

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Designing CAPTCHA for Web Security Using Rotating and Spilting Image against OCR Software

Durgesh Samariya

M.Tech (Software Technology)
VIT University,
Vellore, Tamil Nadu – 632014, India

Abstract: Nowadays internet is used for multiple activities by a huge number of people. These activities include communication, e-commerce, education, and entertainment. For many sites users needs to register to enroll website. However this websites are vulnerable. Hence hackers hack this software available for automatic entry. This software make false entry which occupies the resource of website hence reduces space, performance and efficiency of servers. There is one solution to this problem, use of a software that differentiates between human being and computer robot. The name of that software is Completely Automated Public Turing Test. CAPTCHA is designed in such a way that patterns it shows is easily done by humans but not by computers or Robot. But however lately it is seen that robot using advanced technologies such as Optical Character Recognition (OCR) can easily solve CAPTCHA. Hence this study is focused on new method of generating CAPTCHA. This paper proposes new way of designing CAPTCHA such as splitting image in many parts with random rotation value and adding different kind of random lines.

Keywords: CAPTCHA, Image CAPTCHA, Audio CAPTCHA, OCR, Prediction Algorithm, Web Security

I. INTRODUCTION

CAPTCHA was invented in 2000 by Luis Von Ahn, Manuel Blum, Nicholas J. Hooper and John Langford at CMU. CAPTCHA stands for **C**ompletely **A**utomated **P**ublic **T**uring **T**est to tell **C**omputers and **H**umans **A**part [1]. CAPTCHA follows a reverse Turing test in which CAPTCHA program acts like a judge and participant acts like a user. If the user passes the test, 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 humans 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:

- Human being easily pass through CAPTCHA.
- It should be easy for a tester machine to generate and grade.
- It should be hard for a software robot to pass.

Nowadays all work is mostly done on Internet likewise education, shopping (e-commerce), net banking, Social. On those site users need to fill out registration form by entering all personal information in order to surf that particular website and many kind of other survey form to surf website. Today due to increased technology hacking software is available to fill all details automatically. Hence attacker attacks by false entry on such site to increase traffic, occupy the resource of website reduce the performance and efficiency of server and some time it may stop the entire web service (web site).

CAPTCHA must provide mainly two requirements:

1. Robustness: Robustness is capability to resist or decrease attack on computer.
2. Usability: Usability is the easy for human who pass with this CAPTCHA challenge.

In this proposed work these all requirement are solve at the maximum level. A CAPTCHA is one kind of program that generate and grade test that mostly all human being pass but current computer program not able to pass.

1.1 TYPES OF CAPTCHA:

» Text Base CAPTCHA:

This comes from the most popular CAPTCHA Company out there, ReCAPTCHA. It is reliable, but some of the distorted word images are rather hard to solve. In order to solve that it allows you the option to “reCAPTCHA,” on clicking reCAPTCHA we receive a whole new pattern to solve. Especially for blind people they provided the facility of audio CAPTCHA, if the person is unable to visually make out the word.

Text based CAPTCHAs is very simple to implement. Also it is very effective and requires a large question bank. In Text based CAPTCHA the Number of classes of characters and digits are very small so the problem occurs for user to identify the correct characters and digits. The text based CAPTCHA is possible to identify the character and digit through Optical character recognition (OCR) technique. In Text based CAPTCHAs simple questions asked for example based on arithmetic equation are given below

- a) What is three plus two ($3+2=?$).
- b) What is six minus one ($6-1=?$).
- c) Which of cabbage, apple and table is vegetable?



Figure: 1 Example of text based CAPTCHA.

» Picture Based CAPTCHA :

This CAPTCHA by Picatcha provides the user with an elementary choice of choosing the correct image that they are asked to identify. They never get harder than basic images so you won't have to worry too much about your users not being able to depict the difference between them and the incorrect images.

Graphics-based CAPTCHAs are challenge-tests in which the users have to guess those images that have some similarity. For example: visual puzzles. In image based CAPTCHAs user is required to identify image. The advantage of image based CAPTCHA is that pattern recognition is hard AI problem and therefore it is difficult to break this test using pattern recognition technique. Example of images based CAPTCHA are given below.



Figure: 2 Example of Picture based CAPTCHA

» **CAPTCHA based on puzzle :**

Usually in puzzle based CAPTCHA a given picture is divided to chunks [2, 5]. A user is supposed to combine these chunks so as to form the complete picture same as the original one.

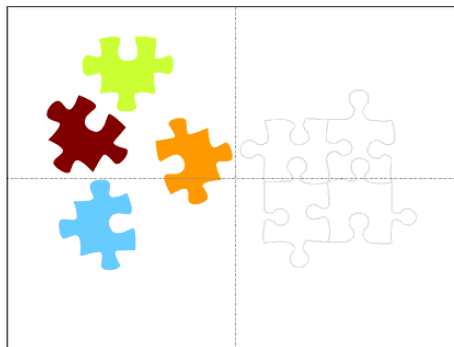


Figure: 3 Example of puzzle based CAPTCHA are given below.

» **Maths Solving CAPTCHA :**

if your users can't solve these basic math problems then maybe you don't want them commenting on your threads anyways. These provide you with easy to read numbers that must be added in order to get past the CAPTCHA.

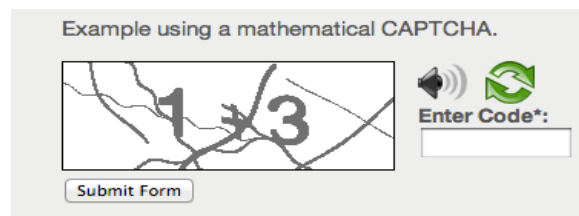


Figure: 4 Example of Math Solving CAPTCHA

» **Sound (Audio) Based CAPTCHA:**

Sound based CAPTCHAs are focused around the sound-based frameworks. These CAPTCHAs are produced for outwardly incapacitated clients. It contains downloadable sound cuts. In this kind of CAPTCHA, first the client listens and after that submits the talked word. The Nancy Chan an understudy from the City University in Hong Kong actualized the main sound-based framework name ECO. The sound built framework is situated in light of the distinction in the capacity between machine machines and people in perceiving talked dialect. The project picks a succession of digits and words haphazardly and renders the words and number digits into sound cuts and bends it. The contorted sound cut is then introduced to the client to enter the right word or number. The client is asked to enter precisely the same words as talked the sound cut.

This paper focused on text-based CAPTCHAs. The new CAPTCHA technique achieves high success rates for humans and low success rates for bots, require text entry.

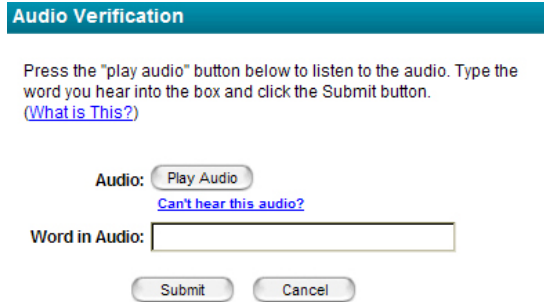


Figure: 5 Example of Sound Based CAPTCHA

This paper is organized as follows. Section 2 discusses related work. Section 3 describes algorithm. Section 5 describes conclusion.

II. RELATED WORK

CAPTCHA was first introduced in 1997 when Andrei Broder devised the CAPTCHA method. In the same year, AltaVista web site used this method to distinguish between computer programs and human user. In this method, a distorted English word was shown to the user and the user was asked to type it. Distortion was needed so that OCR programs could not recognize the word. As mention above CAPTCHA can be classified in two ways OCR based and non OCR based. After CAPTCHA evolved several methods are proposed.

Impressive number of studies was led via scientists on creating new CAPTCHA systems and breaking them. CAPTCHAs were initially created by Altavista to maintain a strategic distance from the accommodation of Urls to the web crawler [6]. It was a straightforward CAPTCHA, which asks clients to sort a mutilated English word.

Carnegie Mellon outlined the Gimpy strategy, which chooses a saying from word reference and asks clients to sort what they see as a picture in the wake of rendering the bended picture containing the content [7].

Yippee utilizes the basic adaptation of this technique; EZ- Gimpy. EZ-Gimpy's picture alteration incorporates foundation frameworks, slopes, non-direct misshapenness, obscuring, and pixel commotion. Most people can read three words from the twisted picture, while current machine projects cannot.

Pessimprint was produced in 2002 by PARC which utilizes the real shortcomings of OCR frameworks, for example, the powerlessness to perceive low quality pictures [8]. It contains just regular English words somewhere around five and eight characters in length. Pessimprint utilized just 70 words, which is low. Pessimprint's CAPTCHA would break with the likelihood of 1/70. In this way, this strategy does not succeed not surprisingly.

Hotmail is a free email benefit by the Microsoft Cooperation, and an alternate CAPTCHA technique is utilized [10]. A string of English characters is arbitrarily chosen, and in the wake of applying a few changes, clients are asked to sort what they see. The significant weakness of this system is a percentage of the characters are perused contrastingly in light of putting bends between characters [5].



Figure 6. Some CAPTCHA words of Yahoo [9], Hotmail [10] and Gmail [11] respectively.

Baffletext utilizes non-English pronounceable character strings to guard against lexicon assaults, and Gestalkt-persuaded picture veiling corruptions to safeguard against picture reclamation assaults [12].

The created word at times brought about challenges to human clients. There are likewise some other CAPTCHA techniques, which are utilizing picture or sound peculiarities to tell human clients and machines separated. PIX is a distinguishment strategy which utilizes regular pictures rather than pictures of words [1]. Be that as it may, this system obliges a huge measured space to store the pictures. In content to-discourse technique, again as opposed to demonstrating a picture, a sound is played for the clients and requested that they remember it [13]. In the wake of comprehension what the statement is by the sound, clients should sort it accurately to proceed with their methodology. Yet, comparable with PIX, this technique additionally obliges an incredible space and cost.

Inadequately actualized CAPTCHAs can be broken effectively even without utilizing character distinguishment programming. A portion of the original CAPTCHAs has as of now been broken, so the new era ought to be all the more influential and unpredictable to stay away from such assaults.

Some of the methods are discussed.

1. A.Krishnashanthi, Dr.K.Kuppasamy was developed and Envolving NEW CAPTCHA using LCG Algorithm and unpridictable Algorithm.They Generate CAPTCHA using this algorithm and provide security to web from bot.

2. Ahmad Salah El Ahmad, Jeff Yan, Lindsay Marshall was developed the robustness of a new CAPTCHA. In this method proposed the security of a new CAPTCHA that was deployed until very recently by Mega upload, a leading online storage and delivery website. The security of this scheme relies on a novel segmentation resistance mechanism.

III. PROPOSED SYSTEM

This paper proposes the IND-OCPA-P model to analyze the security of the proposed EOB and the encryption schemes supporting an efficient range query over encrypted data.

IV. SUGGESTED ALGORITHM

1. Start Session of site
2. Create 2 dimensional w*h rectangle.
Suppose take width = 20cm and height = 10cm
3. Add transparent background in rectangle. With w width and h height so its fit to rectangle.
4. Take random letter and number from
"abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789" and generate one n number String.

NOTE: But for selecting letters avoid some same type of letters so its not make any confusion to users. Like I/1, 0/O, Q/O, C/G, h/b After removing this letters, select string only from

"acdefghijklmnopqrstuvwxyzABCDEFHJKLMNPRSTUVWXYZ23456789"

Suppose we take one String --> dVikOI

5. That random String put in rectangle on transparent background. Make this string font color black only.



Figure: 7 Random String in Rectangle

6. Put this rectangle in session
7. Now give random rotation on that all letters and re-arrange in rectangle.
8. Now divide rectangle in $r \times c$ matrix.

R : Row, C: Column

Here taking $r = 3, c=3$ so rectangle splint into 9 parts.

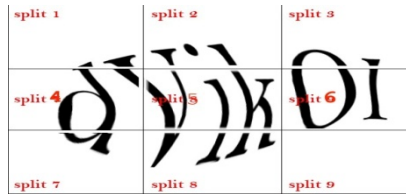


Figure: 8 Divide in 3×3 matrix

10. Insert one background that consist black color's n number of lines.

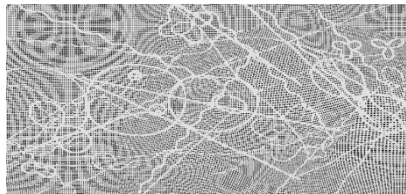


Figure: 9 Background Image

11. Insert this background in Rectangle

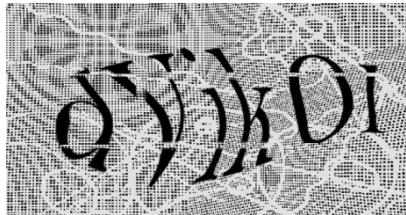


Figure: 10 Combination of Figure 8 & 9

It's a final CAPTCHA

12. Export as jpeg file name: "CAPTCHA.jpeg"
13. On refresh of CAPTCHA or web site destroy the CAPTCHA. Delete "CAPTCHA.jpeg" file.

More CAPTCHA with Different Parameters and with different size of split. This CAPTCHA for testing a algorithm with different different matrix size.



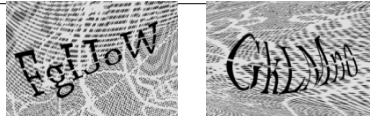


Figure: 11 more Example of CAPTCHA

V. CONCLUSION

In this paper a new technique is proposed for generating CAPTCHAs using combination of number and letters. Random number uses this algorithm to generate CAPTCHA and letters in new way and effective background its user friendly and secure. This algorithm makes more complicity for bot. There are some advantages of this algorithm A) Difficult for computerized bots to tackle and Easy to produce and assess for human. B) to produce characters as different irregular number generator calculations utilized which is more viable. C) In the planning standards different techniques to actualize which is exceptionally confounded for bots.

References

1. M. Tariq Bandy1, N. A. Shah2 by "A Study of CAPTCHAs for Securing Web Services"Vol. 1. No. 2, December 2009
2. Ritendra Datta, Jia Li, and James Z. Wang "IMAGINATION by A Robust Image-based CAPTCHA Generation System"-2005
3. Charles N. Zeeb and Patrick J. Burns by "Random Number Generator Recommendation".
4. Jeff Yan, Ahmad Salah El Ahmad by "Usability of CAPTCHAs Or usability issues in CAPTCHA design"-2008.
5. Ibrahim Furkan Ince, Ilker Yengin, Yucel Batu Salman, Hwan-Gue Cho, Tae-Cheon Yang1 by "DESIGNING CAPTCHA ALGORITHM: SPLITTING AND ROTATING THE IMAGES AGAINST OCRs"-Third 2008 International Conference on Convergence and Hybrid Information Technology.
6. Richard Chow, Philippe Golle, Markus Jakobsson, Lusha Wang, Xiaofeng Wang by "Making CAPTCHAs Clickable".
7. Elie Bursztein, Matthieu Martin, and John C. Mitchell by "Text-based CAPTCHA Strengths and Weaknesses-ACM Computer and Communication security" 2011.
8. Rituraj Soni, Devendra Tiwari by "Improved CAPTCHA Method 2010 International Journal of Computer Applications" Volume 1 – No. 25.
9. Yahoo! mail, <http://mail.yahoo.com>
10. Microsoft Hotmail, <http://www.hotmail.com>
11. Google Gmail, <http://mail.google.com>
12. Chew M. and Baird H. S., "BaffleText: a Human Interactive Proof", Proc of 10th SPIE/IS&T Document Recognition and Retrieval Conf. (DRR2003), Santa Clara,CA,2003, pp. 305-316.
13. Chan, T.Y., 2003, "Using a Text-to-Speech Synthesizer to Generate a Reverse Turing Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence.

AUTHOR(S) PROFILE



Durgesh Samariya was born in Nadiad, Gujarat, India in 23rd June. He is currently pursuing his M.Tech Degree in the Department of Software Technology in VIT University of Vellore, Tamil Nadu in India.