Evaluation of CAPTCHA Efficiency

Passzarkorn Youthasoontorn* · Akarin Phaibulpanich** · Krerk Piromsopa***

Abstract

We propose statistical methods for evaluating the efficiency of CAPTCHA. Most people unfairly assumed that machines are not capable at reading precisely. This fact leads to the invention of CAPTCHA, a distorted word or short phase, which is designed to thwart computers and separate human from machines. However, advances in image recognition technologies mean that machines are constantly getting better at recognizing CAPTCHA. This forces CAPTCHA designers to design even more difficult CAPTCHAs to prevent their systems from being gamed by malicious bots. However, this arm race has an unintended side effect on the common users. Many CAPTCHAs are now so hard that many people are unable to read them. This obviously conflicts with the original purpose that CAPTCHA was invented in the first place. Our analysis shows that some CAPTCHAs are more users friendly. In particular, Yahoo-style CAPTCHA is the most friendliness. This suggests that a good CAPTCHA could be a simple text with some distortion that prevents machines from correctly segmenting characters.

Keywords : CAPTCHA, Challenge-response, Evaluation, Efficiency, CAPTCHA Friendliness, Botnet

Received: 2015. 06. 24.   Final Acceptance: 2015. 09. 13.

* Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand, e-mail: passzarkorn.y@student.chula.ac.th
** Department of Statistics, Chulalongkorn University, Bangkok, Thailand, e-mail: akarin@cbs.chula.ac.th
*** Corresponding Author, Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand, e-mail: krerk.p@chula.ac.th

1. Introduction

People always believe that machines are incapable of reading words or short phases from pictures. This is partly due to the limitations in image processing and character recognition technology in the early days. In addition, advance image-recognition technologies were not available to public at large. With these assumptions in mind, the Completely Automated Public Turing test to tell Computer and Human Apart (CAPTCHA) [Ahn et al., 2004] was invented to separate machines from human.

CAPTCHA is primarily designed to prevent undesirable accesses or malicious attacks from botnets. Over the years, various types of CAPTCHAs have been invented with different level of effectiveness. Recently, the rapid advance in image recognition and machine learning techniques have raised the concern on the efficacy and significant of CAPTCHA.

To evaluate whether CAPTCHA is still a viable tool for differentiating machines from human, we have formulated an evaluation method that contrasts and compares human and machines ability in solving various types of CAPTCHAs. The results of our evaluation show the effectiveness of several types of CAPTCHAs at obfuscating words and phases from machines. Our study also unveils the human ability at recognizing and solving CAPTCHAs.

We believe that an inclusive analysis of CAPTCHA is needed. The approach should take into account both effectiveness and usability. The effectiveness is the ability that a CAPTCHA can prevent machines from malicious accesses. The usability is how efficient a user can recognize a CAPTCHA.

We believe that our approach will reveal the true effectiveness of CAPTCHAs. In our evaluation, CAPTCHA friendliness is a key criterion for measuring efficiency. The CAPTCHA friendliness is defined as a difference between the success rate of human at solving a CAPTCHA and the success rate of machines at solving the same CAPTCHA.

The remainder of this paper is organized as follows. Section II reviews related literatures. Section III describes our experimental method. Our experimental results and conclusion are in Section IV and Section V respectively.

2. Literature Review

Most existing research on CAPTCHA does not focus on efficiency. To our knowledge, the best (if not only) way to determine whether a CAPTCHA can prevent a bot from accessing the system without complicating human is to conduct real experiments. In this section, we will review previous works related to the development of CAPTCHA.

Since the design goal of CAPTCHA is to prevent a machine from entering a system, it has been developed and localized as text CAPTCHA, picture CAPTCHA, voice CAPTCHA, logic CAPTCHA, and others. From the design concept, they should be human friendly [Nanglae and Bhattarakosol, 2012; Shirali-Shahreza and Shirali-Shahreza, 2008; Chandavale and Sapkal, 2011; Truong et al., 2011; Tamang and Bhattarakosol, 2012; Hsieh and Wu, 2013].

To protect a system from potential threat, CAPTCHA designers usually focused on the forms of challenge questions based on human understanding. The HIPs (Human Interaction Proofs) research [Chellapilla et al., 2005] suggests that human can respond better (more accurate) to this kind of question.

Contemporary challenge-response method does not only take the advantage of human vision and recognition, but it also uses the human ability in logic interpretation and voice recognition. Comparing to others, text-based CAPTCHA is the most accurate for human. Though heavily attacked by brute force, it is generally used by popular websites for years.

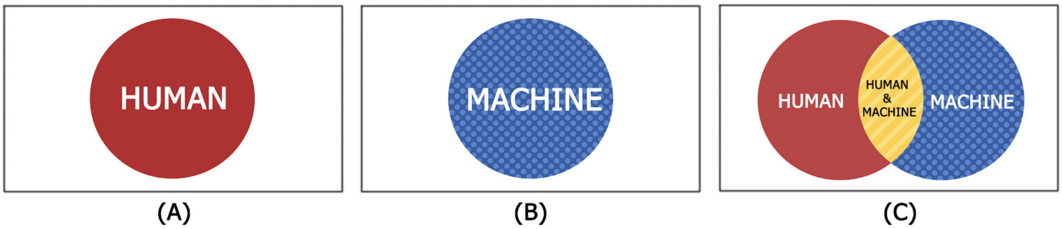
Initially, machines were poor at segmenting characters. Given that segmentation is a key to the success of reading, the designs of CAPTCHA used this fact for separating human and machines. Based on this assumption, we observed that the early day of CAPTCHA in this decade emphasized on applying noise (descend, distort, resize, space reduction, background pattern and cross line) to characters in the CAPTCHA picture. Some research [El Ahmad et al., 2012; Chellapilla and Simard, 2004] looked for methods to obstruct machine by causing difficulty in distinguishing characters. However, the research is not fast enough to surpass the expansion in machine power. Machine can now overcome some technique by applying naive image processing methods [Chandavale and Sapkal, 2010; Yan and El Ahmad, 2009; Li et al., 2010; Chellapilla et al., 2005].

Distorted Text-Based CAPTCHAs (clouding and distorting character) can successfully reduce the learning ability of machines. It, however, prevents human from reading the text correctly [Chellapilla et al., 2005].

Recent advances in machine learning has introduce Deep Convolutional Neural Networks Technique [Goodfellow et al., 2013; Lv, G, 2011; Shu-Guang et al., 2011]. This technique has proven that machines can be trained to appropriately recognize CAPTCHA. This has set a new hurdle between technology and security.

Challenges in creating efficient CAPTCHA is to ideally make it good at preventing bot attack while still being friendly to human (<Figure 1> shows the concept). Good CAPTCHA should only prevent bots without causing trouble to human.

A related study shows that human with different backgrounds (including age, nationality, education, English proficiency, origin, current habitat, and Internet experience) react diffe-



<Figure 1> (A) The Ideal Universe of Well-Designed CAPTCHA (B) The Ideal Universe of Badly Designed CAPTCHA (C) The Universe of Existing CAPTCHA

rently in recognizing CAPTCHA. The study shows that the accuracy decreases as the age increases. People with higher education also gain more success rate. In particular, Google-style CAPTCHA (distortion) has less accuracy and usually require longer period of time for a person to recognize comparing to that of noise-based CAPTCHA [Bursztein et al., 2010].

In the current state of the art, attacks on CAPTCHA are done in several steps. A CAPTCHA image is first preprocessed and separated into chunks (characters). They are then entering the recognition stage [Bursztein et al., 2011]. Noise (stroke and dots) is always an issue. In the early days of CAPTCHA, stroke and dots were added to make CAPTCHA resistance to the segmentation. However, the research has found that eliminating such noise is trivial (fast and accurate [Chandavale and Sapkal, 2010]). In fact, segmentation of a CAPTCHA with noise is proven to be simpler than segmentation of distorted CAPTCHA [Bursztein et al., 2011]. This suggests that noise is not a good obstacle for segmentation comparing to distortion. In general, good segmentation will increase the possibility of recognizing a character (by OCR) to 90% at the average time of 80 milliseconds [Yan and El Ahmad, 2008].

Most studies focus on either human ability or machine ability. Few study focus on both aspects at the same time. There exists a study [Chellapilla et al., 2005] focusing on human and machine ability in reading scruffy character without text localization to test machines. However, the study did not apply to the whole CAPTCHA picture. Thus, it is incomplete since most machines fail to segment localized text.

It is arguably that text-based CAPTCHA is efficient for authenticating human. In next section, we will demonstrate CAPTCHA efficiency to address this concern.

3. Experimental Design

This section provides the details of our methodology. We use the sampling method suggested by Taro Yamane [1967] to determine the size of our sample with some trade-offs in reliability and accuracy. In big picture, representatives of CAPTCHAs are taken from popular web sites. We ask human subjects to recognize them. A simple OCR is used for representing machine ability. By passing the same data set to the OCR, results are then evaluated.

3.1 Sample Size

According to the World Wide Web Consortium (W3C), there are 2,925,249,355 Internet users world wide (data as of July 1st, 2014). We apply the Taro Yamane sampling size method [Yamane, 1967] to find a suitable number of our sample group, with a confidence interval at 95% and allowable error at 0.05. The calculation method is described as follows.

Let

n = size of sample group
N = number of population
e = allowable error

Equation

$$n = \frac{N}{1+N(e)^2}$$

Calculation

$$n = \frac{2{,}925{,}249{,}355}{1+2{,}925{,}249{,}355(0.05)^2}$$

$$n = 400 \text{ persons}$$

Therefore, at the confidence level of 95%, the minimum size of the sample group is 400 persons.

3.2 OCR Engine

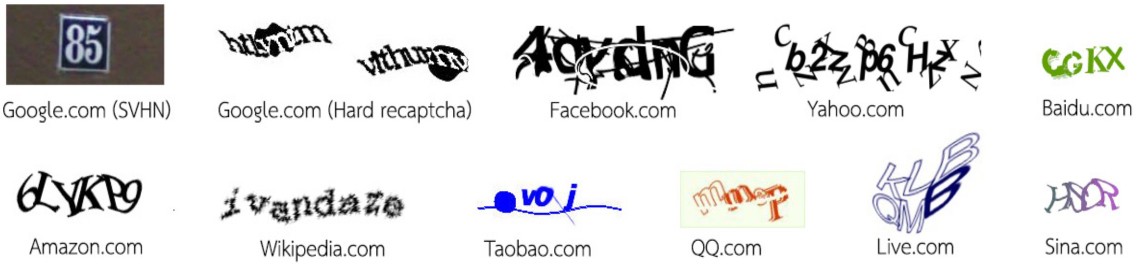
We choose to use Tesseract, a commonly available OCR engine, for our evaluation. Tesseract OCR is an open source OCR engine. It is known as one of the most accurate among other open-source engines. We have primarily validated Tesseract OCR against Cool-php-captcha-0.3.1. The preliminary result shows the accuracy rate of 60%. In our experiment, Tesseract 3.02 is used (The Cool-php-captcha script is taken from https://code.google.com/p/cool-php-captcha/ and the Tesseract software is taken from https://code.google.com/p/tesseract-ocr/).

It is worth clarifying that a bot can be much more powerful in the real world. We choose to use only the plain vanilla Tesseract OCR in our study to represent a novice bot (or poor hacker) only.

3.3 CAPTCHA

To capture the practical scenario, real CAPTCHAs were considered in our evaluation. As suggested by Alexa (Website ranking located at http://www.alexa.com/topsites), 11 sets of CAPTCHAs from 10 most popular websites were chosen (see <Figure 2>). However, only websites with unique CAPTCHAs were selected. For example, Google uses the same reCAPTCHA on the search engine page and the YouTube page. Thus, only one type on CAPTCHA is selected. However, Google also uses street-view CAPTCHA (SVHN) in certain cases. Thus, we also include it in our study. The 11 sets of CAPTCHAs are:

- Google.com (SVHN)
- Google.com (Hard reCAPTCHA)
- Facebook.com
- Yahoo.com
- Baidu.com
- Amazon.com
- Wikipedia.com
- Taobao.com
- QQ.com
- Live.com
- Sina.com



Google.com (SVHN) Google.com (Hard recaptcha) Facebook.com Yahoo.com Baidu.com

Amazon.com Wikipedia.com Taobao.com QQ.com Live.com Sina.com

〈Figure 2〉 Samples of CAPTCHA

For reCAPTCHA, different datasets may be used based on conditions. In our study, we limit the datasets of reCAPTCHA to the text-based datasets, which is comprised of SVGN dataset and hard reCAPTCHA dataset. The SVGN dataset only has numbers from natural scenes, while the Hard reCAPTCHA consists of more complex images that are only used when Google detects anomalies with associated IP address.

In our experiment, a subject must solve all unique CAPTCHAs. Given 400 subjects, 800 CAPTCHA images from the 10 websites are prepared. Additional 200 CAPTCHA images are also stored for backup. Totally, there are 1,000 CAPTCHA images collected.

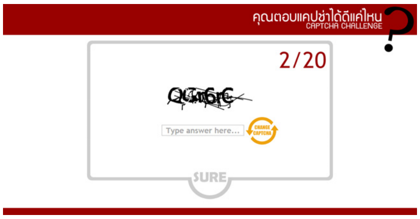
In order to ensure the fairness, human subjects and machines (OCR engine) must solve the same CAPTCHA images.

3.4 Website Challenge Preparation

We created a website, www.captchachallenge.com, to collect minimum data from 400 random Internet users. The participants must first enter their age, gender, nationality, and education. We also collect the IP address of each participant for validation purposes. To avoid any human errors and minimize participant's fatigue, a subject will only solve 20 CAPTCHA images per session. For each session, only 2 CAPTCHA images from the same source will be used. <Figure 5> shows the screenshot of our website.

3.5 CAPTCHA Efficiency Evaluation

To evaluate the effectiveness of CAPTCHA, both machines and human must be assessed. In order to do this, two metrics are introduced. They are Human Success Rate (HSR) and Machine Success Rate (MSR).



〈Figure 3〉 Websites' Picture for Challenge

To further find the CAPTCHA efficiency (CE), we propose to use the subtraction of the MSR value from the HSR value. The idea is to show the difference in the performance of human and machine in solving same type of CAPTCHA. Equation (1) is the function.

CAPTCHA EFFICIENCY is CE

$$CE = HSR - MSR \tag{1}$$

The CE value represents the effectiveness for each type of CAPTCHA. To ease the presentation, the CAPTCHA friendliness is introduced for representing the CE value. The idea is to use simple star rating (ranging from none to 5 stars) for ranges of values. <Table 1> shows the relation between the CE values the friendliness.

〈Table 1〉 CAPTCHA Friendliness

Level	CE	CAPTCHA Friendliness
1	< 0	(N/A)
2	0~10	½
3	11~20	★
4	21~30	★½
5	31~40	★★
6	41~50	★★½
7	51~60	★★★
8	61~70	★★★½
9	71~80	★★★★
10	81~90	★★★★½
11	91~100	★★★★★

The CE value range from -100 to +100. However, the negative value means no human is able to solve the CAPTCHA. This means CAPTCHA of this range misses the purpose. Thus, the CAPTCHA friendliness only shows the range of positive value.

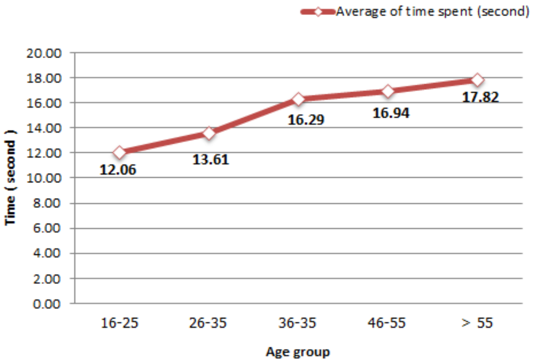
4. Results

We started our experiment in January 2015. It took about 2 months for 652 participants from all over the world. Only 487 participants are applicable since some of them did not complete all 20 questions. On average, each CAPTHA image was solve in 13.40 seconds.

Most participants (48%) age between 26 and 35 years old. About 79.9 percent of them have a college degree or higher. Half of them (55%) are Thai. The ratio of male and female is 54:46. Each participant reached the success rate at 84.9 percent. The details are shown in <Table 2>.

From the result (as expected by the authors), plain-vanilla Tesseract OCR can only solve 2 types of CAPTCHA from Baidu and Wikipedia. The MSR for Baidu and Wikipedia are at 28.54% and 0.62% respectively.

The average of HSR is 84.59%. For Yahoo-style CAPTCHA. The friendliness is as high as 94.15. While Facebook has the lowest HSR, the lowest friendliness goes to Baidu at 62.01. This is because our OCR can also solve Baidu-style CAPTCHA as well.



〈Figure 4〉 Relation between Time Taken to Solve CAPTCHA and Age

We further analyze the relation between the average time taken to solve a CAPTCHA and the age (in <Figure 4>). The result suggested that the younger person were able to solve the CAPTCHA faster then the older person.

〈Table 2〉 Camparision of CAPTCHA Efficiency between Human and Plain-Vanilla Tessaract

CAPTCHA	Human Success Rate (HSR)	Machine Success Rate (MSR)	CAPTCHA Efficiency (CE)	CAPTCHA Friendliness
SVHN	93.63	0	93.63	★★★★★
FACEBOOK	71.66	0	71.66	★★★★
YAHOO	94.15	0	94.15	★★★★★
BAIDU	90.55	28.54	62.01	★★★½
AMAZON	79.16	0	79.16	★★★★
WIKIPEDIA	77.82	0.62	77.2	★★★★
TAOBAO	84.09	0	84.09	★★★★½
QQ	85.93	0	85.93	★★★★½
LIVE	90.25	0	90.25	★★★★★
SINA	80.9	0	80.9	★★★★½
RECAPTCHA HARD	82.34	0	82.34	★★★★½

5. Conclusion

We have proposed a statistical method for evaluating the efficiency of text-based CAPTCHA. To find the suitable CAPTCHA, we propose a CAPTCHA efficiency function and a friendliness value. This suggests that good text-based CAPTCHAs are those that can be solved by human with additional properties (i.e. distortion) that prevent machines from properly segmenting the characters.

Our experiment also suggests that the ability in solving CAPTCHA is depending on age of the person. However, the study of Human Interaction Proofs (HIPs) is beyond the scope of our study. We hope to find the relation between the CAPTCHA efficiency and the HIPs in the future.

Note that this study is based on the plain-vanilla open-source Tesseract OCR. It may not capture the state of the art employed by advance botnets. Nonetheless, we hope that it provides a basis for further improvement in CAPTCHA technology.

References

[1] Ahn, L. von, Blum, M., and Langford, J., "Telling humans and computers apart automatically", *Commun. ACM*, Vol. 47, No. 2, 2004, pp. 56-60.

[2] Bursztein, E., Bethard, S., Fabry, C., Mitchell, J. C., and Jurafsky, D., "How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation", in *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, Washington, DC, USA, 2010, pp. 399-413.

[3] Bursztein, E., Martin, M., and Mitchell, J. C., "Text-based CAPTCHA Strengths and Weaknesses", *Comput. Commun. Secur.*, 2011.

[4] Chandavale, A. and Sapkal, A., "Algorithm for Secured Online Authentication Using CAPTCHA", in *2010 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, 2010, pp. 292-297.

[5] Chandavale, A. and Sapkal, A., "An Improved Adaptive Noise Reduction for Secured CAPTCHA", in *2011 4th International Conference on Emerging Trends in Engineering and Technology (ICETET)*, 2011, pp. 12-17.

[6] Chellapilla, K. and Simard, P. Y., "Using Machine Learning to Break Visual Human Interaction Proofs (HIPs)", in *Neural Information Processing Systems*, 2004.

[7] Chellapilla, K., Larson, K., Simard, P., and Czerwinski, M., "Computers beat humans at single character recognition in reading based human interaction proofs (HIPs", in *In 2nd Conference on Email and Anti-Spam*, 2005.

[8] Chellapilla, K., Larson, K., Simard, P., and Czerwinski, M., "Designing human friendly human interaction proofs (HIPs)", 2005, p. 711.

[9] El Ahmad, A. S., Yan, J., and Ng, W.-Y., "CAPTCHA Design: Color, Usability, and Security", *IEEE Internet Comput.*, Vol. 16, No. 2, 2012, pp. 44-51.

[10] Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V., "Multi-digit Number Re-

cognition from Street View Imagery using Deep Convolutional Neural Networks", *ArXiv-13126082 Cs*, 2013.

[11] Hsieh, C.-C. and Wu, Z.-Y., "Anti-SIFT Images Based CAPTCHA Using Versatile Characters", in *2013 International Conference on Information Science and Applications (ICISA)*, 2013, pp. 1-4.

[12] Li, S., Shah, S. A. H., Khan, M. A. U., Khayam, S. A., Sadeghi, A.-R., and Schmitz, R., "Breaking e-Banking CAPTCHAs", in *Proceedings of the 26th Annual Computer Security Applications Conference*, New York, NY, USA, 2010, pp. 171-180.

[13] Lv, G., "Recognition of Multi-Fontstyle Characters Based on Convolutional Neural Network", in *2011 Fourth International Symposium on Computational Intelligence and Design (ISCID)*, Vol. 2, 2011, pp. 223-225.

[14] Nanglae, N. and Bhattarakosol, P., "A Study of Human Bio-detection Function under Text-Based CAPTCHA System", in *2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS)*, 2012, pp. 139-144.

[15] Shirali-Shahreza, S. and Shirali-Shahreza, M. H., "Bibliography of works done on CAPTCHA", in *3rd International Conference on Intelligent System and Knowledge Engineering, ISKE 2008*, Vol. 1, 2008, pp. 205-210.

[16] Shu-Guang, H., Liang, Z., Peng-Po, W., and Hong-Wei, H., "A CAPTCHA Recognition Algorithm Based on Holistic Verification", in *2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, 2011, pp. 525-528.

[17] Tamang, T. and Bhattarakosol, P., "Uncover impact factors of text-based CAPTCHA identification", in *2012 7th International Conference on Computing and Convertgence Technology (ICCCT)*, 2012, pp. 556-560.

[18] Truong, H. D., Turner, C. F., and Zou, C. C., "iCAPTCHA: The Next Generation of CAPTCHA Designed to Defend against 3rd Party Human Attacks", in *2011 IEEE International Conference on Communications (ICC)*, 2011, pp. 1-6.

[19] Yamane, T., *Statistics; an introductory analysis*. New York: Harper and Row, 1967.

[20] Yan, J. and El Ahmad, A. S., "A Low-cost Attack on a Microsoft Captcha", in *Proceedings of the 15th ACM Conference on Computer and Communications Security*, New York, NY, USA, 2008, pp. 543-554.

[21] Yan, J. and El Ahmad, A. S., "CAPTCHA Security: A Case Study", *IEEE Secur. Priv.*, Vol. 7, No. 4, 2009, pp. 22-28.

■ Author Profile

Passzarkorn Youthasoontorn

Passzarkorn Youthasoontorn received his BA degree in management science in from Silpakorn University, Thailand, in 2007. After working in fields of information and communication technology for 5 years, he decided to study computer science and received his MS degree in computer engineering from Chulalongkorn University, Thailand, in 2015. His research interests are in cyber security and statistical computing.

Akarin Phaibulpanich

Akarin Phaibulpanich received his BS in mathematics, statistics and computer science from University of Wisconsin-Madison, USA, in 2000, and received his MA and Ph.D. in Statistics from University of Michigan-Ann Arbor, in 2005 and 2006, respectively. After working for Office of Atoms for Peace under the Ministry of Science and Technology, Thailand, for 6 years, he has joined department of statistics at Chulalongkorn business school since 2012.research interests include multivariate analysis, data mining, and statistical computing. is currently the director of center for statistical consultation and research at Chulalongkorn business school an dis also a member of international statistical institute.

Krerk Piromsopa

Krerk Piromsopa received the BE and ME degrees in computer engineering from Chulalongkorn University, Thailand, in 1998 and 2000, respectively. In 2003, he was awarded a scholarship from the Royal Thai Government for pursuing his Ph.D degree in computer science, Michigan State University, where received his Ph.D in 2006. He has been a professor of computer engineering at Chulalongkorn University since 2001. His research interests are in cyber security, computer architecture, and high performance computing. He is a member of the IEEE and IEEE Computer Society for 10 years.