Secure and Easy to Read Automated Handwritten CAPTCHA

Mukta Rao*
Dr Nipur*

Abstract

Time has changed, now humans have to pass a test against computer generated spam bots; called CAPTCHA (Completely Automated Public Turing test to tell Computers and Human Apart). The automated scripts, spam bots and robots of web portals can bring down a running website in few minutes. Most of the time the new account creation or online polling suffers from this issue. CAPTCHA is an automated test based on recognition tasks that are easy to pass for human and difficult for current computers to pass. In this paper the recognition difficulty of handwritten text by an OCR is used as a key for secure CAPTCHA formation. The use of numeric and case sensitive letters make the range of possible guesses computationally very large. The application of random image alteration techniques such as noise, color of letters, appearance of letters, length of word etc make the characters more difficult to recognize by current set of OCRs but still relatively easy and readable for a human to respond.

Keywords : Handwritter CAPTCHA, Web Security, OCR

1. Introduction

The idea behind CAPTCHA is more than a half century old. For the first Alan Turing explained a test in 1950 called Turing Test[1]. This was another formulation of his old question "Can Machine Think?". Going further on this Turing proposed a method to test the human v/s machine presence by interrogating both the parties, this interrogation should principally differentiate among them. In November 1999 slashdot.com released an online poll asking which is the best graduate school in computer science. As is the case with most online polls, IP addresses of voters were recorded to prevent single user from voting more than once. I order to get the prize students at Carnegie Mellon wrote a program that voted for CMU thousands of times to stuff the ballots. CMU's score grew very fast. The next day, graduation students at MIT got to know about this and they wrote their own voting program. Now the poll became a contest between voting "bots". The final scores were MIT-21,156 votes, Carnegie Mellon-21,032 and every other school was below one thousand. There arose a question "Can the result of any online poll be trusted?". Not unless the poll requires that only humans can vote. Since then HIP and its challenges have grown many folds. With this incident the Alan Turing's test took the shape of CAPTCHA and started to be utilized on webportals. Alta Vista website was among the first few websites to use CAPTCHAs[2].

*Gurukul Kangri Vishwavidyalaya, Haridwar, India

The tests which distinguishes human from computers are broadly referred as HIP(Human Interactive Proofs) and CAPTCHA is a sub category of HIPs.

One might be familiar with CAPTCHA from online ticket buying sites, signing up for a free email account, trying to reset your passwords or filling up an online enquiry form at any website. In most cases CAPTCHA implementation today presents an image to the user with character string or words in it. The image has some noise applied to it to make the characters appear difficult to read. With the increasing amount of research and kind of bots , there are several CAPTCH techniques available. Few of them are:

a. Natural language processing
Such CAPTCHA ask questions like: "What is 4 times the number of legs a kangaroo has?"

b. Character Recognition :
These are most widely used types of CAPTCHA. Sample are shown below in figure 1. Such CAPTCHA has question of the kind "Type the letters you see in this image."



Figure 1: Sample of Character Recognition CAPTCHA

c. Image understanding:
This category of CAPTCHA first displays an image (examples shown in Figure 2), then asks some questions like "What are these images of?" / "Is this image upright?"



Figure 2: Image Understanding CAPTCHA

d. Automatic speech recognition
Such CAPTCHA has an audio associated with it. One has to understand the text out of it and write in the text box. Some examples are : "1-6-3-9-2-7" / Old radio broadcasts

e. Mathematical

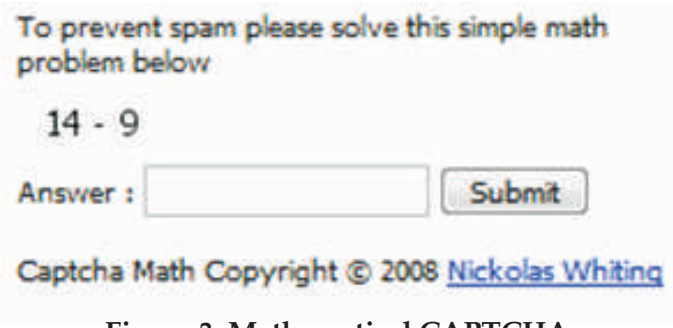

Figure 3: Mathematical CAPTCHA

2. Back Ground

Text based CAPTCHAs are so much popular that the word CAPTCHA has become a synonym for text based CAPTCHA. This paper works on text based CAPTCHA hence, unless otherwise stated the word CAPTCHA refers to text based CAPTCHA. Several CAPTCHA generation techniques are available on internet because the norms of CAPTCHA dictate that the method of generating CAPTCHA should be made public so that those who want to break them can get fair knowledge. This has lead to huge success of spam bots breaking CAPTCHA.

Manuel Blum's group at CMU designed a number of different CAPTCHAs. Some examples of CAPTCHA can be found at www.captcha.net. Variations can be seen at new account opening pages at yahoo, gmail, facebook, orkut, rediff, indiatimes, hotmail and many others. CAPTCHA systems (for example GIMPY, Bongo, Pix….) can be learnt from their respective weportals[3-4]. Few new challenges which are introduced by university of California, Berkley and PARC are more complex than earlier ones[5-8]. Not only the hackers but also the researchers and scientists are working in order to find solutions to break CAPTCHAs. Amit et al [9] use graphs with point feature to recognize deformable objects. Apparently, most of the research work in breaking CAPTCHA has been carried out by researchers working in handwriting recognition domain, but so far only printed text is visible on webportals in CAPTCHA. There has been successfully broken set of CAPTCHA even provided by Microsoft Office Word 2007[10, 11]. The main reason behind this is the increasing technology and excellent success ratio of OCR against printed text[12]. The 100% offline handwriting recognition is still a challenge for OCRs.

Thus use of handwritten text as CAPTCHA image is a revolutionary idea. The work proposed by V. Govind Raju [13] is the first step in this direction but the kind of complexity they have introduced makes the CAPTCHA even non-understandable for most of the users. The trade off between readability and security should be perfectly balanced. The introduced of noise of 5 different arbitrary types have though increased the security [13] but at the same time the readability has decreased resulting in less usability.

Here in this paper we have proposed a method of generating CAPTCHA image using handwritten text and association/ clustering features of artificial Neural Network.

3. Technical Implementation

The implementation of handwritten CAPTCHA generation is based on the following observations: -

1. All CAPTCHA images the authors have seen are made up of a alphabet no numbers.
2. Maximum CAPTCHA images consist of a single word and in most cases same word length.
3. If multiple words [eg-FB] then the words are in same horizontal line.
4. The content of CAPTCHA image is not a word from dictionary, this practically consumes time of the user to type thus become an annoying task. It's always easy and pleasant to type the words from dictionary rather than combination of arbitrary letter set.

Based on these observations we have created CAPTCHA those have: -

1. Numbers (Numerals/digits) in the image.
2. Image those contain two words.
3. The words of CAPTCHA image will necessarily not be in same horizontal line, they can be placed in any order based on a random function result.
4. The content of CAPTCHA shall be a word from English dictionary and a number of 4-7 digits. This would increase the recognition complexity but the user friendliness shall be maintained.

This section details the different stages of handwritten CAPTCHA generation viz. Sampling and Preprocessing, Clustering using ANN and CAPTCHA image generation. The answer sheets of the students at university examination have been scanned and used for sampling in this paper. Using the techniques described in [14] we have removed the noise and segrnented the word image into letters. Poor segmentation is discarded using HNN. The answer sheets of mathematics were scanned to get handwritten numerals and ideally there was no need for any post segmentation verification.

Clustering technique of ANN is used to create clusters. These major cluster are again divided into sub cluster to accommodate 26 letters of English alphabet set and 10 numeric digits.

The result of this process is a well arranged sample set that can be readily used for creation of CAPTCHA. The following algorithm is used in order to create a CAPTCHA images: -

Use a random function twice to get value out of {4, 5, 6, 7} resulting in size of CAPTCHA word.

1. Find a random number between 1- 10000 to get the index of dictionary word say w1.
2. Get the length of w1 and select the corresponding letters from clusters. Again use a random function to pick the letters.
3. At server side maintain a cyclic queue of size 10 containing color codes.
4. Now the array of letters contains colored images and ready to be grouped.
5. Selection of digits:- Add random type of noise as recognition of digits without noise is really easy
6. Use a binary value returning random function for horizontal lining or vertical stacking of the two words as well as their order of appearance.
7. Based on the outcome of the step-6 create CAPTHA.

Example Output:-



4. Test Result and Discussion

The samples for testing purposes were 2000 in numbers. These samples have been tested using three freely available online OCRs namely http://www.free-ocr.com referred as OCR-1, http://www.onlineocr.net referred as OCR-2 and http://www.newocr.com. The correct and incorrect recognition ratio of these three OCRs is provided in Table-1.

	OCR-1	OCR-2	OCR-3
Correct Recognition	126	79	154
Incorrect Recognition	1874 / **(93.5%)**	1921/ **(96.05%)**	1846/ **(92.3%)**

It is clearly evident that introduction of handwritten letters for CAPTCHA generation has brought good success rate for web security, though it has maintained the ease of recognition of text by human being. This shall certainly bring the attention of web developers to use handwritten CAPTCHA in place of machine printed fonts.

5. Conclusion

CAPTCHAs serve dual purpose- they stop robots to fill forms and they help to advance AI research. Although the test results indicate adequate success rates, but 100% accuracy is still a challenge. There are some issues with current implementations which represent challenges for future improvements. The authors are next working on introduction of Natural Language Processing with handwritten CAPTCHA to enhance security of web content. Considering the fact that user does not like to solve complex CAPTCHA and spammers are always there to crack simple CAPTCHAs, more research needs to be carried out in this area to secure web portals.

6. References

1. *A.M. Turing, Computing machinery and intelligence, Mind 59 (236), 1950, pages 433–460.*
2. *Altavista's add url site: altavista.com/sites/addurl/newurl*
3. *www.captcha.net/captchas/gimpy/*
4. *Ahn, L. von, Blum, M., Hopper, N.J., and Langford, J. The CAPTCHA Web page; www.captcha.net.*

5. *H. Baird and K. Popat. Human interactive proofs and document image analysis. Proc. IAPR 2002 Workshop on Document Analysis Systems, August 2002.*
6. *M. Chew and H. Baird. Baffletext: A human interactive proof. Proc. SPIE-IST Electronic Imaging, Document Recognition and Retrieval, pages 305–316, January 2003*
7. *A. Coates, H. Baird, and R. Fateman. Pessimal print: a reverse turing test. Proc. IAPR 6th International Conference on Document Analysis and Recognition, pages 1154–1158, September 2001*
8. *G. Mori and J. Malik. Breaking a visual captcha. Computer Vision and Pattern Recognition, 2003*
9. *Y. Amit and A. Kong. Graphical templates for model registration. IEEE Trans. PAMI, 2006*
10. *http://www.zdnet.com/blog/security/microsofts-captcha-successfully-broken/1232*
11. *http://www.spamfighter.com/News-10165-Spammers-Crack-Microsofts-CAPTCHA-to-Create-Live-Hotmail-Accounts.htm*
12. *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.3684/rep:Drep1/type:Dpdf*
13. *A. Rusu and V. Govindaraju, "Handwritten CAPTCHA: Using the Difference in the Abilities of Humans and Machines in Reading Handwritten Words". Proc. 9th IAPR International Workshop on Frontiers of Handwriting Recognition (IWFHR 2004), IEEE Computer Society, ISBN 0-7695-2187-8, pp. 226-231, 2004*
14. *G. Timar , K. Karacs and Csaba, "Analogical preprocessing and segmentation algorithm for offline hand writing recognition"; Journal Of Circuits Systems And Computers, Volume no.: 12, Issue no.: 6, Page: 783-804, 2003*