reCAPTCHA: Human-Based Character Recognition

[1]Prof. Prerna Kulkarni, [2]Sunil Gode, [3]Shraddha Patil, [4]Sonam Gaikwad

[1]Asst. Professor, [2,3,4]BE Student, [1,2,3,4]Comp. Engg. Dept, SSJCET, Asangaon, India.

[1]prernakulkarni09@gmail.com, [2]sunil.gode007@gmail.com, [3]patilshraddha384@gmail.com, [4]sonam.gaikwad24@gmail.com

***Abstract* — CAPTCHAs for stopping automated attacks and to reduce spam on websites. The idea is to use a puzzle, that only a humans can solve, but not automated programs. Particularly Audio CAPTCHAs, that ask a user to decipher distorted text in an Audio, are widely deployed on websites and users are accustomed to this type of CAPTCHA. reCAPTCHA is one of the biggest provider of text and Audio based CAPTCHA technology and is said to be one of the most secure. Current OCR methods work unreliably on distorted words from reCAPTCHA. The security of reCAPTCHA is analyzed in the presence of automated system solver capable of words. Specialized preprocessing is needed for one of reCAPTCHA's and proposed on the basis of a machine learning algorithm. Existing object recognition methods are modified and extended to work faster in a search space of about 23'000 words. It is experimentally shown that this method works well with all recent.**

***Keywords*— Automated attacks, Captcha, Machine learning algorithm, Traditional segmentation methods.**

I. INTRODUCTION

A CAPTCHA is an automated touring test, that can be taken over the internet and is used to differentiate humans from machines. The captcha stands for "Completely Automated Public Turing test to tell Computers and Humans Apart" and it was coined by Luis von Ahn et al. in the year 2000.A reCAPTCHA should be easy to solve by a human and should be difficult for a computer program. This is a reverse automated public Turing Test: a computer program decides if it deals with a human or a computer. CAPTCHAs have a variety of applications, mostly notably to help prevent spam. Bogus comments are being sub- mitted by automated bots on websites, spiteful users create thousands of accounts on free e-mail services to spread spam. Online-Polls are exploited by automated bots so that their outcome is manipulated. By using a CAPTCHA it is much harder to automate such tasks: if the website can dependable in achievement to detect that a computer program is using it and not a human, it can decide to deny its service to computer programs. Suitable unsolved Artificial Intelligence (AI) problems are used to build CAPTCHAs . The security of a CAPTCHA is based on the supposition, that the underlying AI problem is one that computers cannot yet perform. The recognition problem of characters and words from images under clutter and distortions is often used for CAPTCHAs. Even though OCR has a long tradition and predates electronic computers , the case that humans are considered to be significantly better at recognizing words, particularly when they are distorted and noisy.

The recognition of single an disolated characters on the contrary is seen as a solved AI problem and current recognition algorithms can have a better accuracy than humans . Segmentation into single characters is usually associated with OCR: if the characters of a word are only narrowly separated or not at all, it becomes very hard to segment them with an automated program. Studies have shown that computers are better at solving the recognition problem than the segmentation problem. reCAPTCHA is one of the most widely used CAPTCHAs on websites it is based on the word recognition problem.

CAPTCHAs are used because of the fact that it is difficult for the computers to extract the text from such a distorted image, whereas it is relatively easy for a human to understand the text hidden behind the distortions. Therefore, the correct response to a CAPTCHA challenge is assumed to come from a human and the user is permitted into the website. CAPTCHAs are short for Completely Automated Public Turing test to tell Computers and Humans Apart. The term "CAPTCHA" was coined in 2000 by Luis Von Ahn, Manuel Blum, Nicholas J. Hopper (all of Carnegie Mellon University, and John Langford (then of IBM). They are challenge-response tests to ensure that the users are indeed human. The purpose of a CAPTCHA is to block form submissions from spam bots –automated scripts that harvest email addresses from publicly available web forms. The reCAPTCHA system achieved an accuracy of 99.1% at the word level (216 errors out of 24,080 words), whereas the accuracy of standard OCR was only 83.5% (3976 errors). The percentage of words on which both OCR systems made a mistake was 7.3%. An accuracy of99.1% is within the acceptable "over 99%" industry standard guarantee for "key and verify "transcription techniques in which two professional human transcribers independently type the data and discrepancies are corrected As an anecdote ,the professional manual transcriptions of the articles that were collected as "ground truth" to measure the accuracy of reCAPTCHA originally contained 189 errors ,almost as many as those made by reCAPTCHA That automated program could be part of a larger attempt to send out spam mail to millions of people. The CAPTCHA test helps identify which users are real human beings and which ones are computer programs. Spammers are constantly trying to build algorithms that read the distorted text correctly. So strong CAPTCHAs have to be designed and built so that the efforts of the spammers are thwarted.

Fig I:- captcha image

II LITERATURE SURVEY

CAPTCHA system allows it, a small fraction of the tests can be solved by supplying the same solution to every challenge. The CAPTCHA is then solved at a trivial solving rate. if there are n possible solutions to the CAPTCHA and each solution is equally likely to be used as a challenge. It is widely agreed that a trivial solving rate for a good CAPTCHA should be less than 0.01% .An adversary At is trivial, only solve a fraction of the CAPTCHAs with less than or equal the trivial rate An adversary Ah is partially-human, if leverages the recognition of the CAPTCHA to a human. Such attacks exploit the first principle, that the CAPTCHA must be easily solvable by a human. An adversary As is a strong adversary, if attacks design-flaws in the third principle and is able to pass the CAPTCHA computationally at a solving rate significantly above the trivial one. It is widely agreed that a CAPTCHA is effectively broken if there exists an As that can solve the challenges at a rate higher than 5%. A commonly accepted goal for a good CAPTCHA is that an adversary (without being partially human) should not be able to achieve a success rate of higher than 0.01% for passing the challenges, but that the human success rate should be at least 90%.A text CAPTCHA shall be any CAPTCHA that uses the underlying AI-problems character recognition and segmentation. A text CAPTCHA is segmentation resistant, if it is difficult to segment the CAPTCHA into its individual characters with known character segmentation algorithms. Current research provides further evidence that if a text CAPTCHA can be segmented into its individual characters, it is effectively broken. Thus, to satisfy the third principle above, a good text CAPTCHA should be segmentation resistant. Note that there

is no way to prove segmentation resistance, it can only be shown empirically by testing certain OCR software s or segmentation algorithms. Furthermore a text CAPTCHA shall also be a word CAPTCHA, if I t uses words from natural language for its challenges.

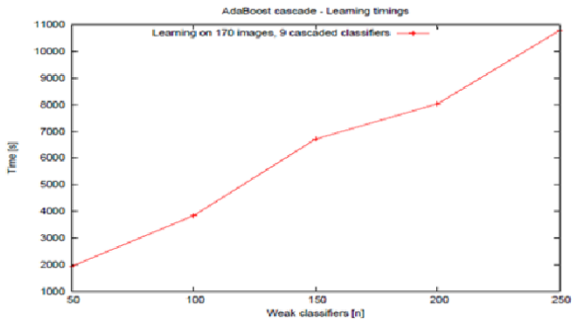
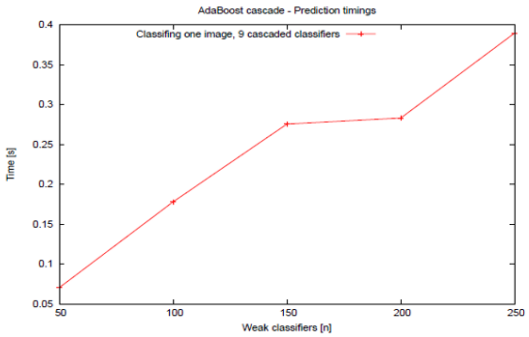


Fig A : Time it takes to train 9 cascaded classifiers for weak classifiers



Fig

B:Times it takes to use 9 cascaded classifiers for predicting one image

A) AIM AND OBJECTIVES

It stands for - Completely Automated Public Turing test to tell Computers and Humans Apart .A captcha is challenge response test used on the world wide web to determine whether a user is a human or a computer. reCAPTCHA puts this invaluable effort to serve humanity by helping digitize the books reCAPTCHA display words taken from scanned texts which could not be recognized by OCR.

CAPTCHAs come from a limited distribution of possible transformations which machine learning algorithms, after some training, can recognize the distorted characters. reCAPTCHA can be used at mail client of sender to prevent spamming. At social websites to prevent crawling. Success of reCAPTCHA report gave birth to duolingo.com

1. Free language education for the world
2. Simultaneously transforming web in various different
3. searching.

III PROPOSED SYSTEM

A) ARCHITECTURE

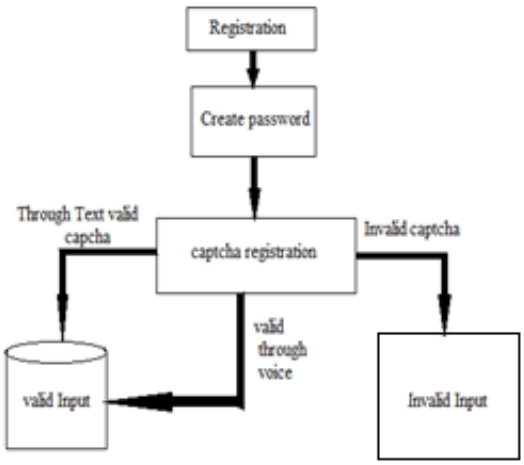


Fig A: - Proposed Architecture

ReCAPTCHA is one of the most widely used CAPTCHAs on websites. Up to this date, it is based on the word recognition problem. Words from scanned books and newspapers are used, most of them are older so that they are subject to an aging process that has degraded, smudged and distorted the words. It can also be misaligned by the scanning process and could be printed in a variety of type faces of which many could be rarely used today. The words used for the test have characters that are most of times not separated at all or leave very thin spaces between the characters. In addition, they are also distorted artificially to make the AI-problem of recognizing these words even harder. By typing in two of this distorted words correctly,

the user proves that he is human. The CAPTCHAs used by reCAPTCHA change from time to time.

B) ALGORITHM

Algorithm 1: An easy algorithm for approximating the center of the ellipse in the CAPTCHA

Input: Single word from a third generation reCAPTCHA as image

1 repeat
2 CAPTCHA ←*erode*(*CAPTCHA*)
3 until *CAPTCHA has a closed shape that is completely filled*
4 needed_dilates ←number of dilates until CAPTCHA is completely white
5CAPTCHA←*dilate needed_dilates*_2(*CAPTCHA*)
6 CAPTCHA ← *threshold*(*CAPTCHA*)
7 x ←*min* (*horizontal_projection*(*CAPTCHA*))
8 y ← *min* (*vertical_projection*(*CAPTCHA*))

Output: x, y : Estimated coordinates of the ellipse center.

Algorithm 3: Better-half-search algorithm using the distance function D_s

Input: img: Image of verification word (query shape), db: (local copy) Database of n shape contexts sets((*M*1,*M*_1, *P*1,_1), . . . , (*Mn*,*M_n*, *Pn*,_*n*)) for n words (*w*1, . . . ,*wn*), *ns*: start size of sample points, *fg*:growth factor of sample points in each iteration, *t f* : threshold value for the final stage, *nf*: sample points for the final stage.

1 (*Ms*, *Ps*, _*s*) generate Shape contexts(img)
2 sample points*ns*
3 repeat

- (*M*0*s* ,*M*0_*s*, *P*0*s* ,_0*s*) random Subset((*Ms*,*M_s*, *Ps*,_*s*), sample points)
- **For each**(*Md*,*M _d* , *Pd*,_*d*) *in db* **do**
- Distance *I* ← *DS*((*M*0*s* ,*M*0_*s*, *P*0*s* ,_0*s*), (*Md*,*M _d* , *Pd*,_*d*))
- **end**
- sort (*w*1, . . . ,*wn*) by (distance1, . . . ,distance *n*) in ascending order
- **for** *wi from* wd *n*=2e **9** *to wn* **do**
- delete(*wi*)
- **end**
- n d*n*=2e
- sample points d *fg*_ sample points e

4 until *n* < *t f*

Best Word NaïveSearch2(random Subset((*Ms*,*M_s*, *Ps*,_*s*), *nf*), db)

Algorithm 3: For Audio Captcha

1. **start**
2. register
3. validate through audio
   - listen number.
   - enter values after hearing number.
   - **If**
     valid then enter into database
     Else
     Error
4. **Repeat** step until valid input.
5. **Stop.**

C) MATHMATICAL MODEL

Let d be the dictionary of reCAPTCHA , that is defined as all unique solutions to all possible verification words, lower and upper case. If the distribution D of a random word from d is uniform. It has the probability 1/|d| of being chosen as a verification word for a CAPTCHA. It is interesting to know an estimate of |d|. for this, the number of collision after solving n CAPTCHAs can be used. A word is a collision to the previous n-1 CAPTCHAS, if they also contain the same word. the number of collision was 44. The probability of a collision for the n-th CAPTCHA is related to variation of the birthday problem.

The probability for this is:

$$q(n) = 1-(364/365)$$

The following generalized formula, on a set of size d:

$$q(n; d)= 1-(d-1/d)^n$$

The probability of a collision to at least one of the n-1 verification words after n CAPTCHAS have been solved is them given by q(n-1; d).The expected value of collisions after n verification words and a dictionary of size d can now be derived from this probability:

$$E = \sum_{i=0}^{n} q(k-1;d) = \sum_{i=0}^{n} (1-(d-1/d)^{k-1} = d\left(\frac{d-1}{d}\right)^n + n - d$$

a collision outcome that is near the expected value

of collisions and that the distribution *D* for reCAPTCHAs verification words is indeed uniform, *d* can be approximated by

$$44 = d\left(\frac{d-1}{d}\right) + 1932 - d$$

This formula can be solved numerically: *d* _ 41'749 words.

IV EXPECTED RESULT

We first allow the user to login in to the search portal where he/she after registration would be allowed to search for the desired web content. Once the user searches for the related web content, he would be provided the information. After fill the all information captcha will be applied. This would allow the user to Show that users human being or robots or computer programs.

Fig A: Expected output

V. CONCLUSION

It is been observed that reCAPTCHA is considered to be one of the most difficult text CAPTCHAs and justly so. Writing a software solver for it proved to be quite a challenging task. But the results presented here show that it is not impossible to build a strong adversary for reCAPTCHA. It will be impossible to increase the difficulty of text CAPTCHAs forever, as a reCAPTCHA should remain easily for humans.

REFERENCES

[1] AS El Ahmad, J Yan, and L Marshall. The robustness of a new CAPTCHA. Proceedings of the Third European Workshop on System Security, 2010. 2.2, 3.1

[2] L Von Ahn, B Maurer, C McMillen, D Abraham, and M Blum. recaptcha: Human-based character recognition via web security measures. Science, 2008. 1, 2, 2.2, 3, 4, 4.3.3, 4.3.4.

[3] U.H. Ai, L. Von Ahn, M. Blum, and J. Langford. CAPTCHA: Using Hard AI Problems For Security. In Proceedings of Eurocrypt, pages 294–311. Citeseer, 2003. 1, 3.

[4] M Alan. Turing. Computing machinery and intelligence. Mind, 1950. 1.

[5] HS Baird and TP Riopka. ScatterType: A reading CAPTCHA resistant to segmentation attack. Proc. SPIE, 2005. 3.

[6] DM Baxter. Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading. British Medical Journal, 1987. 3.2

[7] BBC. Spam weapon helps preserve books. 1

[8] S Belongie, G Mori, and J Malik. Matching with shape contexts. Statistics and Analysis of Shapes, 2006. 1, 3.1, 6.2, 6.3, 6.3.1, 6.3.2, 6.5.2

[9] Blackwidows. Breaking CAPTCHAs (as accessed on Jul 23 2010). 3

[10] Gary Bradski and Adrian Kaehler. Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, 2008. 5.3, 5.5

