Preventing Character Recognition Attacks on CAPTCHA: A Customizable CAPTCHA Approach

Manish Kumar[1], Rajesh Shyam Singh [2] and Hardwari Lal Mandoria[3]

[1] Research Scholar Information Technology/ G.B.P.U.A.&T., Pantnagar-263145, India
[1]mnsh1403@gmail.com
[2] Asstt. Professor, Department of Information Technology/ G.B.P.U.A.&T., Pantnagar-263145, India
[2] rajeshsingh@gbpuat-tech.ac.in
[3] Professor & Head, Department of Information Technology/ G.B.P.U.A.&T., Pantnagar-263145, India
[3]drmandoria@gmail.com

ABSTRACT

CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart. Major challenge of an efficient CAPTCHA challenge is to develop a program which can create and grade challenges that most humans can pass but computers (bots) cannot. This document proposes a new CAPTCHA mechanism which is based on customizations on various security parameters such as distortion, transparency, character set, length and much more to finally develop a more secure alternative to the existing systems.

Keywords — CAPTCHA, Turing Test, Customized CAPTCHA, HIP, Cyber Security, Bot Prevention

1. INTRODUCTION

While using a Gmail Account or making payment using a PayPal account or maybe commenting on some well-known blogs, it is likely that you have come across CAPTCHAs in one form or another. The acronym stands for Completely Automated Public Turing-test to tell Computers and Humans Apart [1]. In simpler terms a CAPTCHA is a security mechanism that is used to prevent autonomous entries on websites (or prevent computerised bots) from doing unlawful activities on web pages.

Gmail improves its service by blocking access to automated spammers that perform auto sign-ups of Google Accounts to degrade its service, eBay improves its marketplace by blocking bots from flooding the website with scams, and similarly Facebook limits creation of fraudulent profiles used to spam and stalk honest users or cheat at games. [3] These are all real world examples of CAPTCHAs in action.

The widely used CAPTCHA schemes use combinations of distorted characters and obfuscation techniques that humans are able to recognize but it is difficult enough for automated scripts. CAPTCHAs are also called “Reverse Turing Tests”: because they are intended to allow a computer to determine if a remote client is human or not. [3] As time has passed bots have evolved into newer mechanisms to find loopholes in the existing CAPTCHA systems. Therefore, CAPTCHAS must be re-engineered time-to-time to prevent these attacks.

Some of the existing CAPTCHA security methods that were being used or are still in wide usage are discussed under next heading. These include the Gimpy, the ez-Gimpy, the MSN CAPTCHA, the Hotmail CAPTCHA, the Google reCAPTCHA. Although these CAPTCHA methods were made to be much secure but still almost all of their security has been compromised from time to time. Therefore, it is obvious that some design flaws were still present in these systems that were exploited.

2. EXAMPLES OF SOME POPULAR TEXT-BASED CAPTCHAs

2.1. The Gimpy CAPTCHA

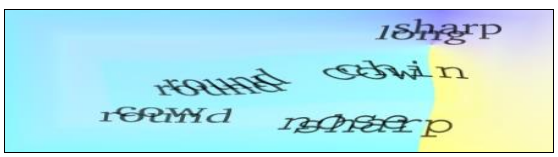

Fig.-1: Gimpy CAPTCHA [3]

GIMPY [1] is one of the many CAPTCHAs based on the difficulty of reading distorted text. GIMPY works by selecting seven words out of a dictionary and rendering a distorted image containing the words (as shown in "Fig. 1"). GIMPY then presents a test to its user, which consists of the distorted image and the directions: "type three words appearing in the image." Given the types of distortions that GIMPY uses, most humans can read three words from the distorted image, but current computer programs can't. The majority of CAPTCHAs used on the Web today are similar to GIMPY in that they rely on the difficulty of optical character recognition (the difficulty of reading distorted text).

Gimpy is a very reliable text CAPTCHA built by CMU in collaboration with Yahoo for their Messenger service. Gimpy is based on the human ability to read extremely distorted text and the inability of computer programs to do the same. Gimpy works by choosing ten words randomly from a dictionary, and displaying them in a distorted and overlapped manner. Gimpy then asks the users to enter a subset of the words in the image. The human user is capable of identifying the words correctly, whereas a computer program cannot [4].

2.2. The Ez-Gimpy CAPTCHA

Fig.-2: Ez-Gimpy CAPTCHA [3]

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 [5].

2.3. The MSN CAPTCHA

Fig.-3: MSN CAPTCHA [38]

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 [5].

Microsoft uses a different CAPTCHA for services provided under MSN umbrella. These are popularly called MSN Passport CAPTCHAs. They use eight characters (upper case) and digits. Foreground is dark blue, and background is grey. Warping is used to distort the characters, to produce a ripple effect, which makes computer recognition very difficult [4].

2.4. The Hotmail CAPTCHA

Fig.-4: Hotmail CAPTCHA [38]

It was used by Microsoft Corp. for its Live Hotmail service which consisted of distorted random alphabets and numbers.

The Google's reCAPTCHA

Fig.-5: Google's reCAPTCHA [2]

reCAPTCHA is a free service to protect your website from spam and abuse. reCAPTCHA uses an advanced risk analysis engine and adaptive CAPTCHAs to keep automated software from engaging in abusive activities on your site. It does this while letting your valid users pass through with ease [7].

reCAPTCHA offers more than just spam protection. Every time a CAPTCHA is solved the human effort incorporated is used in digitizing texts and books, annotate images, and build machine learning datasets. It improves our knowledge base of the physical world by creating CAPTCHAs out of text visible on Street View imagery. As people verify the text in these CAPTVCHAs, this information is used to make Google Maps more precise and complete. It digitizes books by turning words that cannot be read by computers into CAPTCHAs for people to solve. Word byword a book is digitized and preserved online for people to find and read. It also helps solve hard

problems in Artificial Intelligence. High quality human labelled images are compiled into datasets that can be used to train Machine Learning systems. Research communities benefit from such efforts that help build the next generation of ground-breaking Artificial Intelligence solutions [7].

For visually impaired, reCAPTCHA also provides an audio CAPTCHA option that the users find easy to use, bots on the other hand get a much harder audio CAPTCHA designed to block them [7].

3. ATTEMPTS AT BREAKING CAPTCHAS

Since its first appearance in 2000, CAPTCHAs have been subjected to continuous attacks that tend to compromise their efficiency. In spite of their wide spread usage, their extreme importance, and increasing number of research and studies there is yet no systematic methodology for designing or evaluating CAPTCHAs. Many studies have shown that most of the popular websites still depend on traditional unsecure techniques and schemes that are easily vulnerable to automated bot attacks. For example, it is even possible to use the fact that a given CAPTCHA is of a fixed length to make an educated guess so as where to apply segmentation, even if its anti-segmentation technique is impossible to break directly [6].

Various security measures have been proposed time to time covering a wide range of options for formulation of a robust CAPTCHA able to completely resist malicious attacks. Commonly and recently used methods are based on voice-processing, face recognition, complex approaches to recognize click patterns, mathematical and logical questions etc. However text-based systems have proved to be the most efficient way of detecting a bot due to their easy implementation and usability [6].

Let's assume you've protected an online form using a CAPTCHA that displays English words. The application warps the font slightly, stretching and bending the letters in unpredictable ways. In addition, the CAPTCHA includes a randomly generated background behind the word. [4]

A programmer wishing to break this CAPTCHA could approach the problem in phases. He or she would need to write an algorithm- a set of instructions that directs a machine to follow a certain series of steps. In this scenario, one step might be to convert the image in greyscale. That means the application removes the entire colour from the image, taking away one of the levels of obfuscation the CAPTCHA employs [4].

Next, the algorithm might tell the computer to detect patterns in the black and white image. The program compares each pattern to a normal letter, looking for matches. If the program can only match a few of the letters, it might cross reference those letters with a database of English words. Then it would plug in likely candidates into the submit field. This approach can be surprisingly effective. It might not work 100 percent of the time, but it can work often enough to be worthwhile to spammers [4].

4. USABILITY ISSUES IN CAPTCHAS

Sometimes while making a CAPTCHA image harder for bots also makes it difficult for humans too. Thus they tend to become less user friendly. For example, distortion can be a problem in a case if the letter ‘d’ is confused for ‘cl’. In the same way content becomes an issue when the word is not a dictionary word, then the user can’t predict the word and often misspells. Also care must be taken to prevent offensive words. Presentation should also be such that it is pleasant for the users. Good choice of colours and font-size and type should be taken. [4]

TABLE I.

Property	Issues Encountered
Distortion of images	Distortion algo used
	Level of distortion
	Characters confusing?
Content in CAPTCHA Image	Character Set used
	Random/Dictionary word
	Offensive words
	Word length
Look and Feel	Font Type
	Font Size
	Usage of Colours
	Compatibility with webpages

Table-1: Usability issues with text-based CAPTCHAs

5. PROPOSED WORK DESCRIPTION

The plus point of constructing a custom CAPTCHA is that the probability of getting attacked is greatly reduced. This is because spammers go for mass targets, as their success rate is quite low. Simply because we have a custom CAPTCHA we are less of a target. Keeping in mind these things the proposed system was designed that has the features mentioned below.

5.1. Features of Proposed System

- To provide a fully customizable interface.
- Ability to change the character set available for text-based CAPTCHA.
- Addition of special characters and symbols to make the CAPTCHA even stronger and tough to break.
- Ability to control the level of distortion on the CAPTCHA image letters.
- Ability to control the quantity of random lines to confuse bots and to prevent segmentation based attacks.
- Ability to select a random background image in order to make the image blurry and hard to recognize by optical character recognizers.
- Ability to change the theming and colour combination of captcha to match the webpage design.

5.2. Parameters and Their Usage

Backgrounds

A background image is added to the CAPTCHA that contains noise and random fill for example dots, splashes, bubbles or boxes in order to distort the image from being read by bots. A single background image can be used for every CAPTCHA that is generated or random background image can be loaded from a directory where all the usable background images are placed. This even enhances the security of the CAPTCHA as any recognizer would be unable to perform multiple attacks because each time the algorithm applied to break the CAPTCHA needs to be different because of different backgrounds.

Character Set

This is a very important property of the CAPTCHA. It decides whether which characters will be used to display the random CAPTCHA word. For example we can choose 0123456789 as the character set to only display a number based CAPTCHA or we can add alphabets and special symbols to it and make it even more difficult to solve. One more aspect of this parameter is to remove identical and confusing characters in order to ease it for humans. For example the characters I,i,1,l,j could be easily confused for one another. Therefore, they can be excluded from the character set to make it efficient. If we place a character more than once it means that its probability of appearing will be higher than others. Also lowercase and uppercase alphabets can both be added to the character set to add an extra level of security.

Difficulty Level

This is also an important aspect of the CAPTCHA it decides whether the image could be recognized by bots or not, but increasing the difficulty level to maximum also renders the CAPTCHA image unusable and harder to read even by human. Therefore careful selection of this parameter value is required as per need. The two parameters used in the CAPTCHA are the distortion of the CAPTCHA code itself and the number of lines that are drawn randomly over the CAPTCHA image to confuse bots. If the characters are distorted enough the bots will have hard time recognizing the characters correctly and the objective will be successfully achieved. Also drawing random lines greatly confuse bots because then their algorithm goes with the lines thinking of it as a letter or its part and thus failing to decide the right solution of the CAPTCHA.

Customizable Fonts

The usage of custom fonts can also make the CAPTCHA look attractive while adding an extra layer of security to it. Any ‘ttf’ font file can be used.

Colors

Usage of good colors in the CAPTCHA image makes it pleasant and easy for the users. The right choice of colors also helps in matching the CAPTCHA image to the theming of the rest of the webpage. This can be easily done with the ‘color’ attribute.

Image Size

This property is used to change the image size i.e. the height and width of the image in order to match the placeholder where the CAPTCHA image is to appear. The width must be 2.7 times the height of the image in order to keep the characters from flowing out of the area.

String Length

This property is used to vary the length of the CAPTCHA string to be displayed. For example it can be specified that the string generated should be 5 to 8 characters long. A lengthy string helps preventing guessing the word by bots but it also makes the users annoyed.

Signature

A signature as the name specifies is a small text image that appears on each of the CAPTCHA image. It is required in order to trace the usage of CAPTCHA images at other places. It can be our website name or anything else.

5.3. Generated CAPTCHAs using the Proposed Method

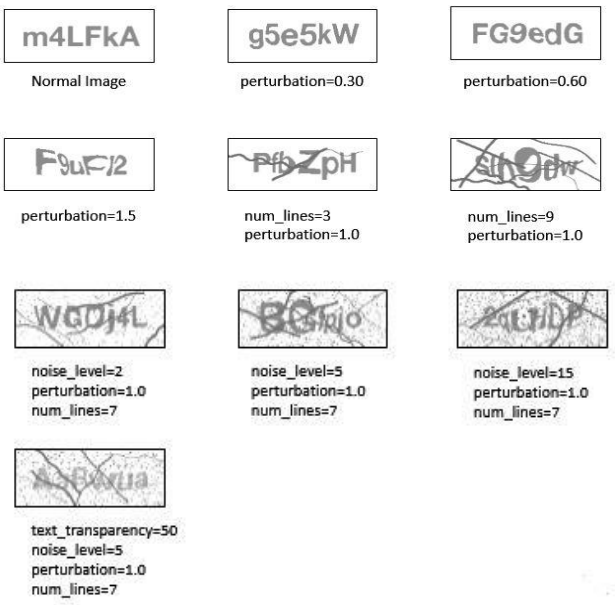

Fig.-6: Snapshots of Generated CAPTCHAs

6. CONCLUSIONS

This paper gives a description of various existing captcha schemes and provides a description for working of CAPTCHA. It also describes the classification of various text-based CAPTCHA schemes and their design flaws.

Form submission security is needed for both the provider and the consumer. No one likes spam or being spammed. At this point, it is clear that the CAPTCHA has run its course and some other security tech needs to step in. While other forms of spam/bot filtering have risen to the surface, none are yet used widely across the Internet and none currently have the backing to supplant the CAPTCHA.

CAPTCHAs are an effective way to counter bots and reduce spam. They serve dual purpose– first they help in advancing Artificial Intelligence Knowledge secondly they keep web data secure from intruders. Some of the issues that are being faced with current implementation of the CAPTCHA prove and represent the challenges for future improvements. Therefore, if we are able to serve the purpose of preventing attacks for the time-being it would serve the purpose for a long time.

REFERENCES

[1] Luis von Ahn, Manuel Blum, Nicholas J. Hopper and John Langford, "The official CAPTCHA website", www.captcha.net, 2000

[2] Greg Mori and Jitendra Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA", 2000

[3] Elie Bursztein, Matthieu Martin, John C. Mitchell, "Text-based CAPTCHA: Strengths and Weaknesses", ACM Computer and Communication Security, CSS'2011

[4] Sarika Choudhary, Ritika Saroha, Yatan Dahiya, Sachin Choudhary , "Understanding CAPTCHA: Text and Audio Based CAPTCHA with its Applications" International Journal of Advanced Research in Computer Science and Software Engineering, ISSN:2277 128X, Volume 3, Issue 6, June 2013

[5] Baljit Singh Saini, Anju Bala, "A Review of Bot Protection using CAPTCHA for Web Security" IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 8, Issue 6, Jan-Feb 2013

[6] Oleg Starostenko, Claudia Cruz-Perez, Fernando Uceda-Ponga, Vicente Alarcon-Aquino, "Pattern Recognition-Breaking text-based CAPTCHAs with variable word and character orientation", www.elsevier.com/locate/pr, 2014

[7] google.com/recaptcha

