

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

CAPTCHA: VULNERABILITIES AND FUTURE ASPECTS

Hardeep Singh

P.G. Department of Computer Science and Applications, BBK DAV College for Women
Amritsar

ABSTRACT

CAPTCHA; now-a-days; is an almost standard security technology, and has found widespread application in commercial websites. CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a challenge-response system test designed to differentiate humans from automated programs. Usability and robustness are two fundamental issues with CAPTCHA, and they often interconnect with each other. This paper discusses various limitations, challenges and vulnerabilities of the current CAPTCHAs being used and addresses the new design of more protected CAPTCHAs. Some of these issues are intuitive, but some others have subtle implications for robustness or security. Section 1 introduces the term CAPTCHA, its uses and some of the work done by others in this field. Section 2 explains how a CAPTCHA is verified for correctness and also addresses different approaches to CAPTCHA designs. Section 3 presents various limitations faced by current CAPTCHAs being used these days. Section 4 discusses some problems in CAPTCHA and introduces with the solution to this problem; namely; SAPTCHA. Section 5 shows the result of using SAPTCHA in place of CAPTCHA. Section 6 concludes the methodology and section 7 discusses the next generation of CAPTCHAs.

Keywords: CAPTCHA, bots, Optical Character Recognition, distorted text, SAPTCHA, HCI (Human Computer Interaction), 3-D CAPTCHA.

I. INTRODUCTION

The term "CAPTCHA" (based upon the word capture) was introduced in 2000 by Luis von Ahn, Manuel Blum, Nicholas J. Hopper (Carnegie Mellon University), and John Langford (IBM). CAPTCHA differentiates between human and internet bots by setting some task that is easy for most humans to perform but is more difficult and time-consuming for current bots to complete. Typically a CAPTCHA has text in different fonts, colors and angles that make it difficult for a computer program to read but hopefully not for a human.



Figure 1: Some Examples of CAPTCHAs

CAPTCHA was first used in 2000, in order to prevent spammers and bots from making spam and generating fake email accounts. The first companies to find itself related in this sort of problems were AltaVista and Yahoo, with yahoo chat, that was bombarded with spam bots entering in chat rooms and spamming (advertising) there sites. Already a lot of work has been done in this field. Some of the work already done in this domain is described below:

AltaVista: AltaVista started a free "add-URL" service. This service was important to AltaVista since it broadens its search coverage. But some users were abusing the service by automating the submission of large numbers of URLs, so that they could get AltaVista's importance ranking algorithms. In 1997, AltaVista's Chief Scientist Andrei Broder

and his colleagues developed a filter to discourage the automatic submission of URLs to their search engine. In this method, an image of printed text (CAPTCHA) was generated randomly so that machine vision (OCR) systems cannot read it but humans still can. This method reduced number of "spam add-URL" by over 95% [1].

Yahoo's Chat Room Problem: Yahoo faced a serious "chat room problem" in which bots join online chat rooms and irritate people by pointing them to advertising sites. The major problem was "How all bots could be refused to enter to chat rooms?" CMU's Prof. Manuel Blum, Luis A. von Ahn, and John Langford decided a solution that anyone who wants to join the chat room must pass through some sort of test. This test was named as CAPTCHA [2].

CAPTCHA is a security requirement of various services provided on the internet. It appears when someone elects to integrate with one of such services. CAPTCHA is a form of security test given to prevent bots from filling out forms on the Internet. Following problems motivated the use of CAPTCHA:

Online Polls: The result of an online poll cannot be trusted because anybody could just write a program to vote for their favorite option thousands of times. In November 1999, an online poll asked for the opinion 'Which was the best graduate school in computer science?' As is the case with most online polls, IP addresses of voters were recorded in order to prevent single users from voting more than once. However, students at Carnegie Mellon found a way to stuff the ballots using programs that voted for CMU thousands of times. CMU's score started growing rapidly. The next day, students at MIT wrote their own program and the poll became a contest between voting "bots." MIT finished with 21,156 votes, Carnegie Mellon with 21,032 and every other school with less than 1,000. Can the result of any online poll be trusted? Not unless the poll ensures that only humans can vote [3].


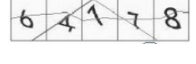

Registration Forms: Many companies like Google, Yahoo and Microsoft etc offer free email services. Up until a few years ago, most of these services suffered from a specific type of attack: "bots" that would sign up for thousands of email accounts every minute. The solution to this problem was to use CAPTCHA to ensure that only humans obtain free accounts. In general, free services should be protected with a CAPTCHA in order to prevent abuse by automated scripts.


E-Ticketing: Ticket brokers like Ticket Master can also use CAPTCHA applications. These applications help prevent ticket scalpers from bombarding the service with massive ticket purchases for big events.

Search Engine Bots: It is sometimes desirable to keep web pages un-indexed to prevent others from finding them easily. There is an html tag to prevent search engine bots from reading web pages. The tag, however, doesn't guarantee that bots won't read a web page; it only serves to say "no bots, please." Search engine bots, since they usually belong to large companies, respect web pages that don't want to allow them in. However, in order to truly guarantee that bots won't enter a web site, CAPTCHA is needed.

Lots of confusing characters are used in CAPTCHAs to make CAPTCHAs safer from bots. Some common confusing character pairs are discussed in Table 1.

Table 1: Confusing characters in CAPTCHAs

Confusing Characters	Description	Example Image	Image Description
Letters vs. Digits	It is hard to tell distorted 0 from O, 6 from G and b, 5 from S or s, 2 from Z or z, 1 from I		Is second last character "2" or "Z"?
Digit vs. Digit	7 is written differently in different countries and often what looks like a 7 may in fact be a 1, and 8 can look like 6 or 9		Is third character "7" or "1"?
Letter vs. Letter	Under some distortions, "vv" can resemble "w", "cl" can resemble "d", "nn" could resemble "m", "rn" can resemble "m". Table 1 shows some such confusing examples		Is first character "d" or connected "cl"?

Characters vs. Clutters	In CAPTCHAs, random arcs are introduced as clutters. There is always confusion between arcs and clutters. For example, it is difficult to tell an arc from characters such as 'J', '7', and 'L'.		Is first character "7" or "L" or "Z"?
-------------------------	--	--	---------------------------------------

Because computing is becoming pervasive, and computerized tasks and services are very common, the need for increased levels of security has led to the development of CAPTCHA for computers to ensure that they are dealing with humans in situations where human interaction is essential to security. Activities such as online e-commerce transactions, search engine submissions, web polls, web registrations, free e-mail service registration and other automated services are subject to software programs, or bots, that mimic the behavior of humans in order to deviate the results of the automated task or perform malicious activities, such as gathering e-mail addresses for spamming [4].

II. DIFFERENT APPROACHES TO CAPTCHA DESIGNS

CAPTCHA; funny-looking letters and numbers that users often have to type on entry forms; are an attempt to separate legitimate entries from those made by bots and scripts. However, as computers have become more adept at deciphering CAPTCHA, they have also become harder and harder to read. In order to validate the digital transaction, using the CAPTCHA system; the user is presented with a distorted word typically placed on top of a distorted background. The user must type the word into a field in order to complete the process. Computers have a difficult time decoding the distorted words while humans can easily decipher the text.

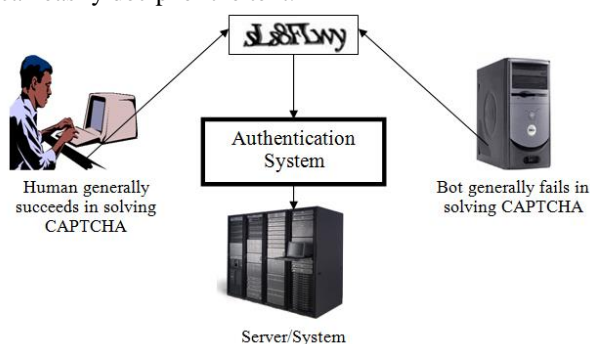
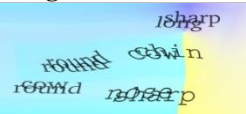


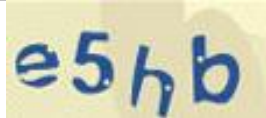
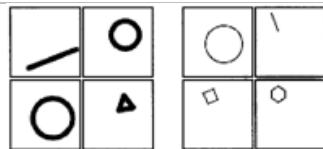





Figure 2: How CAPTCHA is verified

Some CAPTCHAs use pictures instead of words where the user is presented with a series of pictures and asked 'What is the common element among all of the pictures?'. By entering that common element, the user validates the transaction and the computer knows it is dealing with a human and not a bot. While CAPTCHAs may be a minor inconvenience to the user, they can save webmasters a lot of hassle by fending off automated programs. CAPTCHAs are commonly seen at the end of online forms. Fortunately, most CAPTCHAs allow the user to regenerate the image if the text is too difficult to read. CAPTCHA is sometimes described as a Reverse Turing Test, because it is administered by a machine and targeted to a human, in contrast to the standard Turing test that is typically administered by a human and targeted to a machine. CAPTCHA are of various types. Some general approaches are shown in Table 2.

Table 2: Some approaches to CAPTCHAs

CAPTCHA Type	Sub Categories	Description	Image
Text Based	Gimpy	These CAPTCHAs are designed by Yahoo and CMU. In this 5 random words are picked up from dictionary and then they are distorted. User has to recognize at least three words. If the user is correct, he is admitted.	

	<i>EZ-Gimpy</i>	It is a modified version of Gimpy. Yahoo used this version in messenger. It has only one random string of characters. The word is not a dictionary word, so not prone to dictionary attack. The word is then distorted and noise and background is added. EZ-Gimpy is not a good implementation as it is already broken by OCRs.	
	<i>Gimpy-r</i>	Random letter of words are picked, distorted and noise and background is added to them. There is 78% success rate in breaking Gimpy-r.	
	<i>Simard's HIP (MSN)</i>	Here letters and numbers are picked randomly. They are then distorted and arcs are added to them. These are provided for Microsoft's MSN service. It uses up to 8 characters which are distorted. It has very strong implementation as it has not been broken yet.	
Image or Graphics Based	<i>BONGO</i>	It is named after M. M. Bongard (pattern recognition expert). Here user has to solve a pattern recognition problem. This is actually a visual analogy problem. It displays two series of blocks. User must find the characteristic that sets the two series apart. User is asked to determine which series each of four single blocks belongs to .	
	<i>PIX</i>	It is a type of photo recognition problem. It needs large image database of labeled images. Here a set of distorted images is shown and user has to recognize common feature among them. Its poor implementations are easy to crack [5].	
Speech or Audio Based	Here a word or a sequence of numbers is picked randomly. They are then rendered into an audio clip using software. The audio clip is then distorted. User is asked to identify the word or numbers. It is usually used for visually disabled users.		
Animation Based	These use Flash, MPEG, animated GIF. These are often combined with speech. Here too the weaknesses of image CAPTCHA apply. These are usually easier to crack due to extra data for pattern matching to analyze. They result in much higher processing and traffic load. They are not practical in most cases.		

3-Dimensional	A 3D-based Captcha system claims to be both unbreakable and easier for humans to solve than the old text based systems. The system was developed by social website Yuniti.com. It works by asking users to identify 3D objects rather than words or numbers. There are three objects to be identified and the list is endless, making it even harder for scammers to guess correctly. This method is not yet common.	
Mathematical	Here a mathematical equation is given. User has to solve that equation and fill in the answer in space provided.	

III. CHALLENGES FACED BY CURRENT CAPTCHAS

The CAPTCHAs which are being used currently could be broken easily by using different techniques and software [6]. Some ways to break CAPTCHAs are:

- **Exploring bugs in implementation** that allow the attacker to completely bypass the CAPTCHA. This could be achieved in following ways:
 - Reusing the session ID of a known CAPTCHA image.
 - CAPTCHA use hash of the solution as a key passed to the client to validate. Often it is small enough in size that it can be cracked.
 - Implementations use only a small pool of CAPTCHA images which is a limitation in itself.
- **Improving Optical Character Recognition Software.** CAPTCHAs could be broken by using the programs that could perform following functions:
 - Extracting the image from the web page.
 - Removing the background clutter, and then detecting the thin lines.
 - Segmentation: i.e. splitting the image into regions each containing a single letter and then identifying the letter for each region.
- **Man-in-the-middle attack:** Using cheap human labor to process the tests. This is known as sweatshops. Here following method is used:
 - Copying the CAPTCHA from target.
 - Posting it on the attacker's website.
 - Forwarding the answer to the target.

For example, following algorithm is used to break EZ-Gimpy:

1. Locate possible letters at various locations.
2. Construct graph of consistent letters.
3. Look for probable words in the graph.

This algorithm has got 92% success in breaking CAPTCHA [7].

The bottom line is that, the spammers have written some good image recognition software which is great threat to CAPTCHA. Even now days, a lot of good voice recognition software are available which are imposing risks to voice CAPTCHA as well. In this paper, I have tried to find the solution of the above mentioned problems by looking for different alternatives.

IV. INTRODUCTION TO SAPTCHA: A SOLUTION TO CAPTCHA VULNERABILITIES

The main problem with the CAPTCHA is that the spammers have written some good image recognition software which is a great threat to CAPTCHA used today. Many of the text based CAPTCHAs, like Gimpy, EZ-Gimpy, have been broken to date. Even the software is available in the markets which are able to recognize voice, resulting in breaking to voice CAPTCHA also. Moreover, CAPTCHA suffers from many problems:

- First, it is often very unethical as it unnecessarily discriminates against blind and otherwise visually impaired people. For the solution of this problem, many sites offer audio CAPTCHA as alternative. But the problem still remains because if the computers are trained to specific voice or samples, then computers can recognize voice as well.
- Secondly, CAPTCHA is not always very good at keeping spam away. The reason is that the computer software can recognize letters just like humans. For the solution of this problem, an attempt is made to make CAPTCHA harder, for example, by using low text-to-background contrast or bad color combination. But this could do nothing to stop bots, rather makes it harder to read for human. In other words, this result in a CAPTCHA that computer can recognize better than human [8].
- Thirdly, CAPTCHA turns away visitors, and may very well result in loss of revenue.

So the main question is “What could replace CAPTCHA that spammers would have a hard time defeating but at the same time not be too difficult for humans to decipher?” The alternative to the above mentioned problems is SAPTCHA. It stands for **Semi Automatic Public Turing Test to Tell Computers and Humans Apart**. The key concept is that here user is presented with test question or instructions and must give correct answer to that question in order to use a needful resource. These questions or phrases are clear and simple enough to answer without regard to the education level of the website user. These unique test questions on each query are not set by the computer, rather by moderator or owner when SAPTCHA is installed. Here only verification of the answer is automatic. Hence “SA” in SAPTCHA stands for “Semi Automatic” because setting questions is not automatic, but verification of the answer to the questions being asked is automatic. SAPTCHA is best alternative to CAPTCHA problems and works as lightweight CAPTCHA. Human generated questions have much broader diversity and are thus harder for computer to answer as spam-bots have an IQ somewhere around zero and need large question bank for this. So SAPTCHA could be solved only by the human brain and not by the machine, until it knows the answer to the question being asked. Hence SAPTCHA provide more security and protection to the resources available on the internet as compared to CAPTCHA. Example questions for SAPTCHA are:

What is the sum of three and thirty five?

If today is Saturday, what is the day after tomorrow?

Which of mango, table, and water is a fruit?

Any human brain could answer these questions easily, even without having good qualifications. In a way, SAPTCHA can be viewed as light weight disposable CAPTCHA test that is cheap to replace when it get compromised [9].

V. CAPTCHA vs. SAPTCHA

The following comparison of CAPTCHA with SAPTCHA shows that SAPTCHA works much better than that CAPTCHA in avoiding bots that access needful resources in World Wide Web.

In case of SAPTCHA, if a normal user comes across a blog, he can answer question, unless a bad question and/or instructions is made. If a spammer bots comes across a blog, then no spamming happens because the bots can't understand human language yet. And if a spammer human comes across a blog/forum, he can answer question, register account, and possibly can add answer and account to spam bots database or proceeds to spam manually. Here we are spammed and have to take action manually to ban spammer and stop spam like if spamming was done by bots that knows answer to question, the question could be changed.

On the contrary, in case of CAPTCHA, if a normal user comes across a blog/forum and if he can see, and CAPTCHA is simple he can post reply with small hassle. If CAPTCHA is "unbreakable" or uses bad colors, he will need several attempts, especially so if he need to pass it for every reply. If user is blind or otherwise can't see it, there is no way to get rid of CAPTCHA. If a spammer bots comes across a blog, then we might get spammed if bots can recognize images. If a spammer human comes across a blog/forum, he can recognize image and then also we are spammed [10].

Hence we can say that there are a lot of advantages of SAPTCHA over CAPTCHA:

- SAPTCHA software is much easier to implement or replace as compared to CAPTCHA software.
- Textual SAPTCHA does not discriminate against disabled who can use internet. Audio CAPTCHA plus visual CAPTCHA still discriminates against some people, plus the audio part is far easier for computer to break.

There is method for breaking image based CAPTCHAs. If some popular CAPTCHA is used, we may still get spammed by entirely automatic bot. On the other hand, SAPTCHAs can be much more varied and there won't be common method of breaking until it becomes possible for computers to interpret human instructions in normal human language.

Example of SAPTCHA question:

John had one thousand apples. He ate as many of his apples as there is letters in word "apple". How many apples John ate?

Other example:

In a mathematical forum, for example, it could be asked "What is the square root of minus one?"

Even though SAPTCHA will be more useful than CAPTCHA in securing and preventing misuse of resources available on the World Wide Web, but still there are some limitations of using SAPTCHA.

- With SAPTCHA, when banning spammer, moderator must enter new question and answer.
- If SAPTCHA is used to protect registration, it is easier to register many accounts at once.
- Verbal SAPTCHA may be problematic for multi-language resources that need frequent changes.

Even though there are some problems with SAPTCHA, it can be much better bots detector than CAPTCHA. The reason is in the way they are presented to the World Wide Web users. Any human brain could answer to the easy questions being asked, but for this the bots needs a large question bank as the IQ of the bots is zero [11].

VI. CONCLUSION

In the end, it can be concluded that the SAPTCHA could be viable alternative to CAPTCHA for web resources like forums and blogs and in other situations when spammer cannot afford to target resources individually. With textual resources, SAPTCHA does not lessen accessibility of resource to disabled users of World Wide Web. It is suggested that forum and blogging software should offer support for SAPTCHA in addition to existing support for CAPTCHA, thus allowing administrators to use SAPTCHA and switch to CAPTCHA only when SAPTCHA is found to be really inadequate in some situations.

VII. FUTURE DIRECTIONS

In future, more emphasis would be given to 3D CAPTCHA. That will be the next generation of CAPTCHAs. The idea behind these 3D CAPTCHA designs is that a human can recognize an object and manipulate it spatially in his mind. This is a step beyond character recognition/repetition and involves an additional level of cognition and understanding. The challenge is a 3D image of an animal, say of a rabbit's face. The list of answers would display different common animals from different angles, including a photo of the rabbit, this time of its side. Only a human brain would be able to quickly see that the challenge image and the second image on the answer list are of the same animal. In future, CAPTCHAs will be harder to read.

REFERENCES

1. Huang, S.Y., Lee, Y.K., Bell, G. Ou, Z.h. (2008) "A Projection-based Segmentation Algorithm for Breaking MSN and YAHOO CAPTCHAs", The 2008 International Conference of Signal and Image Engineering.
2. J. Yan and A. S. El Ahmad. "Is cheap labour behind the scene? -- Low-cost automated attacks on Yahoo CAPTCHAs", School of Computing Science Technical Report, Newcastle University, England, 2008.
3. Amelio, A.; Jankovi'c, R.; Taniki'c, D.; Draganov, I.R. Predicting the Usability of the Dice CAPTCHA via Artificial Neural Network. In Digital Libraries: Supporting Open Science, Proceedings of the 15th Italian

- Research Conference on Digital Libraries, Pisa, Italy, 31 January–1 February 2017; Manghi, P., Candela, L., Silvello, G., Eds.; Communications in Computer and Information Science; Springer: Berlin/Heidelberg, Germany, 2017; Volume 988, pp. 44–58.
4. C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. Beyah, and T. Wang, “Adversarial CAPTCHAs,” arXiv preprint arXiv:1901.01107, jan 2017.
 5. S. Pramanik, R. P. Singh, and R. Ghosh, “A new encrypted method in image steganography,” Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 3, p. 1412, 2017.
 6. Nicos Isaak and Loizos Michael. Using the Winograd Schema Challenge as a CAPTCHA. In Proceedings of the 4th Global Conference on Artificial Intelligence (GCAI 2017). EasyChair, 2017.
 7. Detchasit Pansa, Thawatchai Chomsiri, "Integrating the Dynamic Password Authentication with Possession Factor and CAPTCHA", Proc. 2017 Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, pp. 530-535, 5-8 December 2017, 2017.
 8. Jeff Yan , Ahmad Salah El Ahmad, Usability of CAPTCHAs or usability issues in CAPTCHA design, Proceedings of the 4th symposium on Usable privacy and security, July 23-25, 2008, Pittsburgh, Pennsylvania [doi>10.1145/1408664.1408671].
 9. Von Ahn L. Blum M., Hopper N. J. and Langford J., “CAPTCHA: Using Hard AI Problems for Security”, downloaded Nov. 2011 (18 pages).
 10. Priyadarshini, I., Cotton, C.: Internet memes: a novel approach to distinguish human and bot authentication. In: Advances in Intelligent Systems. Springer (2017).
 11. Haiqin Weng, Binbin Zhao, Shouling Ji, Jianhai Chen, Ting Wang, Qinming He, and Raheem Beyah. Towards understanding the security of modern image captchas and underground captcha-solving services. Big Data Mining and Analytics, 2:118–144, 06 2017.