ISSN 1512-1232

SECURED ARCHITECTURE STRATEGY FOR FIGHTING AGAINST BOTS

Shouket Ahmad Kouchay, Abdullah Aljumah,Yasir Ahmad

College of Computer Engineering & Sciences, Salman Bin Abdulaziz University
{sahmad1,ayasir}@ksu.edu.sa

Abstract

CAPTCHA stands for Completely Automated Public Turing Tests to Tell Computers and Humans Apart .A CAPTCHA is a program that protects websites against bots –automated scripts by generating and grading tests that humans can pass but current computer programs cannot. The aim is to allow the server to identify the visitor is a human or a computer, and only provide services to human. It can improve the current server system and user information security.The increase in bots breaking CAPTCHAs shows the ineffectiveness of the text-based CAPTCHAs that are used on most websites and Webmail services today. Bots can easily read the distorted letters and words using optical character recognition (OCR) or break the CAPTCHA using a dictionary attack. The weakness of each CAPTCHA scheme is summarized and accordingly we make an approach to build our CAPTCHA scheme. Considering the case study results and including other points which may pose difficulty for the OCR systems. In this paper we proposed a new technique to build a CAPTCHA which is hybrid (both Picture and Text based with multiple fonts). An image is being rendered on the screen and many text labels of multiple fonts drawn over it. A user has to identify the correct name of the underlying image among the set of text labels that are scattered over it, in order to pass a human verification test. We proposed to use multiple fonts for each letter of a single word inside a Captcha which increases the more complexity of training OCR software.

.

1. Introduction

CAPTCHA is an application level anti-bot defence strategy which differentiates human actions from computer activities, a solution to automated network attacks emerged. A

class of tests for this purpose generally known as Human Interactive Proofs (HIP), which defines a proof that a human being can construct with no special equipment whereas a computer cannot easily create. A way to tell apart a human from a computer by a test is known as Turing Test. When a computer program is able to generate such tests and evaluate the result, it is known as CAPTCHA (Completely Automated Public test to Tell Computers and Humans Apart) [1]

A CAPTCHA ensures that a real person is attempting to log in, and not a computer trying random strings. CAPTCHA is a type of challenge-response test to ensure the user is human. CAPTCHA is a test that can be used to reliably differentiate between human users and automated programs on the web. CAPTCHAs allow us to distinguish legitimate requests from automated requests. The concept of a CAPTCHA is motivated by real world problems faced by internet companies such as Yahoo, Hotmail, Google. These companies offer free email accounts intended for human use only. However, they found that many online vendors were using computer programs known as bots to sign up for thousands of email accounts, from which they could send out masses of spams, junk mails etc. By using CAPTCHA in their services the user is required to solve it to create the email. Through the CAPTCHA these companies are now able to stop the program bots in entering their system[2].

Moni Noar was in 1996 the first person who proposed to use automated Turing tests to verify that a human, rather than a bot, is in the loop [3]. Alta Vista patented a similar idea in 1998 (United States Patent 6195698). However, the term of CAPTCHA was coined in 2000 by a team led by Manuel Blum and Luis von Ahn at Carnegie Mellon University, and the popularity of such

ISSN 1512-1232
technology was largely due to this team's efforts. To date, the most widely used CAPTCHAs are the so-called text-based schemes, in which users are asked to recognize a distorted text, which is intended to be beyond the capabilities of the state of the art of pattern recognition programs[4]. It is widely accepted that a good CAPTCHA must be both robust and usable. The robustness of a CAPTCHA is its strength in resisting adversarial attacks, and this has attracted considerable attention in the research community [2][5][6]. As reported recently, humans are being used to solve CAPTCHAs, either in a well organized manner commercially or by the use of games and other methods whereby humans are unaware that their responses are being used for malicious purposes. These attempts make it futile to make harder AI problems, because a CAPTCHA should be solvable virtually by all humans, regardless by their intention. So, CAPTCHAs are and will remain deployed until an alternate scheme of human verification becomes practical. So, far there are the following three CAPTCHAs[7]:

- Text-based schemes: Typically rely on the distortion of text images which are hard to recognize by the state of the art pattern recognition programs but are easily recognizable by humans.
- Sound-based scheme: Typically requires a user to solve a speech recognition task.
- Image-based schemes: Typically require a user to recognize an image based task.

Some examples of text based CAPTCHAs which have been taken from various popular web service providers are shown in figure 1 through 5.

Figure 1. Text-CAPTCHA Google

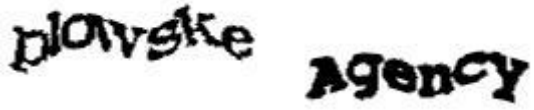


Figure 2. Text-CAPTCHA Hotmail

Figure 3. Text-CAPTCHA Yahoo

Figure 4. Text-CAPTCHA Facebook

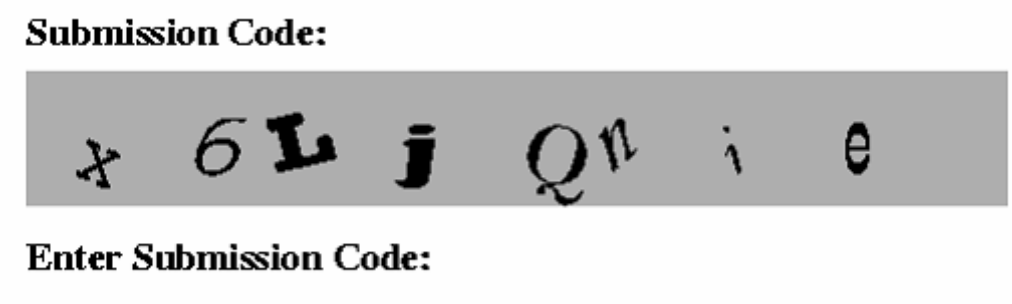


Figure 5. Text-Captcha AltaVista

We propose a new Hybrid CAPTCHA (both Picture and multiple fronted Text) scheme based on the problem of Image recognition. Our key idea is to efficiently use image reorganization with a multiple fronted text that is very easy to answer by human users, but is difficult for automated programs. This test takes advantage of the fact that recognizing image is considered to be a tough task for computers, but is relatively easy for humans. The comfort level of passing these tests is high. The proposed CAPTCHA scheme also provides better protection against spam and has the desirable properties of being easy for humans while being difficult for bots to solve. We choose the

ISSN 1512-1232
text-based scheme and the image-based scheme for the following reasons. First, majority of the websites like Yahoo, MSN are using the text-based CAPTCHAs as shown in figure 1-5 in their websites for human verification purpose. Second, humans find it very easy to read text-based CAPTCHAs. Third, human eyes recognizes the images (animals, fruits, furniture etc) instantly. Program Bots are not able to effectively identify the subject of the underlying pictures and to understand the semantic meaning of the text labels drawn.

2. Related Work

The concept of a CAPTCHA was widely introduced by von Ahn in 2003 [10], hundreds of design variations have appeared. so far, most of them are text-based: The computer generates a challenge by selecting a sequence of letters, rendering them, distorting the image, and adding noise. Text CAPTCHAs are very popular because they are simple and easy to design and implement. Text-based CAPTCHAs seem to suffer from an unfortunate property as, making them hard for computers also makes them hard for humans. This has led some researchers to use images as CAPTCHAs instead. Because general machine vision is a much harder problem than character recognition. Chew and Tygar [4] were among the first to describe using labeled photographs to generate a CAPTCHA. They generated a database of labelled images by feeding a list of easily-illustrated words to Google Image Search [7].

Many CAPTCHA implementations, especially those which have not been designed and reviewed by experts in the fields of security, are prone to common attacks.

2.1. Image Classification CAPTCHAs

Von Ahn et al. entice humans to manually describe images by framing the task as a game. Their “ESP Game” awards points to teams of non-communicating players who can both pick the same label for a random image, encouraging them to use the most obvious label [9]. Their PIX CAPTCHA displays four images from the ESP Game database that have the same label, then challenges the user to guess the label from a menu of 70 possibilities. PIX is clever, but has several potential problems. First, its scale seems insufficient. By solving PIX repeatedly, it is not hard to get repeated images, making the database easy to reconstruct by an attacker. However, perhaps more fundamental, it has a fixed menu of only 70 object classes. This makes it a potential target for brute force attacks (though potentially defensible using our token bucket scheme; see Section 4.2). Even with a large number of categorized images, it may be difficult to add a large number of classes. As the number of classes goes up, so does the number of words that could reasonably be used to describe a set of photos. Finally, PIX photos are sometimes abstract, making it potentially difficult or frustrating as a CAPTCHA. A fascinating use of a large-scale human-generated database is the site HotCaptcha.com. HotCaptcha displays nine photographs of people and asks users to select the three which are “hot.” Its database comes from HotOrNot.com, a popular web site that invites users to post photos of themselves and rate others’ photos as “hot” or “not.” HotCaptcha is clever in its use of a pre-existing motivation for humans to classify photos at a large scale. However, humans may have difficulty solving it because the answers are subjective and culturally relative; beauty has no ground truth. It is also offensive to many people, making it difficult for serious web sites to deploy.

Finally, worthy of mention is the similar-seeming KittenAuth project [11]. Like Asirra, KittenAuth authenticates users by asking them to identify photos of kittens. However, this is a coincidental and superficial similarity. KittenAuth is trivial to defeat because it is has a database of less than 100 manually selected kitten photos. An attacker can (indeed, already has [12]) expose the database by manually solving the KittenAuth challenge a few dozen times. An arbitrary number of challenges can then be solved using an image comparator robust to simple image distortion.

MMC method of CAPTCHA implementation utilizes two existing schemes Text-Based and Image-Based.

ISSN 1512-1232

In this method an image is being rendered on the screen and many text labels drawn over it. A user has to identify the correct name of the underlying image among the set of text labels that are scattered over it, in order to pass a human verification test. [14].Our hybrid captcha helps the user to overcome the drawbacks of the this approach.

2.2. The improved method

The improved method is designed in such a way to increase the resistance of collage CAPTCHA method to hackers attack. In this method the images are displayed on left side and right side of the screen. On the left side of the screen the images consist of different objects like animals, different persons, objects like furniture, flags of countries etc. On the right side of the screen the image consists of the name of the object shown on the left side of the screen. The computer program asks the user to choose the picture on the left side of the screen and then to choose the corresponding image (containing name of the selected object). If the user selected both of them correctly, and after doing this user is allowed to enter the name of the image in the given text box, if user enters the name of the image correctly in the text box ,then it could be guessed that user is human being, not a computer bot. More over the text box remained disabled until user selects both the images correctly then only the text box is enabled and user can enter the name of the image in text box. In this method computer requires four abilities to pass the test

1. To find out the shape of the concerned object
2. To find the concerned object on the screen
3. To find out the object containing the name of the “selected object” on the screen
4. To enter the name of the image in the text box

It is difficult for the computer to realize these tasks in correct order, only a human user can recognize and choose the concerned object. This CAPTCHA program select 8 images (objects) randomly that are different from the previous images. The objects are different from each other and are placed on the left side of the screen. The corresponding images containing the name of the objects already present on the left side appear on the right side of the screen. The user has to select one of the image on the left side of the screen ,say “ a image of chair” and then to select the image containing the word “CHAIR” on the right side of the screen. If user has selected both the objects correctly and also entered the name of the image in the text box correctly, then he is allowed to do the concerned operations[9]. The example of the “chair” is shown in the fig 6.

CHAIR

Text box
(Enter text here)

Figure 6

ISSN 1512-1232

3.1. Our Proposed Method:-Hybrid CAPTCHA Design and Implementation

Hybrid method of CAPTCHA implementation utilizes two existing schemes

- Multi font text Text-Based
- Image-Based

In this method the use of both the schemes are mutually exclusive. We have described in section 2.2 the improved method which also uses both the image and text labels to design the CAPTCHA, but our concept of CAPTCHA implementation is different from the above it renders both the image and multi font text labels together, which increases the more complexity of training OCR software. It helps the user to overcome the drawbacks of the previous approach since it's easy to use and understand. It is one of the innovative methods we have put forth to present the 'recognition of image' as with relevant text label as shown in Fig 7.

CAPTCHA:
(anti-spam code, identify one relevant text label)

Enter the relevant text label here: goat

Submit

Figure 7. Hybrid CAPTCHA

To implement Hybrid CAPTCHA thousands of images (animals, fruits, furniture etc) are collected from the popular search engines like Google, Bing etc. The collected data is publicly available to all humans through these search engines. So, as to maintain the '**P**ublic' feature of the CA**P**TCHAs. A large set of images and text labels are stored in the database. User could easily recognizes the image in our scheme as the images taken are animals, furniture, fruits etc. **Multiple fonts are used for each letter of a single word ..** Our method of CAPTCHA is unique from all the other schemes which exist today. Some of the differences are listed below:

- Image and text labels are exclusively used together.
- Text labels are scattered over the underlying image. (Unique)
- Multiple fonts are used for each letter in a single word which increases the complexity of training OCR software (Unique).

ISSN 1512-1232

There are many methods to break text CAPTCHAs e.g., segmentation, computer-vision, OCR, brute force and dictionary attacks. Breaking our technique is hard enough for these methods because it is still difficult for computer programs to recognize the picture and the text label over it which in our case is written in different font styles and sizes.

3.2. Usability

The comfort level of passing this test for humans is high. Even the people which are suffering from Dyslexia (Seeing disorder) can easily recognize the image and text labels as shown in figure 7.

3.3. Hybrid CAPTCHA - Working

As described in section 3.1 Hybrid CAPTCHA requires that the user type the relevant cursive text to identify the background image to prove he is human. We employed two schemes together to make it hard enough for the computer programs to break the CAPTCHA. An algorithm given below describes the working of our new scheme. Some notations are used as 'L' is final(actual) name of the image, 'T' is any text label with different fonts, N is the image name, 'i' is the subscript variable, 'val' is an integer variable used for human verification.

Algorithm:

a) Read Ti,
b) val = 1;
c) Foreach val to 3
d) Foreach Ti to 4,
e) If Ti is equal to N
L ← Ti;
Goto g;
Else
i ← i + 1;
End Foreach
val ← val + 1
End Foreach
f) Goto h; (Bot)
g) Write L;
h) Stop.

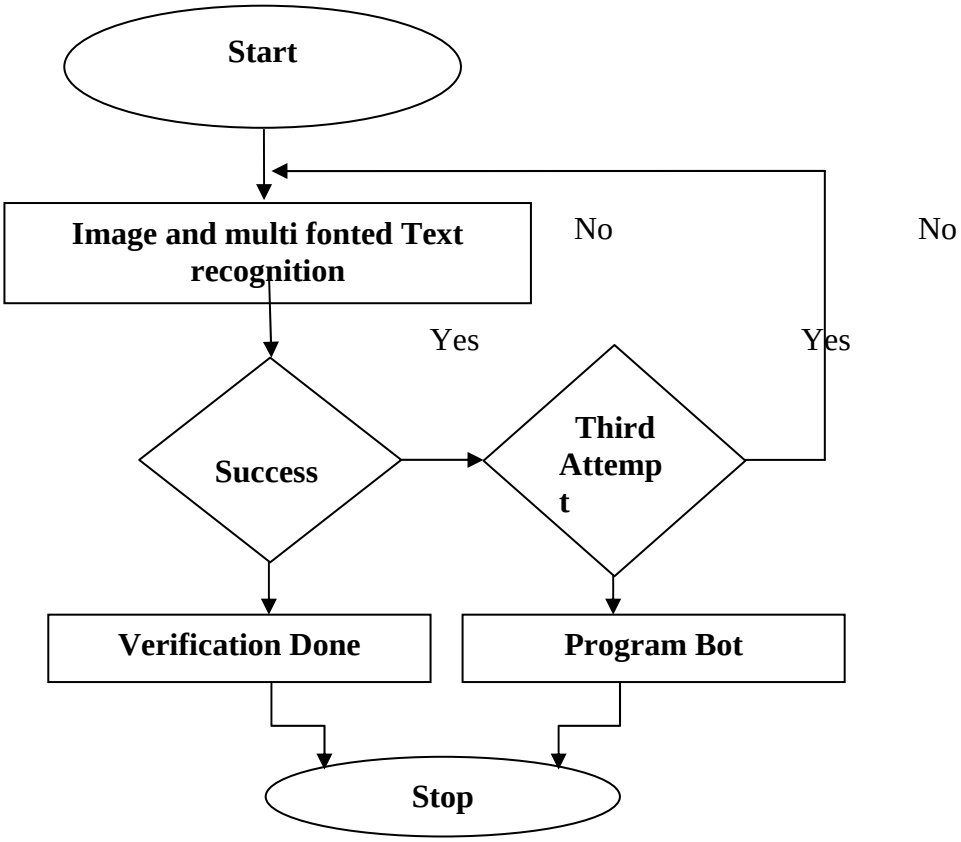
Steps:

A. An image is rendered over the screen (animal, fruits, furniture etc)
B. Four text-labels with different fonts are scattered over the image.
C. Only one label is relevant with the underlying image.
D. A user has to identify the relevant label name and to enter the same in the text box provided.
E. Choosing the right label is the key to verification.

A user has maximum three attempts to recognize the image and to identify the multi fonted text label. If he fails to recognize the image and text label, the underlying system denies any more attempts and treats it as a program Bot.

A program flow diagram of Hybrid CAPTCHA scheme is given below:

ISSN 1512-1232



Program Flow Diagram

3.4. Hybrid CAPTCHA – Application

A. Prevent automated registrations: Hybrid CAPTCHA blocks program bots to register for email, social networking, etc.
B. Prevent spam submission: Hybrid CAPTCHA ensures that an email, comment, blog or social networking message is submitted by a human.

C. Block malicious crawlers: Hybrid CAPTCHA blocks the bots to crawl the websites.

D. Secure login services: Hybrid CAPTCHA prevents bots to login to the secure pages using brute force, dictionary attacks.
E. Prevent online polls: Hybrid CAPTCHA ensures only human users could vote and not the computer programs.

4. Conclusion

CAPTCHA is the method to protect the program bots to access different websites like yahoo, hotmail, Google etc. All these free email service providers used CAPTCHAs in their websites so as to protect program bots to enter into their system. CAPTCHA is an AI problem which only humans can solve current computer programs are not able to solve it. CAPTCHA must be designed as such that the humans could solve it easily but the computer programs could not. We have devised a new technique where we used both the image and multiple fonts of text based CAPTCHAs together to make it difficult for computer programs to break, whereas, humans would find it very easy and interesting to solve this scheme. In our future work we would include the Audio a facility also in our CAPTCHA scheme to make it usable for visually impaired users, and we will consider implementing the text-labels in different languages.CAPTCHA which is hybrid (both Picture and

ISSN 1512-1232
Text based with multiple fonts). An image is being rendered on the screen and many text labels of multiple fonts drawn over it. A user has to identify the correct name of the underlying image among the set of text labels that are scattered over it, in order to pass a human verification test. We proposed to use multiple fonts for each letter of a single word inside a Captcha which increases the more complexity of training OCR software.

References

1. F.Fleuret and D.Geman, "Stationary Features and Cat Detection," J. Machine Learning Research, 9:2549-2578, 20
2. G.Mori and J.Malik. "Recognising Objects in Adversarial Clutter: Breaking a Visual CAPTCHA", IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), Vol. 1, June 2003, pp.134-141.
3. Moni Naor. "Verification of a human in the loop, or Identification via the Turing Test". 1996, online manuscr., avail. at http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.ps.
4. The Robustness of CAPTCHAs: A Security Engineering Perspective [By] J. Yan, A.S. El Ahmad, Newcastle upon Tyne: University of Newcastle upon Tyne: Computing Science, 2009 (University of Newcastle upon Tyne, Computing Science, Technical Report Series, No. CS-TR- 1180).
5. J.Yan and A. S. El Ahmad. "Breaking Visual CAPTCHAs with Naïve Pattern Recognition Algorithms", in Proc. of the 23rd Annual Computer Security Applications Conference (ACSAC'07).FL, USA, Dec 2007. IEEE computer society. pp 279-291.
6. BotBlock. http://www.chimetv.com/tv/products/botblock.shtml.
7. A.Framework to analyze the security of Text based CAPTCHA. Chandavale A.A.; Sapkal A.M.; Jalnekar R.M. Internat. Journal of Computer Applications, vol.1, issue 27, pp. 127-132.
8. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization Published in: Proceeding CCS '07 Proceedings of the 14th ACM conference on Computer and communications security ©2007 table of contents ISBN: 978-1-59593-703-2.
9. Improved Captcha Method: International Journal of Computer Applications © 2010 by IJCA Journal Number 25 - Article 17 Year of Publication: 2010 Authors: Rituraj Soni, Devendra Tiwari.
10. Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using hard AI problems for security. In Eli Biham, editor, Advances in Cryptology – EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4-8, 2003, Proceedings, volume 2656 of Lecture Notes in Computer Science, pp.294–311. Springer, 2003.
11. Oli Warner. Kittenauth. http://www.thepcspy.com/kittenauth.
12. Digg.com user DoubtfulSalmon. http://tinyurl.com/2stwu3, April 2006.
13. http://en.wikipedia.org/wiki/CAPTCHA.
14. A. Almazyad, Y. Ahmad, and S. Kouchay. Multi-modal captcha: A user verification scheme. In Proceedings of International Conference on Information Science and Applications (ICISA), pages 1–7. IEEE, 2011.

Article received: 2013-01-14