Propose Image Captcha System

Rajaa K .Hasoun,M.Sc. (Asst. Lecture) * Soukaena H. Hashem,Ph.D,(Asst. Prof.)*
Rehab F. Hasan,Ph.D,(Asst. Prof.)*

Abstract

Captcha has been designed to be easy for humans and hard on the machines. Captcha are used by many websites to determine human users from indecent programs. This paper will propose two types for generating captcha these are; text and image captcha. In the first type use (6*6) captcha table to store alphabetic characters 'A'..'Z' and '0'..'9' and use random function to generate six numbers to represent the row and colum numbers which used as index to retrieve the character from the captcha table to get the final captcha. In the second types a sample of 100 images from PayPal.com Human Interactive Proofs (HIP) are used. In order to recognize these images three steps for recognize the characters (pre-processing, segmentation and classification) are adopted. The modification was in the preprocessing step where proposing to use Gabor filter to remove the noise in the images which make the recognition accuracy 100% which is good results when compared with the obtained accuracy in reference [1] where the accuracy value in the range (80% .. 98%) for different scheme used in this work.

Keywords: image captcha, optical character recognition (OCR)

* University of Technology

1. Introduction

Stand for "Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs)" which are used by many websites to determine indecent programs from real human user . The most common type of CAPTCHA was first invented in 1997 by Mark D. Lillibridge, Martin Abadi, Krishna Bharat, and Andrei Z. Broder. The term was coined in 2003 by Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Captcha make the user pass through a simple test like reading characters or listen to speech and the user is asked to enter what they saw or heard. The sound or image is usually deformed in some ways in order to be hard for a machine to do and easy for humans. Figure (1) explain example of image captcha. CAPTCHAs are developed as a tools to limit the attackers ability to scale their activities using automated means. In most of the implementation, a CAPTCHA consists of alphanumeric characters which are distorted in such a way to be difficult to segmenting and recognizing the text. At the same time, humans, make some effort in order to decipher the text. CAPTCHAs of various kinds are deployed for guarding account registration, comment posting, and so on [1].

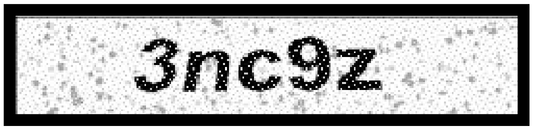


Figure (1) Example for image captcha

Captcha can be generated by using the following approaches [1] :

- Text: easy to crack but better than nothing, which include pick random letters and numbers and distort them in some ways.
- Image: image captcha is more complicated than text captcha because it is required optical character recognition in order to extract the text in that image.
- Speech: Spell in synthesized or recorded voices, required voice recognition.

- Animation: Can use Flash, MPEG, animated GIF. Often combined with speech. Not practical in most cases.
- 3-D: Renders the password in 3D image and more difficult to crack than 2D images, also need more resources on server.
- Combination of all above

In this paper will propose two type for generating CAPTCHA which are (text and image). In the first type use (6*6) captcha table to store alphabetic characters 'A'..'Z' and '0'..'9' and use random function to generate six numbers to represent the row and colum numbers which used as index to retrieve the character from the captcha table to get the final captcha. In the second types a sample of 100 images from PayPal.com Human Interactive Proofs (HIP) are used. In order to recognize these images three steps for recognize the character (pre-processing, segmentation and classification) are used. The remaining sections in this paper include the following: section 2provide over view for the related work. Section 3 explain the background on the OCR . evaluation of OCR accuracy is explained in section 4. Section 5 explained the general description of the proposal. Evaluation of image captcha is explained in section 6. Section 7 explain the conclusions of this paper.

2. Related Works

The following related work will present image captcha techniques as much as related to the proposal:

- **In 2012 , James et al. [2]** ; solve the problem of phishing by proposing new approach by using visual cryptography to divide original image captcha into two shares, this captcha can used as password when original image captcha is revealed to the user. When the shares generated from image captcha , one of them will be sent to the user while the other one will be stored in server site. On other hand for login to website, user must enter valid username, and then he get his share, the user share and server share will be stacked to generate the image captcha. The user must input the text that's appears in image captcha in order to login into the website. The proposal implemented in Matlab.

- **In 2013, Kaousar** [3] ; Propose prevention mechanism to avoid the phishing by using anti phishing image captcha and one time password to identify the fake web site, and used biometric (iris authentication) to authenticate user in addition to use the steganography to preserves the integrity. In this proposal user must enter username, phone number email id, date of birth and eye image is scanned, server generate unique pin code and user ID and stores these data with other details in the database in order to authenticate the user. The pin code variation is generated (combination of pin code and details of user) and this can be used to verify the website during login .
- **In 2014, Jose et al.** [4] ; suggest new method for anti-phishing by using authentication system using image based visual cryptography and to improve the security they use "Blowfish algorithm" to split original image captcha into many blocks and rearrange them , also they use split and rotate algorithm to rotate the rearranged blocks. This proposal provide three levels of security, first level verify if website is secure or not, second level image captcha can read only by human user and not by machine user, the third level it prevent attacks on the account of the user by intruders.
- **In 2015, Mahato et al.** [5] **;** gives an overview and introduction of the captcha based security system as a method for secure web based application, also explain the types of captcha, advantages and disadvantages of each type.

3. Optical Character Recognition (OCR)

In order to extract and match the text in the images from Pay pal.com HIP need OCR which is a method to convert images of printed text , typewritten or handwritten into understood format from the machines. Process of the OCR consists of three steps, which are "pre-processing, segmentation, and classification" [6] .

A. preprocessing

It is important step when image has a noise because of poor printers, bad scanners, etc., this step include remove the noise, and thresholding. Removing the noise performed by filtering the image. Thresholding is performed by making all intensity values greater than some threshold value to "on" and this is binarization process of an image [6] .

B. segmentation

It is operation of fraction the image into segments of a single entity. "Character - based segmentation", it is process of decomposition the image into sub images that have a single character [6] .

C. Classification

Classification of Character is highly dependent on the feature vectors that are taken away from the characters. Feature extraction is a method to transform the input data into a reduced representation. When calculate the feature vectors, classification can be complete. A technique that's used for classification is the nearest neighbor, neural networks, or other techniques. Many classifiers computed correlation coefficients. "The correlation coefficient" (Corr) for two vectors or matrices i and j, can be calculated in equation (1) [6] :

$$\frac{\sum_m \sum_n (i_{mn} - ii)(j_{mn} - jj)}{\sqrt{(\sum_m \sum_n (i_{mn} - ii)^2)(\sum_m \sum_n (j_{mn} - jj)^2)}} \ldots\ldots\ldots \quad (1)$$

Where

ii *is the mean of the input matrix* i *and* jj *is the mean of the input matrix* j.

m and n are the size dimension for two vectors or matrices.

4. Evaluation of OCR Accuracy

The correct word recognition is more important in text retrieval application. If m out of n words is recognized correctly, the word accuracy is found using the following equation:- [7] .

$$word\ accuracy = \frac{m}{n} \qquad \dots\dots\dots\dots\dots..(2)$$

Where m is the number of words that's recognized correctly and n is the total number of words.

5. The Proposal Description

In the following two sections will explain the proposed ways to generate the captcha which are (text captcha and image captcha). The first type (text captcha) consist of five characters results from (6*6) captcha table while the second type (image captcha) demand using samples of images and a way to recognize the text in these images.

5.1 The First Type: Text Captcha

The captcha that results from the proposed method consist of five characters. First initialize (6*6) captcha-table which contains the characters 'A'..'Z' and the numbers (0..9). Then generating two random numbers in the range(1..6), the first number represent row number (row) and the second represent colum number (col), so retrieve the character from the captcha-table that's located in (row, col), i.e. captcha-table (row, col) to get the first character of captcha, this process repeated five times to get the final captcha. Algorithm (1) explains how to generate the captcha and compare it with the string that's input from the user. Figure (2) shows the window to generate five characters using the text captcha algorithm.

Algorithm (1) Captcha Generation

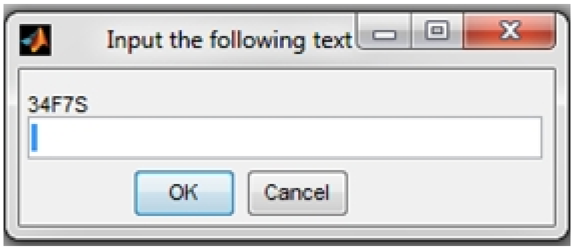
Input: string from the user
Output: Verify if human user or machine user

```
Process
      Initialize captcha-table (6*6) which contain alphabet ('A'..'Z') and
     (0..9)
  Captcha=''

  For i= 1 to length of captcha (5)
```

```
    Generate two random numbers (row and col) in the range (1..6)
      Captcha=captcha+ captcha-table(row, col)
 End
  Display captcha
   Read string from user
    If input string match the generated captcha
                       then " user is authenticated"
       else
           "not human user its machine user"

End process
```


Input the following text

34F7S

Figure (2) The text Captcha generation

The following example will explain how to generate text captcha of five characters. First initialize (6*6) array that's contains the characters from 'A' to 'Z' and numbers (0..9).

```
captcha(1,1)='A';captcha(1,2)='B';captcha(1,3)='C';captcha(1,4)='D';
captcha(1,5)='E'; captcha(1,6)='F';
captcha(2,1)='G'; captcha(2,2)='H'; catcha(2,3)='I';captcha(2,4)='J';
captcha(2,5)='K'; captcha(2,6)='L';
captcha(3,1)='M';captcha(3,2)='N';captcha(3,3)='O';captcha(3,4)='P';
captcha(3,5)='Q'; captcha(3,6)='R';
captcha(4,1)='S';captcha(4,2)='T';captcha(4,3)='U';captcha(4,4)='V';
captcha(4,5)='W';captcha(4,6)='X';
captcha(5,1)='Y';captcha(5,2)='Z';captcha(5,3)='0';captcha(5,4)='1';
captcha(5,5)='2';captcha(5,6)='3';
captcha(6,1)='4';captcha(6,2)='5';captcha(6,3)='6';captcha(6,4)='7';
captcha(6,5)='8';captcha(6,6)='9';
n = 6;
```

second step will be generating n random numbers in the range (1..n), let's consider the generating sequence number is (5,3,1,6,2,4), so this sequence will be used as index for row and Colum in array (captcha) as explain in the following paragraph:

```
captcha(5,3)='0'
captcha(3,1)='M'
captcha(1,6)='F'
captcha(6,2)='5'
captcha(2,4)='J'
```

So the generated captcha will be **"0MF5J"**

5.2 The Second Type :Image Captcha

Using image captcha is more complicated than text captcha because in order to distinguish the text in the image required using OCR. A sample of 100 images from PayPal.com using a Human Interactive Proofs (HIP), an approach for automatically distinguish between human and machines on the internet. Figure (3) show samples of these images.

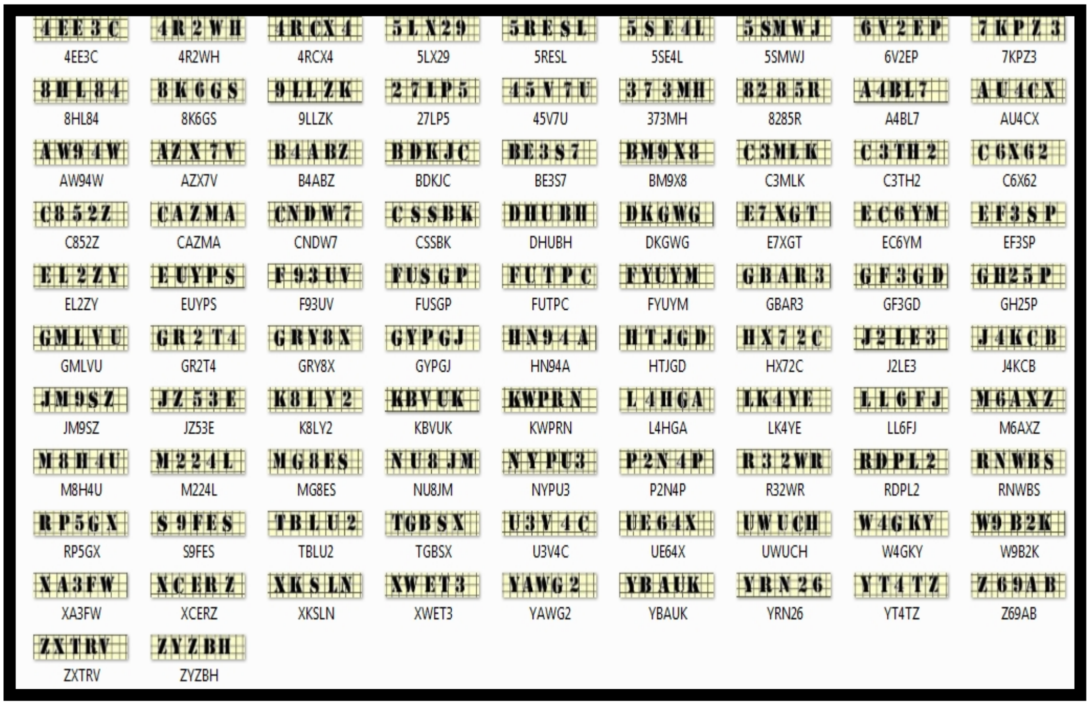


Figure (3) Samples of Pay Pal.com images captcha

First of all enrollment process will done for all 100 entries in order to store template file for each image captcha . The proposed method in this type will generate random number in the range (1..100) as an image index and display it. The Optical Character Recognition (OCR) with the three main steps (pre-processing, segmentation and classification as described in section (3) will be used to match between the image captcha and the string entered by the user.

Algorithm (2) explains the main algorithm to display and recognize the captcha.

```
Algorithm ( 2 )  PayPal Captcha Selection

Input: string from the user
Output: Verify if human user or machine user

Process
  Generate random number in the range  k=(1..100)
  Select the kth image captcha and display it to the user
  read string from the user
 recognize image=Optical Character Recognition (image(k))
    If user string match recognize image that result from OCR
         Then display "authenticated user"
           Else
             Display " not human it is machine user"
End process
```

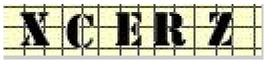
The first step in OCR process the pre-processing as explained in algorithm (3); which convert input image into grayscale image and then applies thresholding process to remove horizontal and vertical lines. But additional noise still remains, so Gabor filter will be applied to remove these noise

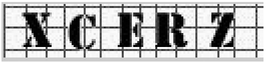
```
Algorithm (3)  pre-process of input image

Input: Input image
Output: preprocessed  image

Process
     1.  Convert the color image to gray scale
     2.  Threshold out the background noise
     3.  Place a bounding box around the image
     4.  delete any random noise by using Gabor filter
  End process
```

The following example explain the steps of pre-processing algorithm

1- Input image: 

2- Gray scale image: 

3- Thresholding :

4- After bounding:

5- Gabor filter :

The next step of (OCR) is determine segmentation boundaries by using vertical projection and candidate split positions as explained in algorithm (4). Empirically each character is of wide ten pixels at least . From the left to the right side of the image do column wise scan. When detecting the beginning of a character, skipped ten pixels and continue with the scan. When examined the end of character , cropped out this character from the image. Finally padded out using 0's to a fixed size (20×20) as a requirement for the classification process (correlation demands agreement of the two matrices dimensions). Repeating this process until the end of the image is reached. The following example explains the segmentation process.

1- Segmented:

2- Padded:

Algorithm (4) segmentation of preprocessed image

Input: pre-processed image
Output: segmented set of characters

```
Process
  Get the size of the input image
  While not end of image
      Scan forward while columns contain data
```

```
        Increment columns counter
     End
   Output  character out of the image
   Add pad after the last element
   Save this character into the returned value
  end
  End process
```

The third step of (OCR) is classification. The segmenter will input the five separated characters to the classifier. The input images must be binary images, The classification procedures as explained in algorithm (5) take the unknown sample image I and the set of template images T. The correlation coefficient (Corr) which explained in equation (1) is computed between two input vectors or matrices. The appearance of image captcha is shown in figure (4).

```
Algorithm (5)  Classification of segmented image

Input: segmented  image (I), Template (T)
Output: classified string of characters

Process
  K=0
  For each template $t_i$ belong to  T
     $K_i$=correlation ($t_i$ ,I)
  End
   Return j such that $K_j$=max(K)
 End process
```

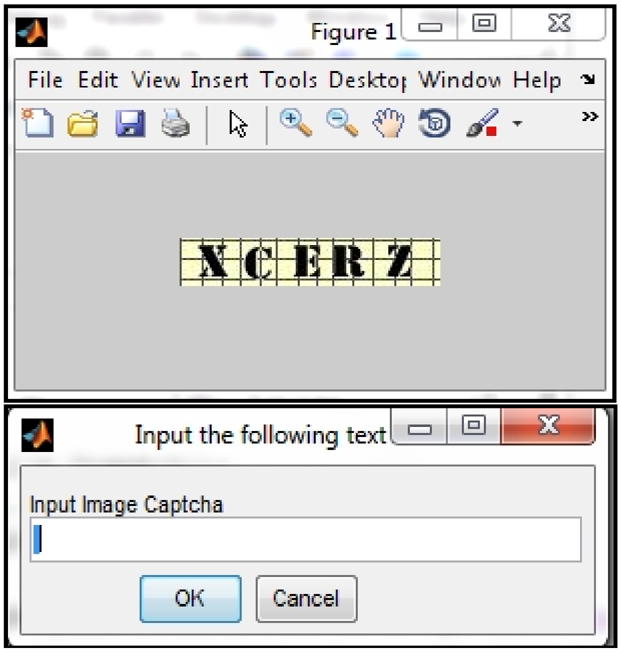


Figure (4) The appearance of image captcha

6. Evaluation the result of image captcha

- **Text type**

Because of the generated captcha is a text , so the matching process required comparing the two text only (the text in image captcha and the text that’s entered from the user) and doesn't required any training or testing process , so when apply a test to check number of misrecognized words, it's clear was zero and word accuracy is 100%.

- **Image type**

To evaluate the OCR accuracy use 20 samples for training purpose and 100 samples for test purpose. The displayed image captcha in this type selected randomly from image dataset. Table (1) show the accuracy test results for number of samples. From this table it can be seen that the word accuracy 100% for all samples because of recognition process was successful for all samples where the use of Gabor filter to remove the noise in the preprocessing steps Play an important role in the success of the process. Table (2) show the result of accuracy and time for different scheme in reference [1] . It is clear that the accuracy for different scheme between (80% and 98%).

Table (1) Word Accuracy for image captcha

No. of samples	No. of Misrecognized words	Word accuracy (%)
30	Zero	100
50	Zero	100
70	Zero	100

Table (2) Accuracy for image captcha for different scheme [1]

Scheme	Time	Accuracy %	Expected time
Authorize	6.8	98	6.9
Baidu	7.1	93	7.6
Captchas.net	8.2	84	9.8
eBay	7.3	93	7.8
Google	9.7	86	11.3
Microsoft	13	80	16.3
Yahoo	10.6	88	12

7. Conclusions

A two type of captcha were proposed: "text and image captcha" with a strong way to automatically recognizing the character strings in the PayPal.com HIP which contain 100 images using a three steps for recognition (preprocess, segment, and classify algorithms). The preprocessing is an important step, there are Gabor filter is applied to remove the noise, and the correlation coefficient was used as classifier. The accuracy of recognize the text captcha and pay pal.com HIP image captcha are 100%. This mean that using Gabor filter in preprocessing step Play an important role in the success of the process which make the image pure from any noise and this contributed to the success of the segmentation and classification process.

References

1- Bursztein E., Bethard S., Fabry C., Mitchell J.C. And Jurafsky D., "How Good Are Humans At Solving Captchas? A Large Scale Evaluation", IEEE Symposium on Security and Privacy, 2010.

2-James D. and Philip M.,"A Novel Anti Phishing Framework Based On Visual Cryptography", International Journal of Distributed and Parallel System, Vol.3, No. 1, January2012.

3- Kaousar K. M. A., "Trio Framework For Secure Online Transaction Using Visual Cryptography", International Journal Of scientific And Research Publication, Vol. 3, Issue 5, May2013.

4- Jose A. and Lakshmi V., "Web Security Using Visual Cryptography Against Phishing", Middle-East Journal Of Scientific Research, Vol. 20, Issue 12, 2014.

5- Mahato S., Saxena V. P., and Mishra R.G., " Securing Web Services and Applications using Captcha Security", HCTL Open International Journal of Technology Innovations and Research (IJTIR), Volume 14, April 2015.

6- Kluever K. A., "Braking the Pay Pal HIP: A Comparison of Classifiers", Article to Rochester Institute of Technology RIT Scholar works, 2008, http://scholarworks.rit.edu/article.

7- Rice S. V., Kanai J., and Thomas A. N. "An Evaluation of OCR Accuracy", Information Science Research Institute Annual Research, 1993.

اقتراح نظام صورة كلمة التحقق

م.م. رجاء كاظم حسون * أ.م.سكينة حسن هاشم * أ.م. رحاب فليح حسن *

المستخلص

كلمة التحقق صممت لتكون سهلة للبشر لكن صعبة للمكائن. كلمة التحقق تستخدم من قبل المواقع لتحديد البرامج غير اللائقة من المستخدمين البشر الحقيقيين . في هذا البحث سوف نقترح طريقتين لتوليد صورة كلمة التحقق: النص والصورة. في النوع الاول تم استخدام جدول (6*6) لخزن الاحرف والأرقام وكذلك تم استخدام دالة عشوائية لتوليد ستة ارقام تمثل ارقام الصفوف والأعمدة وتستخدم كمؤشر لاسترجاع الاحرف من الجدول للحصول على كلمة التحقق النهائية. في النوع الثاني تم استخدام عينة تتكون من 100 صورة المستعملة في موقع PayPal.com , ثلاث خطوات لتمييز الاحرف تم استخدامها، وهي (المعالجة الاولية ، التقطيع والتصنيف). التحديث تم في خطوة المعالجة الاولية حث تم استخدام فلتر كابور لإزالة الضوضاء من الصور مما جعل دقة التمييز تكون 100% وهي نتائج جيدة عند مقارنتها مع النتائج في المصدر رقم [1] حيث ان قيمة الدقة تتراوح بين 80% الى 98% .

الكلمات المفتاحية: صورة كلمة التحقق, تمييز الاحرف البصري.

* الجامعة التكنولوجية