Distinguishing between Humans and Robots on the Web

December 16, 2011

Richard Abrich
richard.abrich@utoronto.ca

Valentin Berbenetz
valentin.berbenetz@utoronto.ca

Matthew Thorpe
matthew.thorpe@utoronto.ca

Abstract

We present a new detection technique for distinguishing humans from bots on the web, comprised of one existing active and one existing passive technique. We arrived at the proposed new detection technique by evaluating each of 10 of the most prevalent existing techniques. Each technique was evaluated in three categories of criteria: administrator experience, user experience, and effectiveness.

The administrator experience was evaluated based on the technique's level of automation, and its library availability. The user experience was evaluated based on whether or not the user's input was correct, and the difficulty rating they assigned. The effectiveness was evaluated based on a literature survey to determine the quantitative efficacy of current techniques. We compiled our data to determine a numerical score for each technique based on a rubric we created. Our proposed new detection technique combines the top scoring user-active and top scoring user-passive techniques.

1 Introduction

The security risks posed by internet bots and other types of malicious web crawlers are well documented. The CAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) was introduced as early as 1997 at AltaVista, with the initial goal of preventing automatic URL submissions designed to skew their search engine's importance ranking. [6]

Since then, the ensuing arms race between web developers and bot programmers has made it increasingly difficult for such tests to correctly identify humans while blocking bots. Some bots still manage to slip by the best defences undetected.

While much study has been devoted to the development and testing of individual detection techniques, there are no significant research efforts to date that perform a comprehensive comparison of these techniques. This paper addresses this research gap by ranking 10 of the most prevalent detection techniques using three categories of criteria: administrator experience, user experience, and effectiveness.

Data was populated in each category for each detection technique through a web-based testing platform [1] and user survey, as well as a preliminary literature review. The final rankings were determined according to an equation we propose to calculate the numerical score of each technique as a weighted sum of each criterion.

After ranking the techniques from highest to lowest, we propose and analyze a combination of two techniques that complement each other's strengths. The subsequent sections of the report are outlined as follows: section 2 provides background and related work, section 3 describes our analysis framework, section 4 reports and analyzes our results, and section 5 presents our proposed solution.

2 Background

2.1 The Current State of CAPTCHA

Most techniques to date have been variations on the original idea of generating an image of printed text such that humans can read it while machine vision systems cannot, and requiring the user to enter the text used to generate the image. We categorize such techniques which require the user to supply explicit proof as active techniques. These include:

Image CAPTCHA The user is required to enter text which appears distorted in an image.

Video CAPTCHA Similar to traditional text-based CAPTCHAs, but with animated instead of static images.

Audio CAPTCHA Generating and replaying audio in which computer-generated spoken words are combined with background noise.

Image SAPTCHA A semi-automated technique, in which the administrator provides one or more images and a description for each, and upon seeing the image, the user must supply the description.

Word SAPTCHA Another semi-automated technique, with questions and answers in text rather than images and descriptions.

Unusual Form Interaction Prompting the user to interact with an HTML form in an atypical manner, such as un-checking an already-checked checkbox.

Confirmation Page Requiring the user to provide confirmation after submitting a form.

Other techniques that do not require explicit user-supplied proof, which we categorize as passive techniques, include:

Hidden Form Elements Hiding form elements from users (e.g. through CSS and/or HTML attributes), with the assumption that if information is entered into the element, the form was submitted by a bot which does not understand the aforementioned CSS and HTML attributes.

Biometric Data Analysis Analyzing biometric data such as form completion time (bots are likely to take less time than humans).

JavaScript Detection Detecting if the user has JavaScript enabled (bots do not typically implement JavaScript engines).

Clearly, there is a wide variety of techniques available. However, very little is known about their relative effectiveness.

2.2 Related Work

2.2.1 Comparing Techniques

Studies into CAPTCHA and its ilk usually focus on one of two main aspects: defeating CAPTCHAs, and evaluating their effectiveness. Few studies, however, have attempted to gather and analyze data for various individual techniques in order to evaluate their relative merits.

One exception to this is *On the Necessity of User-Friendly CAPTCHA*, a 2011 paper published out of the University of Patras, Greece. They confirm the notion that CAPTCHAs are still difficult for humans to solve, with less than half of all participants succeeding on the first try. [5]

While the study gathered data from a questionnaire-based survey as well as real-usage scenarios, their focus was on text-based CAPTCHAs only.

2.2.2 Non Text-Based CAPTCHAs

While many studies have been conducted on the security and efficacy of text-based CAPTCHAs, the novelty and/or obscurity of other techniques has translated into a relative dearth of knowledge in the literature.

A counter-example to this trend is Decaptcha, a tool developed at Stanford University designed to break audio CAPTCHAs. The study reports that Decaptcha was able to successfully break 75% of eBay audio CAPTCHAs. [2] In May 2011, the Stanford University News Service further reported success rates of 50% and 1% for breaking Microsoft CAPTCHA and reCAPTCHA, respectively. [4]

Clearly, audio CAPTCHAs are just as fallible as their text-based counterparts. However, little information is available concerning other aspects of the audio CAPTCHA, such as usability and ease of implementation.

3 Analysis Framework

3.1 Analysis Criteria

In order to perform a comprehensive analysis, we must consider all relevant aspects of a technique. Thus, we analyzed each technique by assigning scores to criteria in three different categories: user experience, administrator experience, and effectiveness at distinguishing between humans and robots. Each criteria was also assigned a weight, as per Table 1.

Table 1: Technique comparison criteria and weighting

Category	**Criteria**		**Weight** (W_i)
Effectiveness	Effectiveness	X_1	5
Administrator Experience	Library availability	X_2	4
	Automation	X_3	3
User Experience	Time to solve	X_4	2
	Difficulty	X_5	1
	% Affinity	X_6	1
	% Correct	Z	-

Once scores for each criteria were assigned, a final score was calculated to directly compare the overall performance of each technique using the following equation:

$$Y_k = Z \sum_{i=1}^{6} W_i \frac{X_i}{\max_k X_{i,k}}$$

Where Y_k is the final score for technique k, calculated as the normalized weighted sum over the criteria X_i.

The weighting of each criteria was obtained by considering the relative impact of that criteria on the overall usability. For example, the effectiveness of a technique represents its ability to accomplish its main goal, which is to prevent bots from impersonating humans, and is thus weighted the highest.

We rank Library Availability and Automation as the next most important effective techniques, as these impact an administrator's ability to implement the technique. Were it not for a large weight on these criteria, then completely manual techniques such as having a user contact the administrator directly would otherwise be scored very highly. These types of techniques do not scale well for services with large user bases, such as those commonly found on the Web.

The criteria relating to the User Experience are ranked the lowest for three reasons. First, these criteria outnumber those in the other two categories. Second, the data assigned to these criteria are the most unreliable, since they are collected directly from users, and are susceptible to manipulation and outliers. Finally, as can be seen from the equation, the weighted sum only applies to the first six criteria, with the final sum being multiplied by the seventh (the rate at which users correctly solve a technique). The intuition behind this is that all other criteria are irrelevant if a user is unable to be provide a correct solution, and by extension be identified as a user instead of a bot.

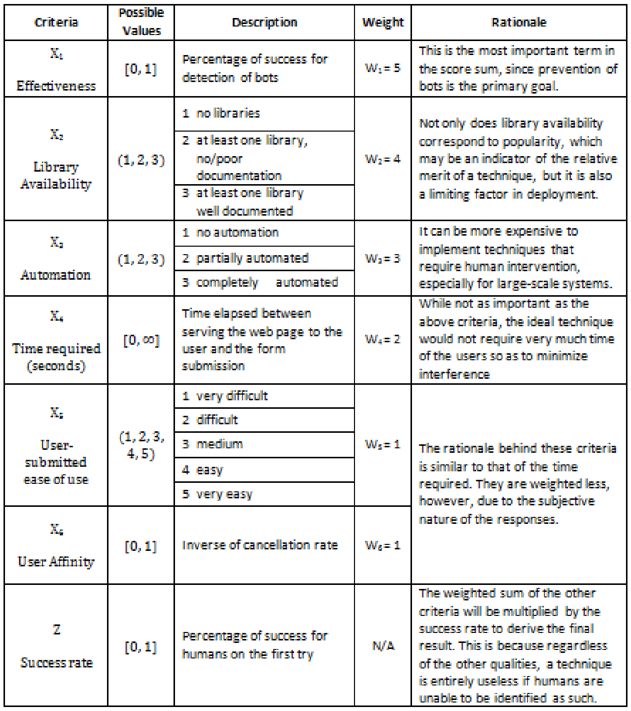
Since the final score is calculated as a sum, each criteria was selected such that a higher value indicates a better technique. For example, the % affinity is calculated as the inverse of the percentage of forms of that technique that were skipped by users (i.e. $1 - \%_{skipped}$).

The guidelines by which the score for each criteria were assigned are shown in Figure 1.

3.2 Data Acquisition

3.2.1 User and Administrator Experience

In order to gather data regarding the administrator experience of the different techniques, we implemented each in a website [1], and recorded our experience. Text CAPTCHA, Video CAPTCHA, and Audio CAPTCHA implementations were available from reCAPTCHA and NuCaptcha. All other techniques, however, were implemented manually.



Criteria	Possible Values	Description	Weight	Rationale
X_1 Effectiveness	[0, 1]	Percentage of success for detection of bots	$W_1 = 5$	This is the most important term in the score sum, since prevention of bots is the primary goal.
X_2 Library Availability	(1, 2, 3)	1 no libraries 2 at least one library, no/poor documentation 3 at least one library well documented	$W_2 = 4$	Not only does library availability correspond to popularity, which may be an indicator of the relative merit of a technique, but it is also a limiting factor in deployment.
X_3 Automation	(1, 2, 3)	1 no automation 2 partially automated 3 completely automated	$W_3 = 3$	It can be more expensive to implement techniques that require human intervention, especially for large-scale systems.
X_4 Time required (seconds)	[0, ∞]	Time elapsed between serving the web page to the user and the form submission	$W_4 = 2$	While not as important as the above criteria, the ideal technique would not require very much time of the users so as to minimize interference
X_5 User-submitted ease of use	(1, 2, 3, 4, 5)	1 very difficult 2 difficult 3 medium 4 easy 5 very easy	$W_5 = 1$	The rationale behind these criteria is similar to that of the time required. They are weighted less, however, due to the subjective nature of the responses.
X_6 User Affinity	[0, 1]	Inverse of cancellation rate	$W_6 = 1$	
Z Success rate	[0, 1]	Percentage of success for humans on the first try	N/A	The weighted sum of the other criteria will be multiplied by the success rate to derive the final result. This is because regardless of the other qualities, a technique is entirely useless if humans are unable to be identified as such.

Figure 1: Technique evaluation rubric

To gather data regarding the user experience, we created a survey on the website allowing users to fill out each technique, and invited users to participate. When a user participated in the survey, they were presented with a randomly chosen technique and asked to solve it. They were then asked to rank the perceived difficulty of that technique on a Likert scale (a scale from 1 to 5). Users were free to skip individual techniques or end the survey at any time.

Table 2: Techniques for distinguishing between humans and robots

Type		Technique
Active	CAPTCHA	Text
		Audio
		Video
		Confirmation Page
		Unusual Form Interaction
	SAPTCHA	Word
		Image
Passive		Biometric Data Analysis
		Honeypot Form Elements
		JavaScript Detection

3.2.2 Technique Effectiveness

Seeing as developing a bot to attempt to break each technique would have been prohibitive in time allotted for this project, we instead surveyed the existing literature to determine the effectiveness of existing CAPTCHA techniques. For each technique, we recorded the highest verified success rate for bots beating the technique and then took the inverse of this number to represent the technique's success rate for thwarting bots. Thus, the values for each technique's effectiveness are in effect a lower bound.

Some of these success rates were adjusted to account for differences between the technique used in our literature survey and a related technique we applied on the website. For example, in the case of Image SAPTCHA, our method involved asking the user to identify three separate images while existing similar techniques generally only used a single image. To account for this difference we adjusted the success rate to account for the fact the probability of correctly guessing three images is substantially lower than the probability of correctly guessing a single image.

For some techniques (Confirmation Page, Hidden Form Elements, Unusual Form Interaction, and Word SAPTCHA) we were unable to determine the success rates due to a lack of research of bot detection/prevention through these techniques. Further information on how we estimated our success rates of thwarting bots for these techniques is outlined in Section 4.2.

4 Results

4.1 CAPTCHA Efficacy for Humans

The results listed in Table 3 are based on our user generated statistics from the website we launched in November 2011, and which accumulated over 10,000 form submissions over four weeks.

Since passive techniques require no user interaction, we use the minimum possible value for user submitted difficulty. Similarly, because these techniques require no user interaction, we did not record an average completion time.

For the confirmation page, we did not gather statistics on the user submitted difficulty. However, we can assume that because the confirmation page requires the least user interaction of all active techniques, the user submitted difficulty would be at most equivalent to the next easiest active technique (Unusual Form Interaction).

Audio CAPTCHAs are generally the least practical technique. They suffer from having both high difficulty (3.65 out of 5) and a low success rate (29.7%), which likely accounts for why 37.6% of all displayed Audio CAPTCHAs were skipped. Comments on forums on which we advertised our website noted that perceived difficulty was the results of users both not being able to hear the voice correctly and not understanding how many of the spoken terms they were supposed to enter. To keep our results consistent with real-world usage scenarios, we did not provide additional instructions beyond that which exists in the implementation itself.

Text CAPTCHAs were generally considered quite challenging, with a user submitted difficulty of 2.76 out of 5 and an average completion time of 11.3 seconds, which was on par with the overall average for this evaluation criteria. Despite this high perceived difficulty, Text CAPTCHAs still maintained a success rate of 91.3%, which can likely be attributed to their prevalence as the current de facto standard, as well as the high skip rate of 16.9%, which may be due to exceedingly difficult instances.

Unusual Form Interaction achieved the best overall results for active techniques, with the highest observed success rate of 98.4%, the lowest user submitted difficulty of only 1.22 out of 5 and the lowest average solve time of 4.24 seconds.

The remainder of the techniques varied greatly over the four criteria; Video CAPTCHAs had the best success rate but a high average solve time, while Word SAPTCHAs had the lowest skip rate, lowest user submitted difficulty and lowest average solve time, but suffered from a lower success rate. Higher difficulty and average solve time for the Video CAPTCHAs can be attributed to their lack of widespread usage (i.e. most users have never encountered them). For Word SAPTCHAs, we cannot identify a major contributing factor for the low success other than user error.

4.2 CAPTCHA Efficacy for Computers

The results listed in Table 4 represent the lower bound or the average of the various success rates of thwarting bots.

reCAPTCHAs audio CAPTCHA had a high success rate of thwarting Stanfords Decaptcha bot due to the use of semantic vocal noise [3]. This confused the bot in its ability to distinguish between vocal noise, and the actual voice targeted.

The Biometric data was assigned a 50% success rate due to the fact that the bot is either programmed to understand the time threshold, or not. This figure represents a lower bound, as we found other Biometric techniques such as were mouse and keyboard dynamics, and obfuscation of session information, which increased the success rate to 95% [10].

As for Biometrics, the Confirmation Page has a success rate of 50% due to the bots ability to either under-

Table 3: User survey results

Technique	Success [%]	Skipped [%]	Average Difficulty [1-5]	Average time [s]
Audio CAPTCHA	29.74	37.57	3.65	23.84
Biometrics*	100	-	1	-
Confirmation	100	10.11	$\leq$ 1.22**	4.44
Hidden Form Elements*	100	-	1	-
Image SAPTCHA	90.63	13.85	2.14	18.24
Javascript Detection*	97.3	-	1	-
Text CAPTCHA	91.31	16.92	2.76	11.29
Unusual Form Interaction	98.35	8.8	1.22	4.24
Video CAPTCHA	97.29	15.21	2.18	12.42
Word SAPTCHA	92.59	11.05	1.66	7.39

* Passive techniques

** Assumed values based on difficulty for more complex technique

Table 4: Technique effectiveness

Technique	Effectiveness Rate [%]
Audio CAPTCHA*	98.5†
Biometrics	50
Confirmation Page	50
Hidden Form Elements	50
Image SATPCHA	78.4†
JavaScript Detection	18
Text CAPTCHA*	82.5†
Unusual Form Interaction	50
Video CAPTCHA**	$\leq$ 82.5
Word Problems	60†

* Using reCAPTCHA library

** Using NuCAPTCHA library

† Lower bound

stand what a confirmation page is and how to treat it, or not.

The hidden form elements follow the same pattern as the Biometrics with its 50% success rate. In order for a bot to achieve a 100% success rate, it will need to be programmed to understand different ways of hiding form elements through CSS or HTML attributes.

The success rate for thwarting bots using Image SAPTCHAs is derived by using the formula:

$$1 - P(success)n$$

Where $P(success)$ is the probability that the bot can properly recognize the image, and n is the number of images used in the same SAPTCHA. In our case, $P(success)$ is 59.36% which is based on six possibilites for an image [9], and $n = 3$.

The success rate of thwarting bots using a test for determining whether the user had JavaScript enabled was 18% [8]. This detection rate was confirmed for bots that were unable to simulate mouse movements.

Like the other passive techniques, Unusual Form Interaction has a 50% success rate because the bot is either equipped with the knowledge of interpreting the form interaction instructions, or simply does not know how to process them.

Recent advances in bot development by research Jonathan Wilkins has led to a bot success rate of 17.5% against an older version of Googles text-based reCAPTCHA [7]. Although the success rate against the newer implementation of reCAPTCHA is higher, these new findings are unverified and therefore not included in our analysis.

The Video SAPTCHA has a success rate that is less than or equal to that of a Text CAPTCHA. This is beacuse a video can be treated as a composition of many images, and therefore a bot would have several datapoints to use when attempting to determine the embedded text.

To evaluate the success rate of the word problems, we used a naive bot simulation technique where the question or statement was entered as a search term in Google, and the first result was treated as the answer. 60% of the time, the first search result did not contain a valid answer or solution to the word problem.

4.3 Rankings

The final rankings for existing CAPTCHA techniques are found in Table 5, which are the result of our evaluation rubric and evaluation equation from Section 3.1. The maximum score any technique can receive is 16, showing significant room for improvement for all existing techniques. The individual score for user submitted difficulty has been readjusted to account for user preference of 1 representing easiest, whereas the chosen evaluation equation requires 1 to represent hardest.

Video and Text CAPTCHAs suffer from steadily decreasing security at the hands of increasing efforts by bot developers. As noted in Section 4.2, recent advances in developing bots has led to Text CAPTCHAs having only an 82.5% bot detection rate [7] and, as outlined earlier, Video CAPTCHAs are assumed to be at most as secure as Text CAPTCHAs. The difference between the rankings for these techniques is largely due to Video CAPTCHAs having a higher user sucess rate and a lower user submitted difficulty.

Putting development efforts into building libraries for other techniques including Image SAPTCHAs and Unusual Form Interaction would have led to these techniques being on par with Video and Text CAPTCHAs, however Image SAPTCHAs require a large bank of images in order to mitigate the effectiveness of a Pictionary Attack. In this attack, a bot maintains a lookup table of images and their associated information, which it constantly updates with newly encountered images.

Among passive techniques, collecting and analyzing Biometric data proved to be far superior to all other techniques. The largest drawback of collecting biometric data is that no publicly available libraries exists, which forces website administrators to either purchase commercial software or develop their own in-house solution, which could lead to a sub-optimal effectiveness at detecting bots.

Table 5: Overall results

Rank	Technique	Score*
1	Video CAPTCHA	12.51
2	Biometrics	12.08
3	Text CAPTCHA	11.7
4	Unusual Form Interaction	9.28
5	Image SAPTCHA	8.93
6	Hidden Form Elements	8.84
7	Javascript Detection	8.01
8	Word SAPTCHA	7.02
9	Confirmation Page	6.8
10	Audio CAPTCHA	3.69

* Maximum score is 16

5 Proposed Hybrid Technique

In order for a technique to be effective, it must be easy for humans to implement and solve, but difficult for bots to understand. In order to fulfill these criteria, we chose to combine the best active technique with the best overall technique for detecting bots.

As noted in Section 4.3, the best overall technique is currently Video CAPTCHA, however this technique suffers from a potentially low bot detection rate. Our best technique at detecting bots was Audio CAPTCHAs, however for technical reasons it would be infeasible to combine these two active techniques together. Instead, we propose combining Video CAPTCHAs (the best overall active technique) with collecting and analyzing Biometric data (the most effective passive technique). Our projected results are displayed in Table 5.

By employing Bayesian reasoning, we may assume that a bot's ability to defeat one of these two techniques does not impact its ability to defeat the other as well. Additionally, by combining Video CAPTCHA with a passive technique, there is no additional increase in perceived difficulty or solution time for the user. Although theoretically superior, our new proposed technique requires further testing to verify our estimated effectiveness.

Table 6: Estimated score of proposed technique

Criteria	Video CAPTCHA	Biometrics	Proposed Technique
Effectiveness	82.50%	95%	99.12%**
Libary Availability	3	1	2
Automation	3	3	3
Time Required	12.28 (s)	-	12.28 (s)*
Difficulty	2.18	-	2.18*
Cancellation Rate	15.21%	-	15.21%*
Success Rate	92.71%	100%	97.21%**
Score	12.51	12.08	13.2

* Calculated based on Biometrics being a passive technique and therefore having a negligible affect on the users experience

** Calculated using Bayesian reasoning

6 Conclusion

In this paper, we collected and analyzed data pertaining to the 10 leading techniques of distinguishing between humans and robots on the web. First, we implemented the techniques in a website, and recorded the results of our experience. Second, we invited users to participate in submitting forms implementing each technique, and gathered data relating to their submissions. Third, we compiled statistics regarding the different techniques' effectiveness at distinguishing between humans and robots. Fourth, we proposed a rubric for evaluating each technique, as well as an equation for assigning a final score to each according to the individual criteria. Finally, we proposed a combination of two different techniques with the highest scores, and presented a preliminary evaluation.

Our results show that while many techniques exist, each with its own strengths and weaknesses, no single technique has a clear advantage over all the others. And our proposed hybrid technique, while performing well according to our evaluation criteria, requires further study.

References

1 R. Abrich, V. Berbenetz and M. Thorpe. The CAPTCHA Experiment. Retrieved from: http://thecaptchaexperimenet.com

2 E. Bursztein and S. Bethard. Breaking 75% of eBay Audio CAPTCHAs. In 3rd USENIX Workshop on Offensive Technologies, pages 1-7, February, 2009

3 E. Bursztein et al. The Failure of Noise-Based Non-Continuous Audio CAPTCHAs. 2011 IEEE Symposium on Security and Privacy. pages 19-31, May 22-25, 2011

4 M. Fellet. Stanford computer scientists find Internet security flaw. Stanford News Service. Retrieved from: http://news.stanford.edu/pr/2011/pr-captcha-security-flaw-052311.html

5 C. Fidas et al. On the Necessity of User-Friendly CAPTCHA. In the Human factors in Computing Systems conference, pages 2623-2626, May 2011.

6 M. Luk. History. Palo Alto Research Center. Retrieved from: http://www2.parc.com/istl/projects/captcha/history.htm

7 J.P. Mello Jr. Google reCAPTCHA cracked. January 5, 2011. http://www.allspammedup.com/2011/01/google-recaptcha-cracked/

8 K. Park et al. Securing Web Service by Automatic Robot Detection. USENIX 2006 Annual Technical Conference Referred Paper. http://www.usenix.org/event/usenix06/tech/full_papers/park/park_html/

9 A. Raj et al. Picture CAPTCHAs With Sequencing: Their Types and Analysis. In the International Journal of Digital Society, Volume 1, Issue 3, pages 208-220, September 2010.

10 D. Stefan et al. Robustness of keystroke-dynamics based biometrics against synthetic forgeries. 2010 6th International Conference on Collaborative Computing: Networking, Applications, and Work Sharing. pages 1-8, Oct 9-12, 2010.