特征码会出现粘连效果。除了使用正弦曲线外，还可以使用余弦曲线、正切曲线、余切曲线、对数曲线，等等。

$$dx = 2\pi i / ImgeHeight \quad (1)$$

$$dy = Asin[B^{*}(dx + C)] \quad (2)$$

公式 1 中 i 表示 x 坐标值，ImageHeight 表示子图验证码的图像高度。公式 2 中，A 表示波形的幅度倍数，A 值越大扭曲程度越高；B 表示波形的相位倍数，B 值越大，扭曲程度越高；C 表示波形的起始相位，取值区间在 $[0,2\pi]$。

（6）生成子图。根据扭曲后的图像来生成子图，可以均匀分割，也可以随机分割，分割的份数越多，识别难度越大。并不是分割的份数越多越好，而是要以不增加人类用户识别难度为前提。本例中把扭曲后的图像从中间均匀分割成四个子图，效果如图 4 所示。6 个汉字随机特征码被分成多个部分。其中，第 3 个汉字被分成 4 个部分，其它汉字均被分成 2 个部分，四个子图组合在一起共同构成一个完整的子图验证码。

（7）输出子图验证码并显示。子图验证码的各个子图要按固定顺序显示，顺序改变后会增加人类用户识别的识别难度，甚至会出现人类用户无法识别的情况。由于各个子图之间有空隙，会出现隔裂的效果，人类用户识别没有问题，但是 OCR 软件就很难了。

根据上述算法，我们采用 C# 语言，结合 ASP. NET 技术在 Microsoft Visual Studio 2005 平台上进行了仿真实现，效果如图 4 所示。

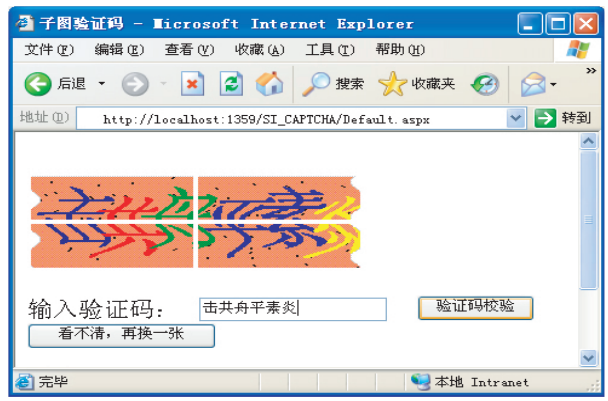

图 4　子图验证码样例

4　性能分析

Gimpy 等传统文字验证码把随机特征码显示在一张图片上呈现给用户，这张生成的图片就是验证码。子图验证码打破了这一思维的限制，把随机特征码显示在多张图片呈现给用户，这些图片共同构成一个验证码。真人用户只要轻轻一瞥就可以很快识别出相互分离的物体，而计算机程序却很难识别，子图验证码正是利用了这一识别上的差别。

子图验证码采用中文作为随机特征码，中文特征码优点是字符集大，计算机程序识别困难。缺点是只能适用于中文环境，而且还可能出现生僻字，给人类用户识别造成困难。1980 年国家颁布了汉字编码的国家标准 GB2312，共包含 6763 个汉字和 682 个其它符号，后来又公布了国家标准 GB18030，对 GB2312 字符集进行了扩充。本文采用了 GB2312 字符集，为了便于用户轻松识别，子图验证码随机生成的汉字应该是人们熟悉的和常用的汉字，应当尽量避免生僻字的出现，而 GB2312 字符集除了包含人们常用的汉字以外，还包含一些不常用的汉字。因此，在随机生成汉字时我们需要根据汉字编码原理对照《汉字区位码表》进行编码，避免生僻字的出现。

在设计验证码时，需要平衡有效抵挡计算机程序攻击和人类用户轻松识别之间的关系。在设计子图验证码时，需要平衡子图个数、子图间空隙大小、噪声密度、噪声大小、字体类型、字号大小、字体颜色、字体间距、背景颜色、扭曲程度等辅助手段和用户识别难度之间的关系。我们采用的方法可以让人类用户轻松通过，而计算机程序却很难攻破。子图验证码属于文本

验证码,此类验证码的共同特点是需要读取验证码图片中的文字,这也给 OCR 软件留下了机会,但相比于其他文本验证码,子图验证码的攻破难度更大,性能更好。

5 结束语

本文提出一种简单的方法用于区分人类用户和计算机程序,即子图验证码。子图验证码利用了计算机程序在识别中文、噪声、粘连字符、扭曲图像、分离图像等方面的弱点,可以很轻易地区分出人类用户和计算机程序,人类用户可以轻松通过验证,而计算机程序却很难攻破。子图验证码在保证网络安全方面能起到积极作用,使用子图验证码相当于为用户登陆设置了一道防火墙,它可以用于阻止计算机程序进行恶意 Internet 注册,可以用于阻止计算机程序自动添加留言和自动发送广告,可以用于阻止计算机程序群发垃圾邮件,可以用于阻止机器人搜索引擎自动索引私人网页,可以用于阻止计算机程序自动投票等。除了以上应用外,子图验证码还可以扩展到 PDA、手机等设备,用于保证 PDA、手机等设备的安全,如何对子图验证码进行扩展是我们下一步工作研究的重点。

参考文献

1 M. Blum, L. von Ahn, J. Langford, The CAPACHA Project, "Completely Automatic Public Turing Test to tell Computer and Humans Apart", www. captcha. net, Dept. of Carnegie - Mellon University, November, 2000.

2 A. Turing. Computing Machinery and Intelligence. Mind, 1950, 59(236):433 - 460.

3 C. Pope, K. Kaur. Is It Human or Computer? Defending E - Commerce with CAPTCHAS. Mind, 2005, 7(2):43 - 49.

4 S. Shirali - Shahreza, A. Movaghar. A New Anti - Spam Protocol Using CAPTCHA. In:IEEE International Conference on Networking, Sensing and Control. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2007. 234 - 238.

5 I. Fisher, T. Herfet. Visual CAPTCHA for Document Authentication. In:IEEE 8th Workshop on Multimedia Signal Processing. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2006. 471 - 474.

6 H. S. Baird, K. Popat. Human Interactive Proofs and Document Image Analysis. In: Proceeding 5th IAPR International Workshop on Document Analysis Systems. London, UK: Springer - Verlag, 2002. 507 - 518.

7 Chan, Tsz - Yan. Using a Text - to - Speech Synthesizer to generate a reverse Turing Test. In: Proceedings of the International Conference on Tools with Artificial Intelligence. Institute of Electrical and Electronics Engineers Inc, 2003. 226 - 232.

8 Greg Mori, Jitendra Malik. Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2003. I/134 - I/144.

9 Greg Mori, Jitendra Malik. Estimation Techniques in Solving Visual CAPTCHAs. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2004. 1123 - 1128.

10 Henry S. Baird, Michael A. Moll, Sui - Yu Wang, ScatterType: A Legible but Hard - to - Segment CAPTCHA. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2005. 935 - 939.

11 M. Shirali - Shahreza, S. Shirali - Shahreza. Drawing CAPTCHA. In: Proceeding of the 28 International Conference on Information Technology Interfaces. Zagreb, Croatia: University of Zagreb, 2006. 475 - 480.

12 Deapesh Misra, Kris Gaj. Face Recognition CAPTCHAs. In: Proceedings of the Advanced International

(下转第 33 页)

（上接第 25 页）

Conference on Telecommunications and International Conference on Internet and Web Applications and Services. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2006.

13 Wen - Hung Liao. A CAPTCHA Mechanism by Exchanging Image Blocks. In: 18th International Conference on Pattern Recognition. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2006. 1179 - 1183.

14 M. Shirali - Shahreza, S. Shirali - Shahreza. Online Collage CAPTCHA. In: Proceeding of the Eight International Workshop on Image Analysis for Multimedia Interactive Services. Piscataway, NJ, United States: Institute of Electrical and Electronics Engineers Computer Society, 2007. 58 - 58.