A New Approach for Labeling Images Using CAPTCHA in Image Semantic Search Engines

Amir Golchini, Abolfazl Toroghi Haghighat and Hasan Rashidi

**Department of Electrical, Computer and IT Engineering, Qazvin Branch, Islamic Azad University
Qazvin, Iran**
a.golchini@qiau.ac.ir, at_haghighat@yahoo.com, hrashi@gmail.com

Abstract

In this paper, we present PICTCHA as a CAPTCHA system (Completely Automated Public Turing test to tell Computers and Humans Apart). This system is a security tool for recognition of human users and instead of a complex text, it uses image labels as a valuable output. PICTCHA is made by Web images. In order to verify this CAPTCHA, users are obliged to enter two words for labeling (naming) a couple of images. In case of convenient names for the images, the presented meaningful names are used for specifying the content of images. Furthermore, meaningful graphics are created for names and images according which we may develop an image semantic search engine. Due to benefiting from images in the proposed system and its architecture, it has higher security level in comparison with other rivals. In experiments with 60 participants, the correctness of PICTCHA words was %98.18 while about %61.26 users verified this challenge successfully.

Keywords: CAPTCHA, Ontology, Image labeling, Security

1. Introduction

Security has been changed into an important aspect of Web within recent years. Most of banking transfers and enrolments are made on web. As a result, web managers are challenging to keep their web sites against harmful attacks. One of the most common and wide security functions intends to prevent from attacking of buts and starting files. It is named as CAPTCHA system. E-mail suppliers used CAPTCHAs in their enrolment forms while weblogs used the same for maintenance of automatic programs against any spam as well.

There are various CAPTCHA systems but most of them are facing with security problems. At present time, re-CAPTCHA [1] is the most resistance system against any attacks. In spite of the mentioned concept, it was hacked by researchers of Stanford University [2]. The proposed system is named as PICTCHA in order to solve this problem by the use of images instead of complex contents. Rather to security, all presented names by users for verifying security paths, it is possible to use WEB for determining the content of made images. It is similar to re-CAPTCHA process by benefiting from words for digitalization of books.

The remaining parts of this paper are structured as follows. Second part is related works. PICTCHA is defined in details in third part. Fourth part is about relevant tests and obtained results. Finally we have relevant conclusion in 5th part as well.

2. Related Works

CAPTCHA [3] is a program designed based upon Turing automatic test [4]. Alan Turing (1950) presented this test for testing the ability of machine and displaying talent behavior. If a person outside a room is unable to recognize machine just in accordance with content deals and their responds, he/she has successfully verified the exam.

Followings are the most effective and fames CAPTCHA systems and their specifications. Finally there is a comparison between PICTCHA and previous systems.

2.1 A review on CAPTCHA projects

According to Turing test [4], CAPTCHAS create various challenges which are not easily created by computers. Such a challenge includes a superficial intelligent like processing of natural language, recognition of character, recognition of speech and image understanding. Therefore there are various types of CAPTCHA systems accordingly. Moni Naor (1997), presented the first idea of separation mankind from computer according to Turing test[4]. Of course it did not publish no more. His handwritten document includes various ideas and theories which finally resulted in creation of CAPTCHAs. The first image sample of Turing Automatic test was a system created by AltaVisa and by the use of complex images which may prevent from automatic enrolments on web pages by bots.

Luis Von Ahn (2000) introduced CAPTCHA [3]. Security challenges of this system obliged the user to solve them in order to verify the page. It means some random and complex characters made by a computer program. Figure 1(a) illustrates a sample of this CAPTCHA.

Von Ahn presented reCAPTCHA [5] in 2007 with this motto that: “Read a book, Stop Spam”. As it is obvious in figure 1(d), reCAPTCHAs challenges include two words one of them is a complex content with an obvious reply and the other is scanned from a physical book with lack of recognizing program with OCR character. Through digitalization process of physical books it is possible to increase searching facility and reducing required resources for reservation or transfer of them. When a user tries to solve a challenge and if there is a correct word for relevant image, it is assumed that the reply is also correct. If the users enter the same variety for an unknown word, the system will be ensured about their reply as well. The effectiveness of this method has been proved with %99 of insurance.

Jeremy Elson (2007) has presented ASSIRA (Animal Species Image Recognition for Restricting Access) Image CAPTCHA. In this CAPTCHA, there are 12 images from among a database with more than 3 million of photos. Then the user is obliged to recognize the image of cats or dogs [6] from among 12 images. Figure 1(e) illustrates a sample of this CAPTCHA.

In another research, a company (2008) presented NuCAPTCHA as a moving CAPTCHA [7]. Figure 1(b) illustrates a sample of which as well. In this system all characters are moving and therefore in order to solve the challenge, user is obliged to type moving characters.

Kluever & Zanibbi (2008) introduced Video CAPTCHA (video CAPTCHA) [8]. This CAPTCHA uses a social video of users on Web. There is a video in this system through which the use is obliged to explain 3 words in order verify a video CAPTCHA page.

2.2 Characteristics of projects in comparison with PICTCHA specifications

All CAPTCHAs follow up a common goal which is providing required conditions for prevention from misuse of bots and automatic starting files in special web pages. By the way, all CAPTCHAS have different specifications and properties. Hereinafter all introduced CAPTCHAS are compared in accordance with 5 important criteria including security, value added, easy application, bandwidth and item counting.

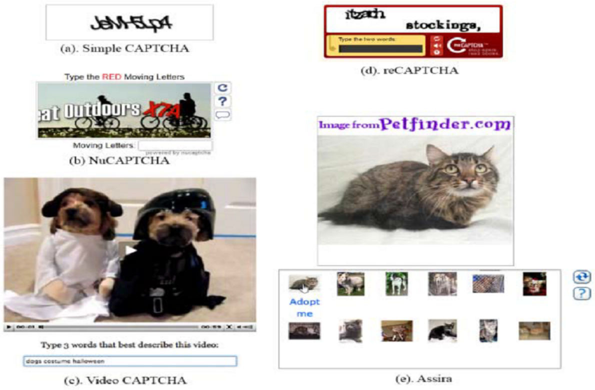

Fig. 1 Various samples of CAPTCHA: (a) a simple CAPTCHA, (b)NuCAPTCHA, (c)VideoCAPTCHA, (d)ReCAPTCHA, (e)Assira

2.2.1 Security

A research team of Stanford University designed and developed DeCAPTCHA for attacking CAPTCHAs of famous Websites including Wikipedia, IB, CNN and so on. This tool is used on 15 websites. Table 1 illustrates its success rate.

CAPTCHA security recommends using black & white characters. In addition, any applying of complex lines on characters may facilitate any prevention from turning of CAPTCHA system. DeCAPTCHA team used the same characteristic at Stanford University for attacking it successfully.

Moving CAPTCHA was hacked by DeCAPTCHA. There are five phases for breaking algorithm of this CAPTCHA. First phase is obliged to extract current frames in relevant animation. The background is omitted in second phase and we have white color words in a black background. Third phase includes extracted frames for determining relevant location of characters. All relevant characters of CAPTCHA are extracted at fourth step. Then a machine learning algorithm may extract all these characters in 5th step [7].

Table 1: Success rate of DeCAPTCHA in various CAPTCHAs

Success Rate	Web Site
1-10 %	Baidu, skyrock
10-24 %	CNN, Digg
25-49 %	eBay, Reddit, Slashdot, Wikipedia
50% or Greater	Authorize, Blizzard, Captcha.Net, MegaUpload, NIH

Any short usage of video CAPTCHAs is related to low function and lack of security. Theoretically, it is possible to turn these CAPTCHAs by the use of image compliance algorithms including SIFT (Scale Invariant Feature Transform) [8].

Finally, there is an important issue for those CAPTCHAs which are used commonly in Web as follows:

- These CAPTCHAs are hacked by most robots. Therefore the major goal is maintenance of websites against any attacks of robots.
- Of course there are lots of benefits out of solving these CAPTCHAs, but they do not have any value added. Even there is not any value added in reCAPTCHA in which we use information for digitalizing of books.

Security is a general and important issue as discussed in all forms of CAPTCHAs. Although reCAPTCHA is one of the safest CAPTCHA, but it has been also hacked. In other words, it is easy to attack any forms of content-based CAPTCHA including reCAPTCHA by the use of optimized versions of OCR algorithms which are present right now. By the way, since there are two different images for the user in a PICTCHA for further labeling and since it is difficult to have automatic image labeling, it is possible to consider a PICTCHA as the most security attitude in this regard.

2.2.2 Value added & Useful output

Except for reCAPTCHA, none of discussed CAPTCHAs have value added. By the way, even a reCAPTCHA is facing with this issue that it is unable to have any priority. Since the word is extracted from e-book without more complexities, generally its recognition is easy. As a result, if a user is aware about fundamental structure of reCAPTCHA it does not assist the system and just presents the required name for solving the challenge. As a result, there is no more value added for digitalization of content.

On the other hand, in reCAPTCHA two different images are presented for the user without any separable signs. Therefore users are obliged to enter the names of both images in order to solve the problem. This is the real reason that PICTCHA is the only CAPTCHA for making value added at %100 of conditions. In other words, the presented names by the users are useful. For instance, searching engines are able to search content of images by the use of these names accordingly.

2.2.3 Easy application

We compared various CAPTCHAs in our studies about PICTCHA and according to a questionnaire in which 60 users answered to some questions. In order to have more security in reCAPTCHA, there are more complex words responsible for separation of mankind and computer. Most of participants were claiming bout complexity of images. They stated that most of the times it is difficult for them to read the contents.

Furthermore, there are some other problems for other CAPTCHAs from viewpoint of time and solving the problems. It is time wasting for selection various photos from among a group of photos in Assira, watching a video clip for labeling and/or solving a moving CAPTCHA in comparison with labeling a unique image.

None of the mentioned problems are defined in PICTCHA. Labeling of two images is really better than recognition a complex content without any need to more time than common CAPTCHAs.

2.2.4 Bandwidth

Web page size is important in today web. Since data programs are expensive, most of managers optimized their websites with movable browsers. Also the number of HTTP requests is important. By remembering all above-mentioned items, a video CAPTCHA is really expensive for users through the time. It is also true in Assira which may use 12 images as well. The size of illustrated image in PICTCHA is not more than 9 kb even in colored form.

2.2.5 Item counting

One of the important factors in designing a CAPTCHA system is the Number of items for which the user is obliged solve it in order to verify the CAPTCHA. Sometimes any effective items on experience and system security of user may cause some trade off (equilibrium and replacement of factors). There are 12 images in Assira which should be processed by the user. NuCAPTCHA needs one word for three characters. Therefore video CAPTCHA needs also presenting three words by the users. Only processing of one item is enough in common CAPTCHAs while it is necessary to label two images in reCAPTCHA and PICTCHA. All presented information in this part may provide various points about other CAPTCHAs as the real idea for creation of PICTCHA. Table 2 includes briefly all presented information in this part.

3. PICTCHA: the proposed system

As it was mentioned before, PICTCHA has been designed for prevention of any spam and also labeling of problem free pictures like other types of it. Figure 2 illustrates PICTCHA in both English (2a) and Persian (2b) languages. There are two images in this system. One of them is labeled by hand and it is expected to have other image labeled by other users. In contrast with re-CAPTCHA, labeled image is not recognizable from the other and it is expected to have both images labeled by the user in order to verify relevant PICTCHA.

Table 2: Comparing of PICTCHA & other CAPTCHAs

Parameter/CAPTCHA	reCAPTCHA	NuCAPTCHA	Video CAPTCHA	Assira	PICTCHA
Security against DeCAPTCHA	Text: 0% Voice: 1%	90%	--	--	--
Added Value	Mostly	No	No	No	Always
Easily Recognizable Items	No	Yes	Average	Yes	Yes
Bandwidth Usage	~5KB	~50KB	~600KB	~130KB	~8KB
Items to Recognize	2	1	3	12	2

In remained parts we may explain relevant structure of PICTCHA and its specifications and then discuss all characteristics of PICTCHA which are lost in other projects.

3.1 PICTCHA architecture

PICTCHA architecture is really important from different aspects. All mentioned aspects are discussed in details in this part.

3.1.1 API

The proposed system will provide an API for enabling all websites use the same. Applying of PICTCHA in a website needs a 11-steps process.
Fort Server model/customer is a process as illustrated in figure 3.

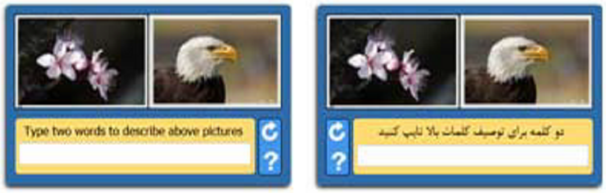
Type two words to describe above pictures
دو کلمه برای توصیف کلمات بالا تایپ کنید

Fig. 2 A sample PICTCHA in both English & Persian languages

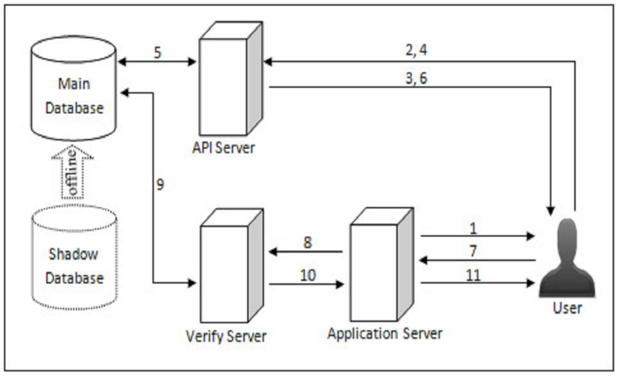

Fig. 3 Customer/Server model of PICTCHA

Figure 3 illustrates the 11-steps process as follows:

1. At first a web page is presented with an empty PICTCHA form accompanied with a general key which is sent by applicable server for the user.
2. User browser will send a general key to API server and request API through AJAX of a PICTCHA for relevant server.
3. All required information are created for making a PICTCHA in API server along with an exclusive code which may be reserved in data bank and sent for the user.
4. Then JAVA script code will write PICTCHA sign in DOM about special characteristics of required elements. As a result, by sending this sign to API server, user browser will send a request for PICTCHA image as well.
5. API server will evaluate information and load the sign information in relevant databank.
6. Both labeled / non-labeled images are loaded in API server for making a unique image which is sent as a PICTCHA image for the user.
7. After solving the PICTCHA problem by the user, he/she is able to send all names along with

general keys and also PICTCHA sign to API server by clicking Submit key in forms.

8. Applicable program will attach its private key to this information and send the same to the server. This private key may prevent third parties to send any requests from applicable server side.
9. Checking server evaluates and confirms any compliance of general and private keys. Then it may evaluate, load and delete any information about PICTCHA signs from databank. In case of entering a correct name, a name will be added to the nameless image.
10. Applicable server will receive the entered name accordingly.
11. If the user has entered an incorrect name, applicable server may send the content of requested item by the user, otherwise it may send another PICTCHA for him/her.

3.1.2 Security

Regarding the system architecture as illustrated in figure 3, there is high level of security for PICTCHA against threatening attacks. Followings are three important factors in security of a PICTCHA:

- All API servers are checked and PICTCHA data banks are separated from each other.
- In lack of security, all labeled images are modified and revised by the use of shade databanks strategy as mentioned before.
- It is impossible to use any images instead of complex contents and also this reality that how difficult is finding the content of an image by the use of computer software. Therefore it is in contrast with OCR software for ignoring the system.

3.1.3 Labeling of images

Images are one of the useful information resources in web. Searching engines are able to search image content. They are also useful tools for presenting this information. One way for finding this goal is labeling of images. In fact labeling is a recognizing process for all current articles in images and explaining them in content. At present, there is little number of techniques for labeling of images.

Luis Von Ahn has introduced one of the mentioned methods of today in 2004 [9]. The names are obtained in this method by the use of a game named as ESP in web. Since it was not an acceptable game for public people and users were obliged to play the game, therefore it was not applicable for great number of current images in web. As it was mentioned in previous section, one of both presented images to user is not labeled.

Therefore users are obliged to solve a PICTCHA which may assist them to make required labeling as well. One image is used for labeling in great numbers then all objects of image are labeled accordingly. For instance, figure 4c illustrates various names like sky, bird and eagle after solving all problems of PICTCHA.

3.1.4 Checking process

As it was stated in part 2, two pictures are presented for the user as follows: One is labeled and other should be labeled by the users. Current implementation could summarize any entered names by the user and if correct, finalize & add them to labeled images.

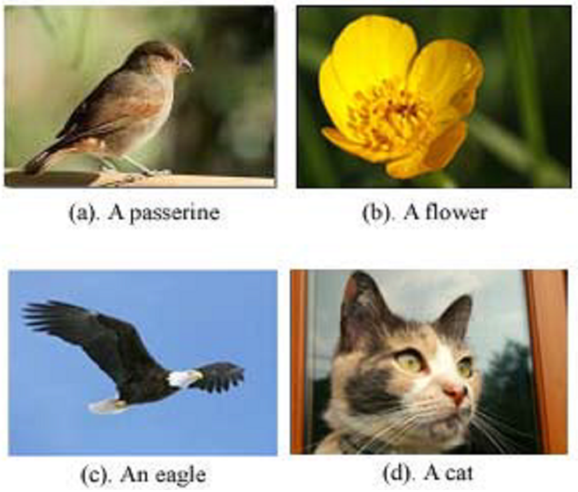
(a). A passerine
(b). A flower
(c). An eagle
(d). A cat

Fig. 4 Four sample images in a PICTCHA

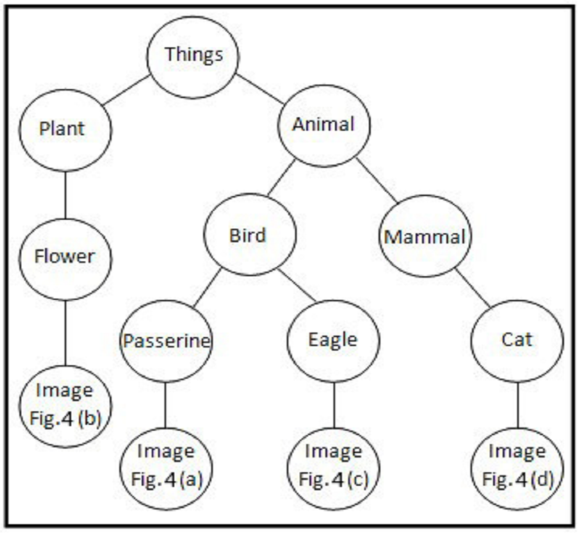

Fig. 5 A sample of logical graph

Assume that labeled T images are available in analysis sample. Then we need a threshold for evaluating a special name like 1 for an image named as i obtained from equation (2).

$$C=\sum_{i=0}^{T} Ci \quad (1)$$

$$Threshold = \frac{C}{T} \quad (2)$$

Quantity of C is calculated in equation 1, by adding Ci which is total number of i images for the first image from among Tth one. As a result, total number of names and images are calculating in equation (2).
Total number of names for calculated images is obtainable at the end of a day. If there is greater number of repetition changes is more than its threshold, this name will be finalized and its image will be added to the labeled pictures. For instance, assume that there are 300 non-labeled images at data banks and also 9270 names for these images. If image 4c is one of the mentioned images with four names of "Animal", "Bird", "Eagle" and "Sky" with repetition process of 40, 35, 32 and 10, then we will have:
CAnimal=40, CBird=35, CEagle=32, CSky=10
Tolerance threshold = 9270/300=30.9

As a result, since the repetition number of animal, bird and eagle is greater than tolerance threshold these names would be finalized for that image.

3.1.5 Forbidden word

According to our studies as mentioned in part 4, people are more interested to enter general words in comparison with special ones for PICTHAs. For instance, there are more chances for entering the word bird more than eagle. In order to enable our system to find out more names for a unique image hereby "Forbidden word" method is introduced as well. In this process, user's system prevents the entrance of any names with more repetitions than a threshold determined by relevant Admin. This means that if a picture is labeled with more repetition numbers as a "Bird", the system will start automatically this specification and request the user not to enter the word "Bird" in replying place.
The fundamental idea for this goal is to process any names of labeled images when there is a PICTCHA. Therefore if one or more numbers of which have greater repetition times than tolerance level it would be inserted in PICTCHA form for further information of user about these forbidden terms. By the way, this may enable a clever user to recognize labeled images from non-labeled ones and provide a name for labeled image.
In order to prevent from this problem, all labeled images in our databases are classified by a field including one or more groups. In case of a "Forbidden" title for a non-labeled image, a labeled image will be selected randomly by system and sent along with forbidden word for both images. Then the user has no chances for recognition of labeled image.

3.2 Other PICTCHA characteristics

In addition to all presented information, PICTCHA has some characteristics which are not available in other CAPTCHAs. The mentioned characteristics are introduced in this part.

3.2.1 Ontology

A considerable point in a PICTCHA is various acceptable names for each image while in re-CAPTCHA and other forms of CAPTCHA there is just one correct word for a complex content. All these names should be analyzed by the use of ontology for further acceptance. For instance if the image of bird named as "cooker" is sent for the user and he/she enters one of the words "Cooker", "Bird" or "Animal", he/she verifies the case successfully.

On the other hands, in lack of an equal for the entered name in ontology, but a great number of users enter the same for a special image, then it may be added to the relevant ontology. This process will assist the system to modify and develop ontology databases. Figure 6 illustrates this process as well.
According to figure 6 when there is a verify order in a system, it may send image sign along with API host keys and entered name by the user towards verify section. This part will send API keys towards language selection part for further analysis. Verify section will use the name and sign for confirming whether the user has entered a correct word or not accordingly. In lack of presence the word in database, system will reserve it as a temporary word.
Furthermore it is possible to apply all ontology databases for specific applications as well. For instance, the system is able to present issued ontology for nature, art or other scopes to RDF files through web services.
Figure 5 illustrates relevant graph in section 5 as a part of ontology of words.

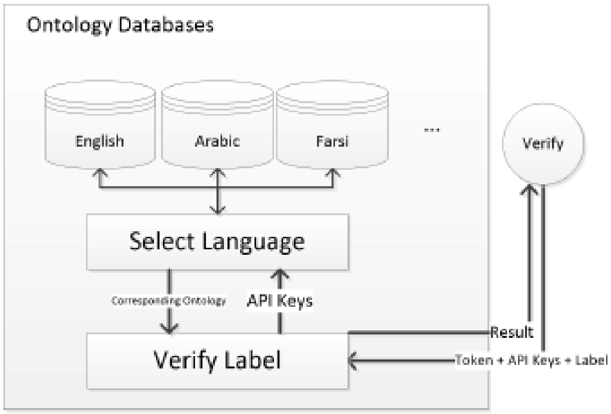

Fig. 6 Illustrating ontology output

3.2.2 Multilingual support

PICTCHA has a complete architecture for supporting various languages such as English, French, Chinese, Arabic and Persian. PICTCHA language is allocated for public/private keys to be selected through API registration process. In case the owners of applicable program prefer to support more than one language in their web sites, they are obliged to allocate a public/private pair keys to each considered language.

In order to have Multilanguage support, all labeled image names are translated by the use of translation software and reserved in multiple tables in databases. Furthermore all collected names for one language would be reserved in a similar table at database.

The other priority of multiple supports is ontology of the concerned language. This means that firstly we have ontology verify as mentioned in part 3-2-1 and then according to the ontology of a special language it is possible to create/develop it for other languages as well.

4. Results

We evaluated PICTCHA in an experiment with 60 participants. There was a web magic including 20 pages available for all users. This magic was developed in PHP language in companying with MySQL databases including 30 labeled images and 300 non-labeled images as well. There is a recommendation for all pages of magic. User is obliged, for verifying all steps, to solve relevant PICTCHA challenges presented in that page.

There were 3706 challenges in our experiment with totally 1758 efforts for solving the challenge from which about 955 were successful. From among effective factors in non-successful efforts, rather than incorrect recognition or entering an invalid word by user, there are little cases in which the user is unable to have easy recognition of image. These are including these cases in which the image has more details in great scales.

Upon omission of three images from databases and in spite of their results, we could consider 1077 (%61.26) efforts as successful one for meeting the needs of a PICTCHA system.

We evaluated all 955 presented names in successful efforts for measuring the success rate of PICTCHA in labeling of all 955 pictures. There were 130 separated names for different images and different repetition times. From among 300 images, minimum 110 images had different names with repetition times more than tolerance threshold. Therefore it was possible to be labeled. There were just two incorrect names and one of them was empty. According to this reality that recognition of labeled & non-labeled images is impossible for PICTCHA challenges, it is only possible by chance. By the way, PICTCHA success in this experiment was estimated as %98.18 in worst condition.

Rather than 130 presented names by users, public words like bird, flower, sea, animal and tree had more repetition times than other names. This shows that users are intending to apply any words with more public meanings for easier solving of challenges.

Although such a tendency may facilitate labeling of unknown images, on the other hand it may limit any details of system as effectively as possible. Therefore we have a list of forbidden words in part 3-1-5 as discussed before.

5. Conclusion

On-line private scope has a daily-increasing importance. On-line banking, enrolment processes and generally all recognition methods need some services for separation of mankind from automatic programs. CAPTCHAs are a famous methods of today and through websites. By the way, the current systems are unable to create acceptable experience for users. Also they are involved with various security problems. On the other hand, reCAPTCHA is the only project with required value added.

The PICTCHA system, defined in this paper, provides suitable and safe experience along with value added for supplier system and finally all users. The presented digits in part 4 illustrate system’s ability in successful labeling and introduce a practical CAPTCHA system with high rate of security.

Although it is useful and applicable but is facing with a fundamental problem which is required starting data base for labeled images. By the way, it is necessary to mention these databases are developing by labeling of images by users and as a result that is enough to have a small database from the first.

References

[1] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCaptcha: Human-based character recognition via web security measures," Science, vol. 321, pp. 1465-1468, 2008.

[2] ElieBursztein , Matthieu Martin , John Mitchell, Text-based CAPTCHA strengths and weaknesses, Proceedings of the 18th ACM conference on Computer and communications security, October 17-21, 2011, Chicago, Illinois, USA .

[3] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using hard AI problems for security. In Eli Biham, editor, Advances in Cryptology – EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4-8, 2003, Proceedings, volume 2656 of Lecture Notes in Computer Science, pages 294–311. Springer, 2003.

[4] Turing Test, http://en.wikipedia.org/wiki/Turing_test, visited Date: 26/06/2013.

[5] What is reCAPTCHA?, http://www.google.com/recaptcha/learnmore, Visited Date: 26/06/2013.

[6] J. Elson, J. R. Doucerur, J. Howell, and J. Saul.Asirra: A Captcha that exploits interest-aligned manual image categorization. In Proceedings of the 14th ACM conference on Computer and communications security, CCS ’07, pages 366–374, New York, NY, USA, 2007. ACM.

[7] How we broke the NuCaptcha video scheme and what we propose to fix it, http://elie.im/blog/security/how-we-broke-the-nucaptcha-video-scheme-and-what-we-propose-to-fix-it/, Visited Date: 26/06/2013.

[8] It’s 1999. Lowe, David G. (1999). "Object recognition from local scale-invariant features". Proceedings of the International Conference on Computer Vision. 2. pp. 1150–1157. DOI:10.1109/ICCV.1999.790410.

[9] Von Ahn, L.; Dabbish, L. (2004)."Labeling images with a computer game".Proceedings of the 2004 conference on Human factors in computing systems - CHI '04. pp. 319–326. DOI:10.1145/985692.985733. ISBN 1581137028.