Volume 3, No. 6, June 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
TECHNICAL NOTE
Available Online at www.jgrcs.info

SURVEY ON CAPTCHA SYSTEMS

Rizwan Ur Rahman
Maulana Azad National Institute of Technology, Bhopal, M.P, India
rizwan.rahman12@gmail.com

Abstract: CAPTCHA stands for completely automated public Turing test to tell computer and human apart. Due to the enormous growth of Internet, security of web application has become a vital issue and many web applications facing a threat of Internet bot also known as Internet Robot is an automated script which executes over the web forms and wastes precious web space. CAPTCHA has become de facto standard for securing web applications from Internet Bot and almost all the registration web forms use this test. During a last decade many researchers has done a work on CAPTCHA systems. This paper is a collective survey of work done on CAPTCHA systems. In first section typical applications of CAPTCHA has been discussed and in the next section strengths and weaknesses of text based and image based CAPTCHA are discussed.

INTRODUCTION

Nowadays web application like email, social networking sites, blogs, e-governance sites etc has become everybody's need. With rapid growth of Internet, Security is also becoming critical issue. Many websites uses CAPTCHA or completely automated public Turing test to tell computer and human apart [1] to block Bot or automated script execution. For example in figure 1 human can identified the distorted text but not the current computer programs [2]. Although the term CAPTCHA was introduced by John Langford of Carnegie Mellon University [3] but the ground work was already done by Moni Noar who introduced the concept of Turing test to identify the difference between a human and BOT in 1996 [4]. The original motivation of CAPTCHA came from an online poll [1] asking which is best graduate school of computer science. In one sense CAPTCHA can referred as "reverse Turing test" since its task is to determine whether the remote user is human or script. We can broadly categorize CAPTCHA into four schemes [5] and these are given below.

Text based CAPTCHA: Text based CAPTCHA are the most widely used in web applications. In this system server renders set of characters after distorting the text and with a noise addition. Many web sites like Yahoo (www.yahoo.com), Microsoft (www.hotmail.com), Google (www.mail.google.com) and Wikipedia (www. Wikipedia.com) use their own CAPTCHA. Another interesting text based CAPTCHA is reCAPTCHA [2]. It is free web service available for major web development languages like ASP.NET, PHP and JSP. It helps digitizing the books and news papers that were written before the computer ages. reCAPTCHA improves the process of digitizing books by rendering a word that cannot be read by computer or OCR (optical character recognition) technique in the form of CAPTCHA for humans to solve. Each word that cannot be read correctly by OCR is placed on an image and used as a CAPTCHA. 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 to server. 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.

reCAPTCHA is shown in figure 1. Table I presents major text based CAPTCHA that are available today.

Figure: 1 reCAPTCHA.

Table: 1Text based CAPTCHA

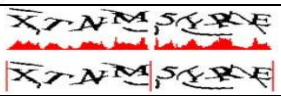
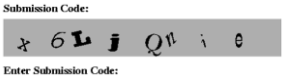
Sr. No.	Text based CAPTCHA	Website
1		Wikipedia[6]
2		Yahoo[7]
3		Microsoft[8]
4		Mailblocks[9]
5		Ticket Master[10]
6		Alta Vista [11]
7		Facebook [12]

Image based CAPTCHA: In this scheme the user is required to identify some image recognition task. ESP Pix is first image CAPTCHA and it was developed at Carnegie Mellon University [13]. A snapshot of ESP Pix CAPTCHA is shown in Figure 2. In ESP Pix the user has given four images and in order to pass this test the user has to select word related to those four images from drop down list of 72 choices.

Figure: 2 ESP Pix CAPTCHA

Another Image CAPTCHA is Asirra Stands for Animal Species Image Recognition for Restricting Access is a cat or dog labeling based CAPTCHA design [14]. In this test user has to select all the pics of cat. Asirra is randomly choosing images from petfinder.com. Snapshot of Asirra is shown in figure 3.

Figure: 3 Asirra CAPTCHA

One CAPTCHA is available as paid service is CAPTCHA the dog [14]. It Shows nine images in 3 by 3 grid and user is asked to chose all the images of cat one by one until images become dog's images. A snapshot of it is shown in Figure 4.The dog is randomly placed among nine cats and the process is repeated for three times. Multi Model CAPTCHA Combines text and image based System together. In this end user is shown an image and four text labels associated with the image. Text labels are embedded in the image and the user is asked to select a relevant text label [15]. A snapshot of Multi Model CAPTCHA is shown in Figure 5

Figure: 4 CAPTCHA The Dog

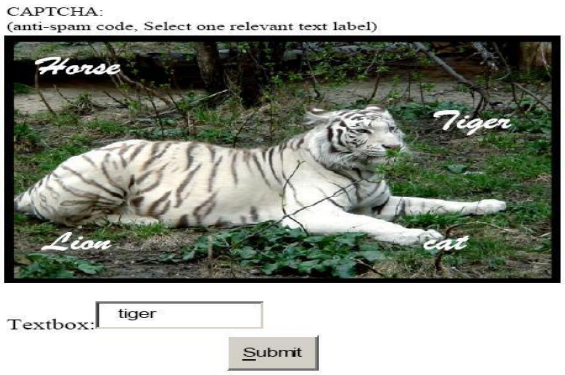


Figure: 5 Multi Modal CAPTCHA

Another improved image CAPTCHA is Dynamic Image Based CAPTCHA (DIBC).In this CAPTCHA system user is required to recognize the exact matching image or images to pass the Turing test. An image is selected randomly from image database and is placed in a grid of six images random number of times. User is supposed to submit all the correct version of the filtered image for clearing the Turing test in maximum of 5 attempts [16].It is shown in Figure 6

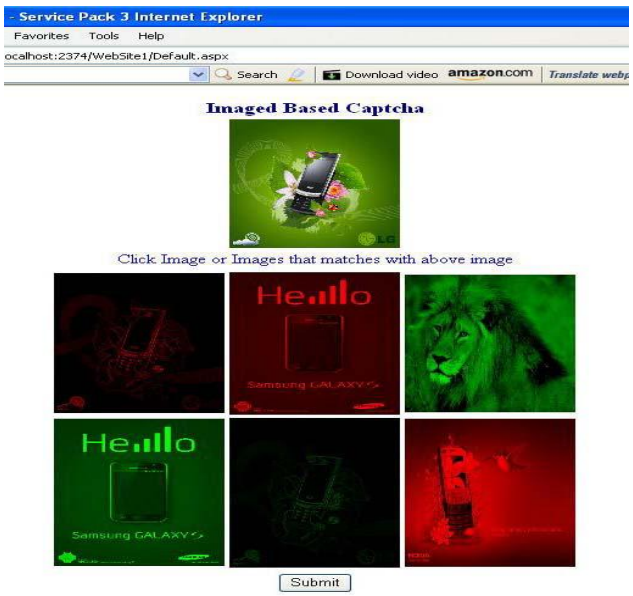


Figure: 6 Dynamic Image Based CAPTCHA

IdentiPic is photo based CAPTCHA system where user has to identify picture [17].Three pictures are shown and corresponding to each pic there is drop down list having ten options. A snapshot of Identipic is shown in Figure 7.

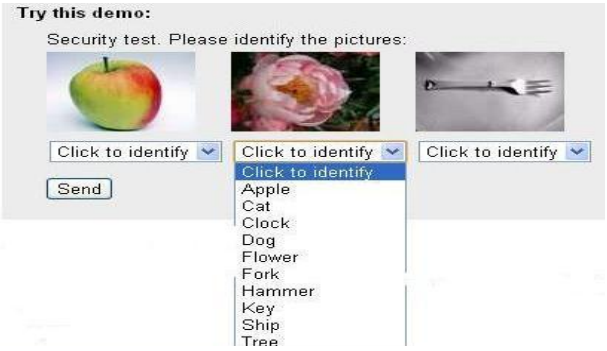


Figure: 7 IdentiPic CAPTCHA

Puzzle based CAPTCHA: It is also referred as question based CAPTCHA [18]. In this test, a small mathematical problem is generated according to some predefined rules. The problem then rendered by the server to the user answer of which is already known to server. Solving of this problem requires an ability of understanding text of question, only a human user can answer this question. Figure 8 illustrates the GUI of Question based CAPTCHA.

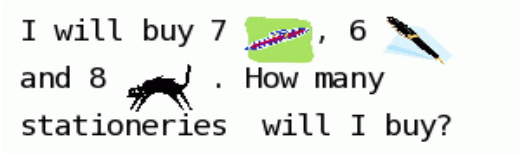


Figure: 8 Question based CAPTCHA

Audio CAPTCHA: In this, a user is asked to pass typically audio or voice recognition task. A typical audio CAPTCHA is shown in Figure 9.

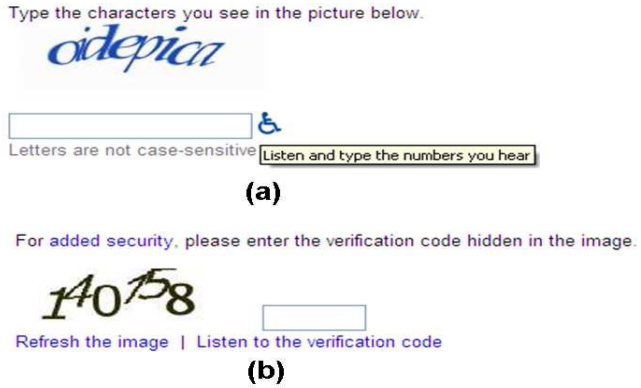


Figure: 9 Audio CAPTCHA

APPLICATION OF CAPTCHA

a. Free Registration through web forms: Millions of websites on internet offer free registration to services such as Social Networking sites, Email services, Web blogs etc. Unfortunately many web sites are attacked by web robots. Web robots are typically scripts which registered thousands of email account on the internet wasting precious web space [19].
b. Online polling: The Original Motivation of CAPTCHA came from an online poll asking "which is the best graduate School in computer Science?" Students of Carnegie Mellon University wrote a program that voted for CMU thousands of times. The next day, students at MIT wrote their own program and the poll became a contest between voting bots. Can the result of any online poll be trusted? Not unless the poll ensures that only humans can vote [20].
c. Web crawler: A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion [21]. It provides reasonable solution to web pages that we want not to be index by search engines.
d. Online Games: Another application of CAPTCHA is online games [22, 23, 24] where it is preventing web Robots from playing games.
e. Dictionary Attack: In cryptanalysis and computer security, a dictionary attack is a technique for defeating a cipher or authentication mechanism by trying to determine its decryption key or passphrase by searching likely possibilities [25]. CAPTCHA is used in preventing dictionary attacks in many applications [26, 27].
f. Phishing Attack: Phishing is attempting to get information such as bank details, usernames, passwords, and credit card details by masquerading as a trustworthy entity [28]. CAPTCHA also provides plausible solution to phishing attacks [29].
g. Worms and Spam: Last but not the least CAPTCHA provides a solution against worms and spam i.e., it receives mail only if it is sure that there is human behind it not the computer bot.

STRENGTH AND WEAKNESSES OF CAPTCHA

Text Based CAPTCHA: Although Text based CAPTCHAs are the most widely used CAPTCHA that are used in web application but there are some common weaknesses. The Number of classes of characters and digits are very small. Since characters and digits have limited geometry (limited font families) so it is possible to identify them through OCR or Optical Character Recognition Technique. When the noise and distortion is added to the text based CAPTCHA they often creates a problem in recognizing them. Although some alphabets and digits have very different shapes, but when they are distorted, it is become difficult to recognize them. This problem is most common in Text based CAPTCHA. Given below is a list of common confusing character pairs [30].

***Digit vs digit*:** In many cases 8 may look like 6 or 9. Depending on what type of font is used by the system 7 may look like 1 and vice versa.

***Letter vs letters*:** If distortion is applied for example "cl" can be confused with'd'; "nn" can be confused with 'm' or 'rn' and "vv" can be confused with 'w'. Given below list in Table:2 of some confusing character in Google CAPTCHA [30].

Table : 2 Confusing characters in Google CAPTCHA

Sr. No.	CAPTCHA	Problem
1		There is confusion first 2 characters are "cl" or 'd'
2		Another confusion of "cl" and 'd'
3		Whether 2nd and 3rd character are 'I' and 'v' respectively or it 'w'
4		First two characters are 'rn' or it is continuous m
5		A real headache: is the first part "m" or "rn,the middle part inv"or"nw"?

Decaptcha tool breaks the Wikipedia scheme, illustrated in figure 2, approximately 25% of the time. 13 out of 15 of the most widely used current schemes are similarly vulnerable

to automated attack by Decaptchal [31].Efficiency of Decaptcha against real captchas from Authorize, Baidu, Blizzard, Captcha.net, CNN, Digg, eBay, Google, Megaupload, NIH, reCAPTCHA, Reddit, Skyrock, Slashdot, and Wikipedia. None of these captcha schemes had been reported broken prior to this work. Of these 15 CAPTCHAs, Decaptcha achieve 1%-10% success rate on two (Baidu, Skyrock), 10-24% on two (CNN, Digg), 25-49% on four (eBay, Reddit, Slashdot, Wikipedia), and 50% or greater on five (Authorize, Blizzard, Captcha.net, Megaupload, NIH) [31].

The robustness of Text CAPTCHA studied in the field computer vision, Image Processing and document analysis. For instance, Mori and Malik [32] have broken from object recognition algorithms EZ-Gimpy CAPTCHA with a success 92% and Gimpy CAPTCHA with success 33%. J Yan and A S El Ahmad [33] have broken a number of CAPTCHAs with almost 100% success by simply counting the number of pixels of each segmented scharacter, although these schemes were all resisting against OCR software on the available in market. Chellapilla and Simard [34] attacked a number of text based CAPTCHAs of different web application by applying machine learning algorithms, achieving a success rate ranging from 4.89% to 66.2%. Moy et al developed as in [35] distortion estimation techniques to break EZ-Gimpy achieving success rate of up to 99% and 4-letter Gimpy-r with a success rate of up to 78%. J Yan and A S El

Ahmad [35] have implemented a low-cost segmentation attack that has achieved a success rate of higher than 90% on the latest version of this Microsoft CAPTCHA. They have shown that the Microsoft scheme can be broken with an overall segmentation and then recognition) success rate of about 60%.

In January 2008 article publish in informationweek.com claiming Yahoo's CAPTCHA security had been broken [36].In February2008 in www.theregister.co.uk claiming that Google's CAPTCHA had been cracked by spammers [37]. In May, Microsoft's CAPTCHA security had been broken [38].

Image based CAPTCHA: The advantage of using Image based CAPTCHA is that pattern recognition of in image is a hard AI Problem and therefore it is very hard to break this test using pattern recognition technique. The weaknesses of ESP Pix CAPTCHA shown in figure 2 has summarized below: ESP Pix is available only in English language so the end user must know the English vocabulary but there are only twenty seven percent internet users are English speaking [39].

a. It creates a problem to users having low vision or learning disability [40].
b. Most of the time object recognition becomes cumbersome due to the ambiguity presents in image objects. Instead of Turing test it has become almost an IQ test.
c. Probability of an automated bot entering into a site is 1/number of choices and here the choices are 72 so the probability is 1/72 that means one attempt will be successful out of seventy two attempts.

The weaknesses of MMC scheme shown in figure 5 are same as described for ESP Pix CAPTCHA such as it is available only in English. Text labels that are embedded to the image object are in simple fonts so it can be easily recognized by OCR technique. In this case the probability of entering into a site is 1/4 that is 25%, since the number of choices are 4.

The probability of guessing an image has been increase in CAPTCHA The Dog shown in figure 4 method, the probability is (1/9*9*9) that is 1/729 or 0.00137, since one image is selected out of 9 images and it repeats 3 times. The disadvantage of using this CAPTCHA is that it creates an extra overhead on the server since it requires 3 extra round trips to server.

Dynamic Image Based CAPTCHA shown in figure 6 is not English dependent and the probability has also been increases from 0.00137 to 0.000061
Summary of Image based CAPTCHA features discussed above is presented in tabular format below.

Table: 3 Comparasion of Image based CAPTCHA

CAPTCHA System	#Choice	#Attempt to Pass the test	Probability	English dependency
ESP-Pix	72	10	0.0138	Yes
MMC	4	3	0.25	Yes
CAPTCHA THE DOG	9	Not defined	0.00137	No
identiPic	10	Not defined	0.001	Yes
DIBC	6	3	0.000061	No

CONCLUSION

CAPTCHA has become de facto standard for security measures on the World Wide Web that prevent automated scripts from abusing online services. In this paper we surveyed the research done on CAPTCHA systems during last decade. We carried out systematic study on text based CAPTCHA and image based CAPTCHA and tried to identify strengths and the weaknesses of the CAPTCHA Systems that are available today. It is observed that lots of research to be done on text based CAPTCHA systems for the usability. As the field of Artificial intelligence advances more CAPTCHA systems will break in future. We expect this survey will help the researchers in field of web security and Artificial intelligence to easily get the research done previously.

REFERENCES

[1] CAPTCHA Project, available at, http://www.captcha.net.

[2] reCAPTCHA project, available at http://recaptcha.net.

[3] L Ahn, M. Blum and J. Langford.Telling Humans and Computers Apart Automatically. Communications of the ACM, 47(2):57–60, 2004.

[4] Moni Noar."Verification of a human in the loop, or Identification via the Turing test",taken from wisdom.weizmann.ac.il, Available at http://www.wisdom.weizmann.ac.il/`nor/PAPERS/human.ps.

[5] Chandvale, A.A; Sapkal, A.M; Jalnekar, R. M., A Framework to analyze the security of Text based CAPTCHA, International Journal of Computer Applications, Vol 1 issue 27, pp. 127-132..

[6] Wikipedia, www.wikpedia.com.

[7] Yahoo, www.yahoo.com.

[8] Microsoft, www.hotmail.com.

[9] Mail Blocks, www.mail-block.com.

[10] Ticket Master, www.ticketmaster.com.

[11] Alta Vista, www.altavista.com.

[12] Facebook, www.facebook.com.

[13] ESP Pix CAPTCHA, taken from cmu server, Available at http://server251.theory.cs.cmu.edu/cgi-bin/esp-pix/esp-pix.

[14] Captcha the dog, available at, http://www.captchathedog.com.

[15] Almazayd, A.S. Ahmed, Y Kouchay, "Multi Modal CAPTCHA: A User Verification Scheme". Proceeding of IEEE Int Conf on Information Science and Application, Korea, 2011, pp. 1-7.

[16] Rizwan ur Rahman, Deepak Singh Tomar, Sujoy Das, "Dynamic Image Based CAPTCHA". Proceeding of IEEE Int Conf on Communication Systems and Network Technologies, India 2012, pp. 90-94.

[17] identiPic CAPTCHA, available at , http://www.identipic.com.

[18] M. Shirali-Shahreza and S. Shirali-Shahreza, "Question-Based CAPTCHA,"Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), Sivakasi, India, December 13-15, 2007, Vol. 4, pp. 54-58.

[19] H. S. Baird and K. Popat. Human interactive proofs and document image analysis. Proc. of 5th IAPR Int. Workshop on Document Analysis Systems (DAS 2002), vol. 2423 of LNCS, pp. 507–518, 2002.

[20] CAPTCHA Using a hard problems for security. Luis Von Abn, Manual Blum, Nichlas Hoper and John Langford of Carnegie Mellon University.

[21] Web Crawler, available at http://en.wikipedia.org/wiki/Web_crawler.

[22] Roman V. Yampolskiy, Venu Govindaraju. Embedded noninteractive continuous bot detection. Computers in Entertainment (CIE), vol. 5 Issue 4. March 2008. (ACM).

[23] P. Golle and N. Ducheneaut. Preventing bots from playing online games. Comput. Entertain., 3(3):3, 2005.

[24] Hilaire, S.; Hyun-chul Kim; Chong-kwon Kim; How to deal with bot scum in MMORPGs?; IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR), 2010. Pp: 1 – 6.

[25] Dictionary Attack, available at http://en.wikipedia.org/wiki/Dictionary_attack.

[26] S. Chakrabarti and M. Singhal. Password-based authentication: Preventing dictionary attacks. Computer, 40(6): pp. 68–74, June 2007.

[27] B. Pinkas and T. Sander. Securing passwords against dictionary attacks. Proc. of 9th Conf. on Computer and Communications Security, pp. 161–170, Nov. 2002.

[28] Phishing Attack, http://en.wikipedia.org/wiki/Phishing.

[29] Tak, G.K.; Badge, N.; Manwatkar, P.; Ranganathan, A.; Tapaswi, S.; Asynchronous Anti Phishing Image Captcha approach towards phishing; 2nd International Conference on Future Computer and Communication (ICFCC), 2010. Vol 3, pp: 694 – 698.

[30] Jeff Yan and Ahmed Slah ," Usability of CAPTCHAs or usability issues in CAPTCHA design ", Proceedings of the 4th symposium on Usable privacy and security Pages 44-52 ACM New York, NY, USA.

[31] Elie Bursztein, Matthieu Martin, John Mitchell "Text-based CAPTCHA strengths and weaknesses", Proceedings of the 18th ACM conference on Computer and communications security Pages 125-138 ACM New York, NY, USA.

[32] G Mori and J Malik. "Recognising objects in adversarial clutter: breaking a visual CAPTCHA", IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 2003.

[33] J Yan and A S El Ahmad. "Breaking Visual CAPTCHAs with Naive 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.

[34] G Moy, N Jones, C Harkless and R Potter."Distortion estimation techniques in solving visual CAPTCHAs", IEEE CVPR, 2004.

[35] J.Yan and A. Ahmad. A Low-cost Attack on a Microsoft CAPCTHA. The 15th ACM Conference on Computer and Communications Securtity (CCS), 2008.

[36] http://www.informationweek.com/news/internet/webdev/showArticle.jhtml?articleID=205900620.

[37] http://www.theregister.co.uk/2008/02/25/gmail captchacrack/

[38] http://blogs.zdnet.com/security/?p=1232&tag=nl.e550..

[39] English Language taken from wikipedia and available at http://www.en.wikipedia.com/English_Language.

[40] Move & Select: 2-Layer CAPTCHA based on Cognitive Psychology for securing web services , Taken from ijens.org, available at http://www.ijens.org/Vol_11_I_05/117005-8383-IJVIPNS-IJENS.pdf.

