Journal of Applied Sciences Research, 8(5): 2506-2509, 2012
ISSN 1819-544X
This is a refereed journal and all articles are professionally screened and reviewed

ORIGINAL ARTICLES

A de-CAPTCHA to show the vulnerabilities in CAPTCHA

[1]B. Thamotharan, [2]Dr. V.Vaithiyanathan, [3]R.Aparna

[1]Assistant Professor, School of Computing, SASTRA University, Thanjavur, Tamilnadu, India.
[2]Associate Dean-Research, School of Computing, SASTRA University, India
[3]Student, School of Computing, SASTRA University, Thanjavur, Tamilnadu, India.

ABSTRACT

A CAPTCHA an acronym for "completely automated public Turing test to tell computers and humans apart" is a challenge-response test which is used in web pages to determine whether the user is human. The CAPTCHA is mostly made of distorted image which might contain alpha-numeric letters in them. The CAPTCHA was brought up to avoid automatic form submission done by the spam bots. In this paper we are designing an Intelligence algorithm to showcase the vulnerabilities which are still available in them. The intelligence program created consist of three main modules namely noise removal, segmentation and classification. We go for noise removal as the CAPTCHA is present in the form of image. In the segmentation process we segment the characters present. During classification we compare the segmented characters using the templates created already and then print the output.

Key words: CAPTCHA, filters, segmentation, spam bots.

Introduction

Where ever submission is involved a CAPTCHA is been placed there in order to reduce the postings of spam bots, computer scientist came up with a concept called CAPTCHA, which effectively tests whether the user is a computer or a human. A CAPTCHA is enforced in a webpage in order to control the automated submission. Some websites use CAPTCHA to reduce the spams. A CAPTCHA is nothing but a set of alpha numeric characters arranged in random fashion which may or may not give meaning. CAPTCHA can take any form of color and orientation. It can also have further background color in order to improve the obfuscation. If the user is unable to understand the characters present in the CAPTCHA he/she can ask the server to generate next CAPTCHA characters.

These random texts can be interpreted by human easily but for a spam bot this is rather difficult.

Another important thing to be remembered about CAPTCHA is the limitation it has for the number of characters it can consist. The available CAPTCHA's can have a character length of a maximum of six or eight. As the number of character being less the complexity involved in breaking the CAPTCHA is comparatively less which also makes them vulnerable for the attack.

In this paper, the intention is to develop an intelligence program which can break the CAPTCHA image using a systematic process. The image at first goes through a filtering process which removes the extra additive noise present in the image or which is created by the server. Then the image undergoes a segmentation process in which each character is segmented. The segmented text is compared with the already created templates and the most possible match is found and is displayed as the output.

Materials and Methods

1 Proposed Work:

Even though CAPTCHA has been used to prevent the automated intervention there is upcoming threat towards this security. And it goes by the acronym OCR (optical character recognition). OCR is one which is capable of translating a typed text or generated text into machine encoded text. Hence when it comes to breaking strength OCR becomes a major threat factor in such case is high which may lead to more number of spammers and hackers(Matthieu Martin,2011).

Another important threat towards CAPTCHA is segmentation. Segmentation makes the CAPTCHA into a much simpler text and can easily convert into machine encoded text(Rich Gossweiler Google, Inc.1600.). This is

Corresponding Author: B. Thamotharan, Assistant Professor, School of Computing, SASTRA University, Thanjavur, Tamilnadu, India.
E-mail: balakrishthamo@gmail.com

where the vulnerability that is available in the CAPTCHA is shown out. This is done in three phases like noise removal, segmentation and classification.

2.Noise Removal:

As the CAPTCHA is an image it needs some preprocessing techniques to be applied which includes removal of noise. The CAPTCHA image is fed into Wiener filter for filtering of noise(Mukesh C. Motwani.). We go for Wiener filter as it best suits for digital image and also gives minimum mean square error value. Mathematically wiener filter is given as follows,

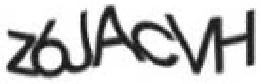
$$b(n_1,n_2)=\mu+\frac{\sigma^2-v^2}{\sigma^2}(a(n_1,n_2)-\mu),$$

Where μ is the local mean around each pixel,
σ^2 is the local variance around each pixel,
v^2 is the noise variance.

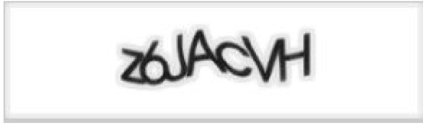
$$\mu=\frac{1}{NM}\sum_{n_1,n_2\in\eta} a(n_1,n_2)$$

$$\sigma^2=\frac{1}{NM}\sum_{n_1,n_2\in\eta} a^2(n_1,n_2)-\mu^2,$$

When the image is passed through the wiener filter the noise present in the CAPTCHA is removed.
The sample input and output of this phase is given below,



Input image



Output after noise removal [9 9]

The difference between the input and the output image is so evident. Even the noise created by the server or the noise created by the spill of character edge is been removed after the application of Wiener filter.

2.1 PSNR - Peak Signal to Noise Ratio:

PSNR is the ratio between the maximum power of signal to that of the noise(A.K.Jain,.1989). Unit of PSNR is in 'db'. The tabulation to various window size to PSNR is given below,

Table 1: After Wiener filter on same image with different window size

Window	PSNR
[3 3]	36.4119
[4 4]	33.5490
[5 5]	31.8791
[6 6]	30.0495
[7 7]	28.6260
[8 8]	27.1297
[9 9]	25.9801

3.Segmentation:

Segmentation is a process in which the digital image is partitioned into smaller individual pixels which helps in simplification of image into units which are easily analyzable(A.K.Jain,.1989) In case of CAPTCHA's we go for segmentation so as to remove the adjoining of characters from one another and for easy breaking of the characters.

Also all CAPTCHA's available online are present in RGB format in the first step of our segmentation process we convert the rgb image into gray image there by retaining the luminance and retaining the hue and saturation.

Then the gray scale image is then converted into binary image where the converted binary image has pixels values of 0 (black) for all pixels in the input image with luminance less than level and 1 (white) for all other pixels(Feng Ge, Song Wang Tiecheng Liu. 2007).

And now the enormous difference available with the pixel values the letters can be sensed. From the change noted in the pixel value the alphanumeric letters are traced and is kept ready to compare with the six set of alpha-numeric templates created earlier. These templates are created with the characters which are tilted to 5 and 10 degree to the obtuse and acute angle both with the numeric and alphabets of English in both upper and lower case.

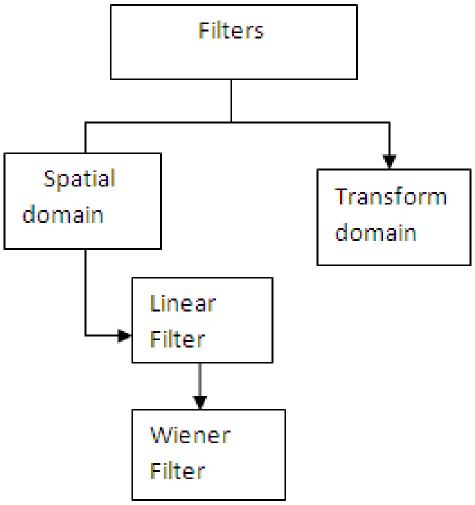


Fig. 1: showing the filter domain of wiener filter

4.Classification:

A template folder is already created which has the template characters of both numbers and alphabets. The traced letters segmented from the image are then compared with the templates. During comparison if there exits any occurrence of characters matching with the template characters then the most matching letter is taken as the output and is printed.

5.Algorithm:

1. Read the CAPTCHA image.
2. Apply filters to the image in order to remove the noise present.
3. Convert the noise removed RGB image into gray scale image.
4. Convert the gray scale image to binary image.
5. Trace the characters using pixel value of images.
6. Compare the traced characters with the template available.
7. Print the output.

Results and Conclusion:

This paper thus pinpoints how to break the CAPTCHA thereby showing the threat applicable to the millions of users who are using it in day to day online transactions. Thus by show causing the vulnerability it has in itself. This problem can be overcome by creating a strong, complex system of CAPTCHA thereby increasing the complexity to break them which would be tougher than ever before to reframe it.

References

Mukesh, C., Motwani., Mukesh C. Gadiya, Rakhi C. Motwani, Frederick C. Harris, A.K. Jain, 1989. “Fundamentals of digital image processing”,Prentice Hall.

Antoni Buades, Bartomeu Coll Jean Michel Morel “On image denoising methods “Exploiting the Human-Machine Gap in Image Recognition for Designing CAPTCHAs, 2009.

Feng Ge, Song Wang Tiecheng Liu. 2007 ”New benchmark for image segmentation valuation”,*Journal of Electronic Imaging* Jr., “Survey of Image Denoising Techniques”

Manuel Egele CAPTCHA Smuggling: Hijacking Web Browsing Sessions to Create CAPTCHA Farms,

Mathlab Works, Inc.

Matthieu Martin Stanford University, 2011”Text-based CAPTCHA Strengths and Weaknesses, Stanford University, 2011”.

What’s Up CAPTCHA? A CAPTCHA Based On Image Orientation, Rich Gossweiler Google, Inc.1600.