Global Journal of Computer Science and Technology
Neural & Artificial Intelligence
Volume 13 Issue 3 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

CAPTCHA: Attacks and Weaknesses against OCR Technology

By Silky Azad & Kiran Jain
Doon Valley Institute of Engineering

Abstract - The basic challenge in designing these obfuscating CAPTCHAs is to make them easy enough that users are not dissuaded from attempting a solution, yet still too difficult to solve using available computer vision algorithms. As Modern technology grows this gap however becomes thinner and thinner. It is possible to enhance the security of an existing text CAPTCHA by system-apically adding noise and distortion, and arranging characters more tightly. These measures, however, would also make the characters harder for humans to recognize, resulting in a higher error rates and higher Network load .This paper presents few of most active attacks on text CAPTCHAs existing today.

Keywords : CAPCHA, human interactive proofs, recaptcha, optical character recog-nition, tessaract, security.

GJCST-D Classification : I.2.7

CAPTCHA: Attacks and Weaknesses against OCR Technology

Silky Azad [α] & Kiran Jain [σ]

Abstract - The basic challenge in designing these obfuscating CAPTCHAs is to make them easy enough that users are not dissuaded from attempting a solution, yet still too difficult to solve using available computer vision algorithms. As Modern technology grows this gap however becomes thinner and thinner. It is possible to enhance the security of an existing text CAPTCHA by system-apically adding noise and distortion, and arranging characters more tightly. These measures, however, would also make the characters harder for humans to recognize, resulting in a higher error rates and higher Network load .This paper presents few of most active attacks on text CAPTCHAs existing today.

Keywords : CAPCHA, human interactive proofs, recaptcha, optical character recog-nition, tessaract, security.

I. Introduction

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart)[1] , also known as Human Interactive Proof (HIP), is an automated Turing test in which both generation of challenges and grading of responses are performed by computer programs. CAPTCHAs are based on Artificial Intelligence (AI) problems that cannot be solved by current computer programs or bots, but are easily solvable by humans.

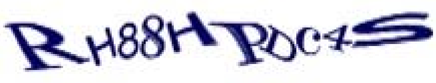


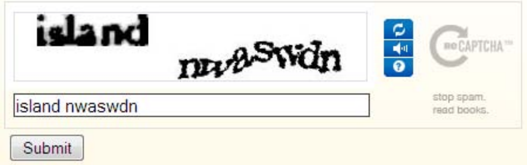
Figure 1 : CAPTCHAs from Internet Source Hotmail.com

The term "CAPTCHA" was first introduced in 2000 by Von an et al., [1] describing a test that can differentiate humans from computers. Under common dentitions, the test must be

1. Easily solved by humans
2. Easily generated and evaluated,
3. But, Not easily solved by computer

Over the past decade, a number of different techniques for generating CAPTCHAs have been developed, each satisfying the properties described above to varying degrees. The most commonly found CAPTCHAs are visual challenges that require the user to identify alphanumeric characters present in an image Obfuscated by some combination of noise and distortion. Figure 1 shows examples of such visual CAPTCHAs.

Another example of an excellent CAPTCHA service is re CAPTCHA[2] , re CAPTCHA is a user-dialogue system originally developed by Luis von An, Ben Maurer, Colin McMillan, David Abraham and Manuel Blum at Carnegie Mellon University's main Pittsburgh campus.



island nwaswdn

It uses the CAPTCHA interface, of asking users to enter words seen in distorted text images onscreen, to help digitize the text of books, while protecting websites from bots attempting to access restricted areas. [1]

a) Applications of CAPTCHAs

CAPTCHAs are used in attempts to prevent automated software from performing actions which degrade the quality of service of a given system. CAPTCHAs are also used to minimize automated postings to various sites.

- Preventing Comment Spam in Blogs.
- Protecting Website Registration.
- Protecting Email Addresses.
- Prevention from Scrapers.
- Online Polls. IP addresses of voters are recorded in order to prevent single users from voting more than once.
- Preventing Dictionary Attacks. CAPTCHAs can also be used to prevent dictionary attacks in password systems.
- Search Engine Bots. It is sometimes desirable to keep WebPages un-indexed to prevent others from finding them easily.
- Preventing Worms and Spam. CAPTCHAs also offer a plausible solution against email worms and spam.

II. Related Work

Marti Motoyama, et al [2] The Authors describes the reverse Turing tests, or CAPTCHAs, have become a

Author α : Department of Computer Science, Doon Valley Institute of Engineering. E-mail : silkyzd15@gmail.com
Author σ : Department of Computer Science, Doon Valley Institute of Engineering. Kurukshetra University, Kurukshetra.

ubiquitous defense used to protect open Web resources from being exploited at scale. An effective CAPTCHA resists existing mechanistic software solving, yet can be solved with high probability by a human being.

They have argued that CAPTCHAs, while traditionally viewed as a technological impediment to an attacker, should more properly be regarded as an economic one, as witnessed by a robust and mature CAPTCHA-solving industry which by passes the underlying technological issue completely.

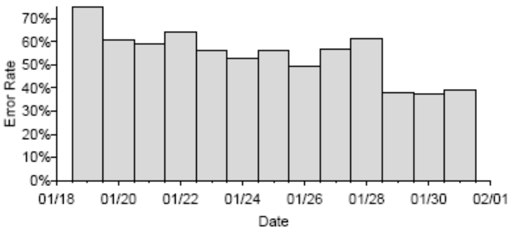


Figure 2 : Image to Text error rate for the custom Asirra CAPTCHA over time [2]

CAPTCHAs are suitable for use with standard solver image APIs. The authors wrote the instructions "Find all cats" in English, Chinese (Simplified), Russian and Hindi across the top, as the majority of the workers speak one of these languages. They submitted this image once every three minutes to all services over 12 days.

Image to Text displayed a remarkable adaptability to this new CAPTCHA type, successfully solving the CAPTCHA on average 39.9% of the time. Figure 2shows the declining error rate for Image to Text; as time progresses, the workers become increasingly adept at solving CAPTCHA.

Elie Burstein, et al, [7] describe that the CAPTCHAs are designed to be easy for humans but hard for machines. However, most recent research has focused only on making them hard for machines. In this paper, they presented what is to the best of their knowledge the first large scale evaluation of CAPTCHAs from the human perspective, with the goal of assessing how much friction CAPTCHAs present to the average user. For the purpose of this study they have asked workers from Amazon's Mechanical Turk and an underground CAPTCHA breaking service to solve more than 318000 CAPTCHAs issued.

In the paper, "Text-based CAPTCHA Strengths and Weaknesses" by Elie Burstein and Mathieu Martin [4], The authors carry out a systematic study of existing visual CAPTCHAs based on distorted characters that are augmented with anti-segmentation techniques.

Applying a systematic evaluation methodology to 15 current CAPTCHA schemes from popular web sites, the authors find that 13 are vulnerable to automated attacks. Based on this evaluation, we identify a series of recommendations for CAPTCHA designers and attackers, and possible future directions for producing more reliable human/computer distinguishers.

In the paper, "Attacks and Design of Image Recognition CAPTCHAs" [6], the authors systematically study the design of image recognition CAPTCHAs (IRCs). They first reviewed and examine all IRCs schemes known to them and evaluate each scheme against the practical requirements in CAPTCHA applications, particularly in large-scale real-life applications such as Gmail and Hotmail. Then the authors present a security analysis of the representative schemes the authors have identified. For the schemes that remain unbroken,

In the paper they presented their novel attacks. For the schemes for which known attacks are available, the authors propose a theoretical explanation why those schemes have failed. The authors have attempted a systematic study of image recognition CAPTCHAs.

The authors provided a thorough review of the state-of-the-art, presented a novel attack on a representative scheme, and analyzed successful attacks on the other representative schemes. Learned from these attacks, the authors defined for the first time a simple but novel framework for guiding the design of robust image recognition CAPTCHAs.

The framework led to their design of *Cortcha*, a novel CAPTCHA that exploits semantic contexts for image object recognition. Their usability study showed that Crotch yielded a slightly better human accuracy rate than Google's text CAPTCHA. Cortcha offers the following novel features.

III. Conclusion

It is possible to enhance the security of an existing text CAPTCHA by systematically adding noise and distortion, and arranging characters more tightly. These measures, however, would also make the characters harder for humans to recognize, resulting in a higher error rate. There is a limit to the distortion and noise that humans can tolerate in a challenge of a text CAPTCHA. Usability is always an important issue in designing a CAPTCHA. With advances of segmentation and Optical Character Recognition (OCR) technologies, the capability gap between humans and bots in recognizing distorted and connected characters becomes increasingly smaller. This trend would likely render text CAPTCHAs eventually ineffective.

In the paper," The Robustness of Google CAPTCHAs" [5], they reported a novel attack on two CAPTCHAs that have been widely deployed on the Internet, one being Google's home design and the other acquired by Google (i.e. re CAPTCHA). With a minor change, their attack program also works well on the latest Re CAPTCHA version, which uses a new defense mechanism that was unknown. When they designed their attack.

This suggests that their attack works in a fundamental level. Their attack appears to be applicable to a whole family of text CAPTCHAs that build on top of the popular segmentation resistant mechanism of "crowding character together" for security. They also proposed a novel framework that guides the application of their well-tested security engineering methodology for evaluating CAPTCHA robustness, and proposed a new general principle for CAPTCHA design.

They concluded that CAPTCHAs are still a new research area. Open problems include the mislabeling problem. Of all the problems they discussed, mislabeling causes the most human errors. The authors may be able to solve this using collaborative filtering, where known human users rate images according to how well they evoke their label. They presented more images per round in the anomaly detection CAPTCHA to debate computer performance.

IV. Future Scope

A lot of work has been done in Enhancing CAPTCHA usability and Security one such example is use of re CAPTCHA[2], However emergence of recent advents and techniques made it more difficult to prevent automated bots and other dangerous spammers against CAPTCHA attacks. some techniques we have discussed in this paper provide more than 40% success rate, and as the faulty CAPTCHA requests are re-evaluated by the server and absence limiting count means that CAPTCHA decryption will be successful in consecutive attacks. In future we would like to use open source OCR Engines to validate such claims.

References Références Referencias

1. Luis Von An, Manuel Blum, Nicholas J. Hopper, and John Langford. 2003. CAPTCHA: using hard AI problems for security. In Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques (EUROCRYPT'03), Eli Biham (Ed.). Springer-Overflag, Berlin, Heidelberg, 294-311.
2. Luis von An, Ben Maurer, Colin McMillan, David Abraham and Manuel Blum (2008). Re CAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321 (5895): 1465–1468.
3. Marti Motoyama, Krill Levchenko, Chris Kanich, Damon McCoy, Geoffrey M. Volker, and Stefan Savage. 2010. Re: CAPTCHAs: understanding CAPTCHA-solving services in an economic context. In *Proceedings of the 19th USENIX conference on Security* (USENIX Security'10). USENIX Association, Berkeley, CA, USA, 28-28.
4. Elie Bursztein, Steven Bet hard, Celine Fabry, John C. Mitchell, and Dan Jurafsky. 2010. How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy* (SP '10). IEEE Computer Society, Washington, DC, USA, 399-413.
5. Elie Bursztein, Mathieu Martin, and John Mitchell. 2011. Text-based CAPTCHA strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security* (CCS '11). ACM, New York, NY, USA
6. Ahmad Salah El Ahmad, Jeff Yan, and Lindsay Marshall. 2010. The robustness of a new CAPTCHA. In *Proceedings of the Third European Workshop on System Security* (EUROSEC '10). ACM, New York, NY, USA.
7. Bin B. Zhu, "Attacks and Design of Image Recognition CAPTCHAs", ACM, Microsoft Research, Temple University, Computer and Information Science, OCT 2010.
8. Marti Motoyama, Krill et al. Re: CAPTCHAs: understanding CAPT-CHA solving services in an economic context. In *Proceedings of the 19th USENIX conference on Security* 28-28. v. 2010.

This page is intentionally left blank