IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 6 (Jan. - Feb. 2013), PP 36-42
www.iosrjournals.org

A Review of Bot Protection using CAPTCHA for Web Security

[1]Baljit Singh Saini, [2]Anju Bala
Asst. Professor, Deptt. CSE/IT, Research Scholar Lovely Professional University (Punjab), INDIA
Lovely Professional University (Punjab), INDIA

Abstract: *Today several daily activities such as communication, education, E-commerce, Entertainment and tasks are carried out by using the internet. To perform such web activities users have to register regarding the websites. In registering websites, some intruders write malicious programs that waste the website resources by making automatic false enrolments that are called as bots. These false enrolments may adversely affect the working of websites. So, it becomes necessary to differentiate between human users and Web bots (or computer programs) is known as CAPTCHA. CAPTCHA is based on identifying the distorted text, the color of image, object or the background. This paper examines CAPTCHAs and its working and literature Review. This paper also provides classification of CAPTCHAs, its application areas and guidelines for generating a captcha.*

I. Introduction

CAPTCHA was invented in 2000 at CMU by Luis Von Ahn, Manuel Blum, Nicholas J. Hooper and John Langford. CAPTCHA stands for **Completely Automated Public Turing Test to tell Computers and Humans Apart** [1]. CAPTCHA follows a reverse turing test in which captcha program acts like a judge and participant acts like a user. If the test is passed by the user, then he is considered as human otherwise it is a machine. CAPTCHA is a defensive system that acts as a tool to prevent web bots from abusing online services on the internet including free e-mail providers, wikis, blogs etc. It is a HIP system that is widely used to secure the internet based applications [10]. It is also called as a challenge response test which gives a challenge to the users, when the user gives accurate answer he is considered as human otherwise a web bot.

An HIP system like CAPTCHA is a defensive mechanism to secure the human users from bots in an online environment. There are 3 basic properties that CAPTCHAs must satisfy:

- It should be easy for human users to pass.
- It should be easy for a tester machine to generate and grade.
- It should be hard for a software robot to pass.

II. Literature Review

1. In 1997, Andrei Broder et al. designed a new system for differentiating between human users and computer programs [4]. This method is used by Alta vista website. In this method, a simple distorted English word is presented to the user and then the user is asked to submit it correctly. If a match is found, then he is considered as human otherwise a bot.

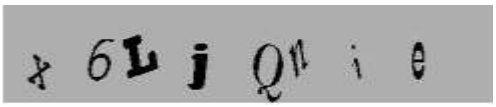

Figure 1: Alta vista Example [4]

2. In 2007, Shirali-Shahreja, M.H. & Shirali-Shahreja, have been proposed Multilingual CAPTCHAs [4]. In this paper a method is based on basis of choice of an object shown on the screen. The user interface of this method is multilingual. At first, the user selects his/her native language. After that, all messages are shown in the selected language. All the messages are translated using an online translator. The advantage of this method is that the user doesn't need to be familiar with English language. In this method some objects are chosen randomly and the pictures about these topics are searched and downloaded from the Internet. Then all of the pictures are shown on the screen. After that, the user is asked to choose a specific object. The main advantage of this method is that non-English users can use it easily, even if they don't know English language. This method has been implemented by the PHP language.

In this method there are 2 two stages (recognition of the object and finding the object) and each of these two operations cannot be done by computer appropriately, so this method can resist the computerized attacks efficiently. On the other hand the users can work with this method easily because all of messages are shown in their native language. This method can also implement on other devices such as mobile phone, PDA (Personal Digital Assistant), and the devices which have touch screens.

3. In 2007 Ahn et al. created system called reCAPTCHA that channels the necessary cognitive work associated with human verification into a useful purpose: correcting ambiguous portions of text scanned from books using optical character recognition (OCR) software [12]. It overcomes the drawbacks of existing implementations. The authors justify their word-based CAPTCHA in light of the fact that they use two state-of-the-art OCR programs in an agreement-based approach to scanning text. Words that both OCR programs fail to agree on are used in the CAPTCHA to make use of human cognitive effort. Thus, if the CAPTCHA is broken automatically, state-of-the-art OCR will have advanced beyond that used in reCAPTCHA. reCAPTCHA works by showing a user two words: a control word and an unknown word. The control word is used to validate that the user is human, as in a standard word-based implementation. If the user is validated as human, the unknown word is also assumed to be valid, pending agreement from other users.

Figure 2: reCAPTCHA Example [12]

4. In October 2007 Elson et al. developed an image-based authentication system called Asirra that relies on a large database of images of pets from various animal shelters [13]. In order to pass the CAPTCHA, the user must select all images depicting either cats or dogs from a set of random images from both categories. The system takes advantage of the fact that users can easily differentiate between semantically different visual content, while the problem is difficult for computers.

Figure 3: Microsoft ASSIRA Interface [13]

5. In 2011 Yadava P, Sahu C and Shukla S. introduces a new TIME-VARIENT CAPTCHA [8]. In this paper the focus is not on the effective development of CAPTCHA but targeting a display of CAPTCHA over the webpage for a fixed time, CAPTCHA replaces itself until the final CAPTCHA is filled by user. Refresh process just work with CAPTCHA and don't affect the web page. So, now, automated program has to cover one more area to breach the CAPTCHA: to determine the final entered CAPTCHA.

6. The paper "Cyber Security Using Arabic CAPTCHA Scheme" is proposed by Bilal Khan et al [6]. The proposed scheme uses Arabic script to generate an image. The image is distorted by adding various types of noises in the background in the form of dots, lines and arcs. The background and foreground colors are selected so that the overall CAPTCHA image is attractive for the user. The varying number of characters, font types and font sizes make it extremely hard for the OCR to read our CAPTCHA. The Algorithm is efficient and the user does not have any problem while interacting with the system. To evaluate the readability rate of CAPTCHA images, a survey was conducted consisted of over one hundred and fifty individuals. Those survey participants were from Arabic speaking countries as well as non-Arabic South Asian countries who can understand the Arabic script. It was found from the survey that the overall readability rate of the images was high. So this proposed CAPTCHA scheme can be used in non- Arabic speaking countries where languages use Arabic script such as

Urdu, Pashto and Persian etc. Also some experiments were conducted to find out the robustness of the CAPTCHA that were encouraging.

Figure 4: Arabic CAPTCHA Scheme Example [6]

7. In 2011 Sushma Yalamanchili and Kameswara Rao proposed a DevaCAPTCHA that is highly usable as it is easy for humans to successfully provide the response [17]. Since we are using words from books and newspapers or an assortment of characters that are non-words, it is not difficult for humans to visually perceive these characters despite the distortions and noise due to their superior visual capabilities and cognitive abilities to make connections with words that they have encountered in some context. Distorted images containing random strings are still easy for humans to read while computers spend endless time processing information. The implementation of DevaCAPTCHA and the participation in OCR testing efforts related to Indian language scripts is to be taken up as future research work. Handwriting recognition and testing for Devanagari script is another future research activity.

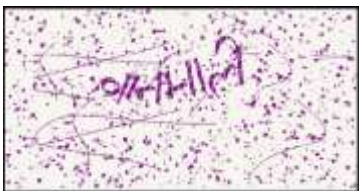

Figure 5: DevaCAPTCHA Example [17]

8. In the paper "What's Up CAPTCHA? A CAPTCHA Based on Image Orientation" by Rich Gossweiler, Maryam Kamvar, Shumeet Baluja a novel CAPTCHA system has been presented that requires user to adjust randomly rotated images to their upright orientation [15]. This is a task that will be familiar to many people given the use of early digital cameras, cell phones with cameras, and even the simple act of sorting through physical photographs. This system further improves traditional text-based CAPTCHAs in that it is language and written-script independent, and supports keyboard-difficult environments. It is important that random images are not chosen for this task; they *must* be carefully selected. Many typical vacation and snapshots contain cues revealing upright orientation. *A priori* knowledge of the image's label is not needed, which makes examples for this system easier to automatically generate than other image-based CAPTCHA systems. Furthermore, it is harder for bots to solve than the image-based CAPTCHAs that require a user to identify a common theme across a set of images, since the set of images to compare against is not closed.

Figure 6: Images with various orientation properties (left column: the image randomly rotated, right column: the Image in its upright position) [15].

III. Classification Of Captchas

CAPTCHAs means presenting a challenge response test to the users or humans. They are classified based on what is distorted that is whether characters, digits, or images.

1. Text-based CAPTCHAs
2. Image based CAPTCHAs
3. Audio-based CAPTCHAs
4. Video-based CAPTCHAs

1. **Text-based CAPTCHAs:** Text-based CAPTCHAs are very easy to implement. It is very effective and requires a large question bank. In Text-based CAPTCHAs simple questions are asked [2]. For example:

- What is four minus two?
- Which of cabbage, apple, and table is a vegetable?

Various forms or implementations of Text-based CAPTCHAs are as follows:

- **Gimpy:** Gimpy is a reliable text-based CAPTCHA developed by Yahoo and Carnegie Mellon university (CMU). In Gimpy, ten words randomly are picked from a dictionary. Then these random words are displayed in distorted and overlapped manner. After this the user has to enter three words in the image box. If the user enters three words correctly, then he is presumed to be a human.

Figure 7: Gimpy Example [14]

- **Ez-Gimpy:** Ez-Gimpy is a simplified version of Gimpy that is developed by Henry Baird. It is used by Yahoo in Messenger in case of their signup page. In case of Ez-Gimpy, a single word is chosen from a dictionary and then distortion is applied. The main task of user is to identify the distorted text correctly. It is not a good implementation and already broken by OCRs.

Fig 8: Ez-Gimpy Example [14]

- **Baffle-Text:** It is also designed by Henry Baird at California University at Berkeley [11]. It is a modified version of Gimpy.It It eliminates the Gimpy's drawback and is not prone to dictionary attacks because it does not include dictionary words. In Case of Baffle-text, a random characters or alphabets are picked to create a pronounceable text. Then the user is challenged to enter the correct word.

- **MSN CAPTCHA:** MSN CAPTCHA is used as a different CAPTCHA for providing services under MSN umbrella. These are also known as MSN passport service CAPTCHAs. In this type of CAPTCHA 8 characters (upper case) and digits are used. And the color of background is grey and of foreground is dark blue. In order to produce the ripple effect and to distort the characters, warping is used.

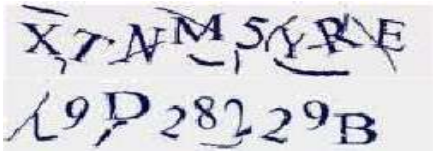

Figure 9: MSN CAPTCHA Example [14]

2. Image or Graphics-Based CAPTCHA: Graphics-based CAPTCHAs are challenge-tests in which the users have to guess those images that have some similarity [14]. For example: visual puzzles. Various implementations of Graphic-based CAPTCHAs are as follows:

- **Bongo:** Bongo is developed by M.M. Bongard who is a pattern recognition expert. The user has to solve the visual pattern recognition problem. In Bongo CAPTCHA, two series of blocks are displayed, the block in one series differ from those that in other series. The user has to find the distinct characteristics between these two series.

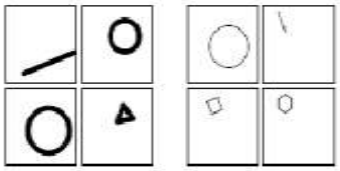

Figure 10: Bongo CAPTCHA Example [14]

- **PIX:** PIX includes large database of labeled images. In PIX the user has to find the common feature among the set of images. These set of images are distorted randomly before displaying them to the users [14].
  For example - pick the common features among the following 3 pictures =" AEROPLANE"

Figure 11: PIX Example [14]

3. Audio-Based CAPTCHAs: Audio-Based CAPTCHAs are based on the sound-based systems. These CAPTCHAs are developed for visually disabled users. It contains downloadable audio-clips [14]. In this type of CAPTCHA, first the user listens and after that submits the spoken word. The first sound-based system named ECO was implemented by the Nancy Chan of the City University in Hong Kong. The sound-based system is based on the gap in ability between computer machines and humans in recognizing spoken language. The program chooses a sequence of digits and words randomly and renders the words and number of digits into sound clips and distorts it. Then the distorted sound clip is presented to the user to enter the right word or number.

Figure 12: Google's Audio Enabled CAPTCHA Example [14]

4. Video-Based CAPTCHAs: In Video-Based CAPTCHAs, three words (tags) are provided to the user which describes a video [9]. If a user's tag belongs to a set of automatically generated ground truth tags then a challenge is passed.

Figure 13: Video CAPTCHA Example [9]

IV. Working Of Captcha

The following steps show the working of CAPTCHA [14].

1. **Create Random Value:** The first step in generating CAPTCHAs is to create some random words or sequence of digits. These random values are often hard to predict and guess.
2. **Generate an Image:** Images are used because these are harder to read by the web bots and are nice and readable to humans. It is an important step in CAPTCHA as simple text in images can be cracked easily.
3. **Store It:** The random string generated that is also in the image is stored for matching user input. For this session variables are used.
4. **Matching:** After these steps, CAPTCHA is drawn and shown on some form which one want to protect from being abused. User fills form along with CAPTCHA text & submits it.
   Now one has the following:

a. All submitted form data.
b. CAPTCHA string input by the user.
c. CAPTCHA string (real one) from session variables. Session variable is generally used as to keep stored values across page requests.

5. If match is found, then it is ok, otherwise not, in that case the message displayed to the user is that the CAPTCHA they had submitted are wrong.

V. APPLICATIONS OF Captchas

It is significant important to differentiate between a human and a machine over the internet in the fields of AI, Internet Security and human computer interaction. Various applications of CAPTCHAs are as under [3]:

1. Protecting Online Polls
2. Preventing E-mail Spam
3. Web registration Protection
4. E-Ticketing
5. Preventing Dictionary Attacks
6. Search Engine bots
7. Automating Document Identification

VI. Comparison

In this paper, the current review research is on CAPTCHAs. Currently, there are mainly three kinds of methods to implement the CAPTCHA mechanism: OCR (Optical character recognition) visual method, non-OCR visual method.

The CAPTCHA based on OCR visual method takes advantage of superiority in language barrier, security and easy use, becoming the most widely used CAPTCHA. However, with the fast development of OCR technology based on neural network, as well as the emergence of a variety of character segmentation technology, CAPTCHAs of lots of websites have been attacked. A Russian programmer has ever cracked the CAPTCHA mechanism of Yahoo with 35% success rate. Also, the CAPTCHA mechanism of Microsoft live mail has been bothered by junk mails many times. Given facts like these, newly designed CAPTCHAs have become increasingly complex, so that some of those are extremely difficult to identify.

Though there are many different kinds of specific implementations for non-OCR visual method, it eventually comes down to the OCR problem in general, requiring users to identify images. It is not so widely used. Up to now, except some research sites, commercial sites rarely use it. Non-OCR visual method is designed for special occasions and certain user groups, thus it has very limited applications.

In conclusion, the OCR-based visual method is the main way to implement current CAPTCHA mechanism. However, it could no longer strike a balance between security and easy use, calling for a new kind of CAPTCHA to address this increasingly prominent problem.

VII. Conclusion

This paper gives a description for various existing captcha schemes with a literature survey and provides a description for working of Captcha. It also describes the classification of various captcha schemes. And finally, the various applications of Captchas and comparison between OCR and NON-OCR based captchas are also presented.

Acknowledgement

I would gratefully and sincerely appreciate my supervisor: Assistant Prof. Baljit Singh Saini. Their inspiring guidance, rich experience and sustained encouragement enabled me to develop an intensive understanding of my research area. Without the generous help of my supervisor, this work would not have been possible. I am honored to have Prof. Baljit Singh Saini from Lovely Professional University as my opponent. I thank him for his kind support and helpful suggestions during the discussions in my M.Tech study.

References

[1] L. von Ahn, M. Blum, and J. Langford, "Telling Human and Computers Apart Automatically," in Communications of the ACM, vol. 47,February 2004, no. 2, pp. 57-60. DOI:10.1145/966389.966390
[2] L. von Ahn, M. Blum, N. Hopper, and J. Langford, The CAPTCHA web page: http://www.CAPTCHA.net,2000.
[3] Ahn, L. von, Blum, M., Hopper, N. J., & Langford, J., (2003), "CAPTCHA: Using hard AI problems for security", Proceedings of Eurocrypt 2003.
[4] Shirali-Shahreza, M.H. & Shirali-Shahreza, M., (2007) "Multilingual CAPTCHA", ICCC 2007 IEEE International Conference on Computational Cybernetics, 19–21 October, Gammarth, Tunisiali, pg 136.
[5] M. Blum, L. von Ahn, and J. Langford, The CAPTCHA Project (Completely Automated Public Turing Test to Tell Computers and Humans Apart), School of Computer Science, Carnegie- Mellon University, November 2000,http://www.CAPTCHA.net.
[6] M.H. Shirali-Shahreza and M. Shirali-Shahreza, "Persian/Arabic Baffletext CAPTCHA," Journal of University Computer Science (J.UCS), vol. 12, no. 12, December 2006, pp.1783- 1796. DOI: 10.3217/jucs-012- 12-1783
[7] H.S. Baird and J.L. Bentley, "Implicit CAPTCHAs," Proceedings SPIE/IS&T Conference on Document Recognition and Retrieval XII (DR&R2005), San Jose, 2005, pp. 191-196. DOI: 10.1117/12.590944
[8] Yadava P, Sahu C and Shukla S. (2011), "Time-Variant CAPTCHA: Generating Strong CAPTCHA Security by Reducing Time to Automated Computer Programs",JETCIS VOL. 2, NO. 12, pp 701 704.
[9] Kurt Alfred Kluever. Evaluating the Usability and Security of a Video CAPTCHA. Master's thesis, Rochester Institute of Technology, Rochester, NY, August 2008.
[10] H.S. Baird and K. Popat, "Human Interactive Proofs and DocumentImage Analysis," Proceedings of the 5th IAPR International Workshop on Document Analysis Systems, Princeton, LNCS 2423, 2002, pp. 507- 518.
[11] M. Chew and H. S. Baird, "BaffleText: a Human Interactive Proof," Proceedings of the 10th SPIE/IS&T Document Recognition and Retrieval Conference (DRR2003), Santa Clara, CA, 2003, pp. 305-316.
[12] Luis von Ahn, Ben Maurer, Colin McMillen, Mike Crawford, Ryan Staake, and Manuel Blum. reCAPTCHA Project. Online, http://www.recaptcha.net, May 2007.
[13] John Douceur, Jeremy Elson, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In Proceedings of the 14th ACM Conference on Computer and Communications Security, pages 366{374, New York, NY, USA, October 2007.
[14] www.slideshare.net/kunalkiit/seminar-**report-on-captcha**
[15] Rich Gossweiler, Maryam Kamvar, Shumeet Baluja: What's Up CAPTCHA? A CAPTCHA Based on Image Orientation, pg: 2
[16] Richard Chow, Philippe Golle, Markus Jakobsson "Making CAPTCHAS Clickable" at Palo Alto Research Centre, pg: 1-4
[17] Sushma Yalamanchili and Kameswara Rao: A FRAMEWORK FOR DEVANAGARI SCRIPT-BASED CAPTCHA, International Journal of Advanced Information Technology (IJAIT) Vol. 1, No. 4, August 2011 DOI: 10.5121/ijait.2011.1404 47.