

CAPTCHAS: SURVEY OF EXISTING TECHNIQUES AND A NEW APPROACH

Vaishakh B. N.

4th semester Bachelor of Engineering

Department of Computer Science and Engineering

R.V College Of Engineering, Bangalore

Harish G.

Assistant Professor

Abstract - CAPTCHA makes a distinction between humans and computers by employing Turing Tests. This paper discusses about the existing CAPTCHAs. A new type of CAPTCHA design which is different from the existing CAPTCHAs is also discussed.

Index Terms - CAPTCHAs, nuCAPTCHAs, reCAPTCHAs, IMAGINATION, Website specific approach.

I. INTRODUCTION

Colorful images with distorted text in them at the bottom of Web registration forms are called *CAPTCHAs*. CAPTCHA stands for “Completely Automated Public Turing Test to Tell Computers and Humans Apart.” CAPTCHAs are widespread security measures on the World Wide Web that prevent automated programs from abusing online services. They do so by asking humans to perform a task that computers cannot yet perform, such as deciphering distorted characters or identifying images. In 1997, Andrei Broder and his colleagues at the DEC Systems Research Center developed a scheme to block the abusive automatic submission of URLs to the AltaVista web-site. Their approach was to ask each user to read an image of printed text formed specially so that Optical Character Recognition, or OCR systems, could not read it but humans could. Since then, many such CAPTCHAs have been developed. But since then the computer bots have also increased causing a constant threat to the internet. This paper discusses about the presently used CAPTCHAs and also a new type of approach to the CAPTCHAs which is a website specific approach.

II. A BRIEF HISTORY

In November 1999, www slashdot.org conducted a poll to determine the “best” graduate program in Computer Science. The only defense against robotic voting was a check on IP addresses. Groups at Carnegie Mellon University (CMU) and Massachusetts Institute of Technology (MIT) rose to the challenge and automated the process of casting votes for their schools. By the close of the poll MIT had 21156 votes, CMU had 21032 votes, while all other schools had fewer than a thousand votes. This event is often used to motivate traditional entry CAPTCHAs, i.e., a voter must pass an explicit test before casting a vote.

III. A REVIEW OF EXISTING CAPTCHAS

1) reCAPTCHA

reCAPTCHA is a free CAPTCHA service that protects a site against spam, malicious registrations and other forms of attacks where computers try to disguise themselves as a humans. This CAPTCHA comes in the form of a widget that can be easily added to blogs, forums, registration forms, etc. [14]

In addition to protecting one’s site, reCAPTCHA also helps one digitize old books and newspapers. reCAPTCHA improves

the process of digitizing books by sending words, that cannot be read by OCR, to the Web in the form of CAPTCHAs for humans to decrypt. It is mutually beneficial in that the web masters can secure their websites against spammers and get old books digitized. More specifically, each word that cannot be read correctly by OCR is placed as an image and is used as a CAPTCHA. This is possible because most OCR programs alert the user when a word cannot be read correctly.

Each new word that cannot be read correctly by OCR is given to a user in conjunction with another word for which the answer is already known. The user is then asked to read both words. If they solve the one for which the answer is known, the system assumes their answer is correct for the new one also. The system then gives the new image to a number of other people to determine, with higher confidence, whether the original answer was correct. The disadvantage is that there is always a chance for spammers to prove their point of surveillance.

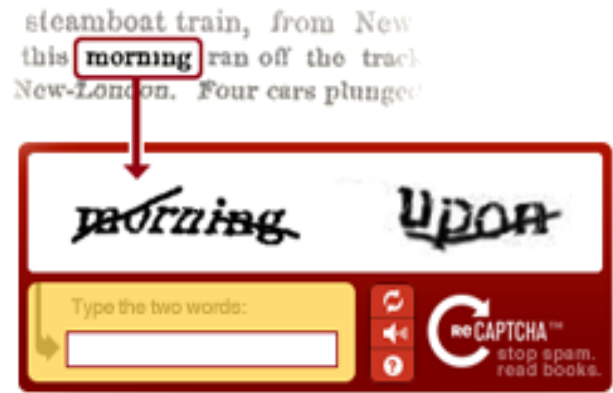


Figure 1: A screen shot of the reCAPTCHA

2) EZ- GIMPY

This is one of the earliest CAPTCHAs used on the Yahoo! website. These are examples of word-based CAPTCHAs. In the EZ-Gimpy program, user is presented with an image of a single word. This image has been distorted, and a cluttered, textured background has been added. The distortion and clutter is sufficient to confuse current OCR (optical character recognition) software. But this kind of CAPTCHA has been broken by Greg Mori and Jitendra Malik of UC Berkeley Computer Vision Group [12].

3) TEXT BASED CAPTCHAs

These tell humans and robots apart by asking a simple logic question [11]. These questions are aimed at a user whose age is equal to or above seven, so they can be solved easily by all but the most cognitively impaired users. Some of the questions that are typically asked are

a) The 7th letter in the word “central” is?

b) Which word contains “z” from the list: zoologist, midwifery, spiderweb, crimps?

But the disadvantages with these text-based CAPTCHAs are that all these questions are stored in a database which also contains the solutions. There is a website called <http://www.wolframalpha.com> [10] which answers questions; does math; instantly gets facts, calculators, unit conversions, and real-time quantitative data and statistics. This website can give approximate answers to the questions asked by the CAPTCHA and the spammers can narrow the answers down.



Figure 2: A screen shot of the EZ- GIMPY CAPTCHA

4) ESP PIX CAPTCHA

This CAPTCHA uses a large database of photographs and images. Instead of typing letters, a user authenticates himself as a human by recognizing what object is common in a set of images. This was the first example of a CAPTCHA based on image recognition.

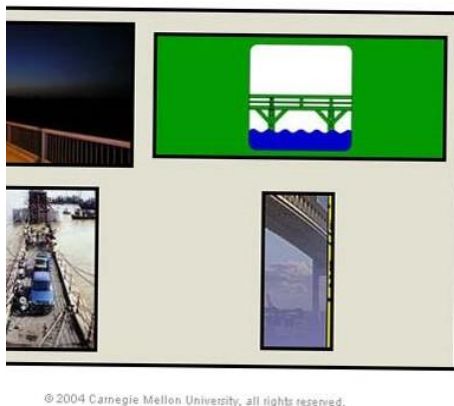


Figure 3: A screen shot of ESP PIX CAPTCHA

5) IMAGINATION CAPTCHA

IMAGINATION (IMAGE Generation for INternet AuthenticaTION) [8] is a system for the generation of attack-resistant, user-friendly and image-based CAPTCHAs. This system produces controlled distortions on randomly chosen images and presents them to the user for annotation from a given list of words. The distortions are performed in a way that satisfies the incongruous requirements of low perceptual degradation and high resistance to attack by content-based image retrieval systems. Results demonstrate the attack-resistance and user-friendliness of this system compared to text-based CAPTCHAs.

6) Nu CAPTCHA

It is a CAPTCHA technique that does not involve twisting, tilting, overlapping, dashing or flying of words to bypass bots. Rather, NuCAPTCHA has a small video in the background and a couple of words flying in the front. As a test to prove that the user

is human, the user is required to enter only the letters that appear in red, which seems to be an easy task since the text is straight and clear. [4] [7]



Figure 4: A screen shot of NuCaptcha

7) FACEBOOK'S SOCIAL CAPTCHA

This is a relatively new announcement made by the social networking site www.facebook.com. In situations where the platform security identifies a possible bot or hack instead of seeing a traditional CAPTCHA of words or characters, the user will see a picture of a friend that they will have to identify.

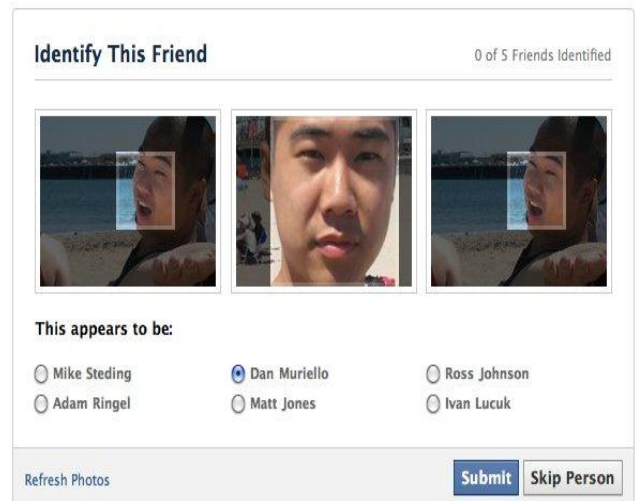


Figure 5: A screen shot of the facebook captcha

But this has seen a lot of criticism from the users of this social networking site. An average Facebook user has around 250 friends. These “friends” are people whom a person just adds when he/she comes across them and may not actively engage with them. Based on the distorted photos or other photos, the user may have trouble identifying some of the “friends”. [6]

8) AUDIO CAPTCHAs:

Typical audio CAPTCHAs consist of one or several speakers saying letters or digits at randomly spaced intervals. A user must correctly identify the digits or characters spoken in the audio file to pass the CAPTCHA. The audio CAPTCHAs of www.digg.com, www.google.com have been hacked with the

accuracy of nearly 60% by a group of four persons from Carnegie Mellon University. [5]

IV. THE PROPOSED TECHNIQUE

Since none of the above CAPTCHAs are foolproof, there are two solutions to this.

1) Website specific approach

This method is an improvement of the Text CAPTCHAs explained earlier. In the traditional text CAPTCHAs, the logical questions asked are the same to all websites and these questions are stored in a database. In the modified technique, instead of asking the user logical questions the CAPTCHA can ask questions related to the content of the site.

Some examples for this are as follows. A website which is related to the Java programming language can ask the following question: "Hello World".substring(3,8). A human will understand this as the substring containing all the characters from the beginning index up to the ending index.

In another website which is based on literature, the following question can be asked, 'Who wrote the epic "Ramayana"?' A number of such examples can be stated. The user with adequate knowledge in that particular field will be able to answer these questions, but not a bot.

Another approach is to use the information provided by the user during the site registration to ask questions. An example will illustrate this as follows. Consider a user who has stated the languages known as Hindi and Kannada, the city they live in as Bangalore and hobbies as Cricket.

In such situations, the user can be asked questions on his city or hobbies rather than the logical questions.

Another approach is to use CAPTCHAs in local languages like Kannada or Tamil and ask the user to decode the same in English. This kind of approach can especially be used for blog comments.

2) Combine 2 or more different types of CAPTCHAs

A combination of the CAPTCHAs in Section can be effective in warding off computer bots. Use a normal reCAPTCHA to show a set of words on the screen. The audio captcha should contain the instructions for the user as to what he is supposed to do with the string. A typical audio message could ask the user to type the second letter of the first word thrice followed by the second word.

V. CONCLUSION

This paper surveys many of the existing CAPTCHAs and proposes a new one. A continuous research in this field is required because the best CAPTCHA of today can be bypassed tomorrow by the spammers. At the same time one should keep in mind that CAPTCHAs are basically for keeping away the spammers and not the humans. Hence the complexity of CAPTCHAs should not exceed a level above which the human feels frustrated.

VI. REFERENCES

[1] Shah N.A., Bandy M.T. *Drag and Drop Image CAPTCHA*. Published online in Working Papers on Information Systems ISSN 1535-6078, 2008. Available at: http://sprouts.aisnet.org/877/1/Sprotus_Paper_Dran_n_Drop_CAPTCHA.pdf

[2] von Ahn L., Blum M. and Langford J. *How lazy cryptographers do AI - telling humans and computers apart automatically*. Communications of the ACM, February 2004, Vol. 47, No. 2.

[3] Baird S.H. and Bentley J. L. *Implicit CAPTCHAs*. In Proceedings of the IS&T/SPIE Document Recognition & Retrieval XII Conference, San Jose, CA, January 16-20, 2005.

[4] Lardinois F. *Video CAPTCHAs Promise Better Security, Less Frustration*. June 30, 2010. Available at:

http://www.readwriteweb.com/archives/nucaptcha_introduces_video_captchas_better_security.php

[5] Tam J., Simsa J., Hyde S. and von Ahn L. *Breaking Audio CAPTCHAs*. Available at:

http://www.captcha.net/Breaking_Audio_CAPTCHAs.pdf

[6] Bains P. *Facebook Captcha?* Blog article posted on January 26, 2011. Available at:

<http://www.nucaptcha.com/captcha/facebook-captcha>

[7] Baweja T. *NuCaptcha: A Motion Based Captcha That Is More Human*. Online article posted on August 06, 2010. Available at:

<http://techie-buzz.com/webmaster-tips/nucaptcha-a-motion-based-captcha-that-is-more-human.html>

[8] Li J. and Wang J.Z., *Automatic Linguistic Indexing of Pictures*, November 1, 2006. Available at: <http://alipr.com/captcha>

[9] *CAPTCHA: Telling Humans and Computers Apart Automatically*, Web article available at: <http://www.captcha.net>

[10] van Horn J. *Using WolframAlpha to Hack Text CAPTCHA*. November 10, 2010. Online article available at:

<http://joelvanhorn.com/2010/11/10/using-wolframalpha-to-hack-text-captcha>

[11] Tuley R., *Accessible Text CAPTCHA Logic Questions*. Online article available at: <http://textcaptcha.com>

[12] Mori G. and Malik J. *Breaking a visual captcha*. Online article available at: <http://www.cs.sfu.ca/~mori/research/gimpy/>

[13] *Some Advantages and Disadvantages to remember about Recaptcha*. Online article available at:

<http://gazolinia.blogspot.com/2009/04/some-interesting-points-to-remember.html>

[14] Abilla P. *reCAPTCHA: A Problem, A Solution, an Innovation*. August 13, 2010. Online article available at:

<http://www.shmula.com/recaptcha-a-problem-a-solution-an-innovation/3419>