A REVIEW OF USAGE & SECURITY OF CAPTCHAS TO PROTECT WEB RESOURCES AND WEB SERVICE'S PERFORMANCE

[1]Mehul S. Patel, [2]Manish I. Patel
[1]PG Scholar, [2]Assistant Professor
[1]Department of Information Technology, [2]Department of Electronics & Communications,
[1,2] Sankalchand Patel College of Engineering, Visnagar, India

Abstract - To protect web resources and web services from computer bot program, various types of CAPTCHAs are used. In all types of CAPTCHAs, Text-Based CAPTCHA is widely used by many web service providers. In this review paper, different types, applications, security issues and features about the CAPTCHA system are discussed, including the techniques involved. As a performance parameter, the breaking ratio is summarized from various research articles.

Index Terms—CAPTCHA, Text-Based CAPTCHA, Segmentation, Recognition, CAPTCHA Breaking Ratio

I. INTRODUCTION:

Day to day, the usage of internet becomes a part of our daily life. Many websites and web applications on advance smart phone usage has increased drastically in the last few years. Due to the large usage of web services as well as increase in the number of users, performance of web resources becomes the current challenging issue. In such environment, if resources are wrongly engaged by dummy users or unwanted requests are generated frequently by any computer program, then web service may not correctly perform the required operations and also not possible to respond within the time. To protect web resources and to maintain performance of web services, researcher has developed a CAPTCHA (Completely Automated Turing test to Tell Computer and Human Apart) security in which human identify the secure code or solve a puzzle asked to web users which is very difficult to solve by any computer program within some time limit.

CAPTCHA is used to protect resource utilization and for performance from bot program. CAPTCHA is also known as a Human Interaction Proofs (HIP) [4]. In current website trend, CAPTCHA becomes an important part of many web services as security. In CAPTCHA technique, different types of CAPTCHAs are automatically generated by a computer program in such a way that it's more difficult to solve by computer bot program on user side but easily solved by humans.

Innovation in character recognition is an issue for CAPTCHA security. Using character recognition techniques, lots of bot programs developed to break the CAPTCHA with a high breaking ratio; still it is a more effective security tool to protect web resources and web services from bot program [3]. In 1997, DEC Systems Research Center developed the design of CAPTCHA to block automatic submission for AltaVista website [4]. Since then lots of developers designed CAPTCHA in many ways for different website to protect services on the web server. Feasibility of CAPTCHA is provided in Section II. Section III introduces the types of CAPTCHA. Summary of all types of CAPTCHA is presented in Section IV. Section V describes about the various applications of CAPTCHA. Section VI highlights the main CAPTCHA breaking issue. A CAPTCHA breaking procedure steps mentioned in Section VII. Security features of CAPTCHA highlighted in Section VIII. Review of research articles of attack on Text-Based CAPTCHA are summarized in Section IX. Section X concludes the review of Text-Based CAPTCHA security and issue.

II. FEASIBILITY OF CAPTCHA:

CAPTCHA is considered as secured if it can be solved successfully by humans more than 90% and it cannot be solved by computer more than 1% [1]. In Fig. 1, Sweet spot represents the scope of CAPTCHA algorithm to generate efficiently secure code. Outside the Sweet spot boundary, CAPCTHA code is not readable or easily breakable by a computer program.

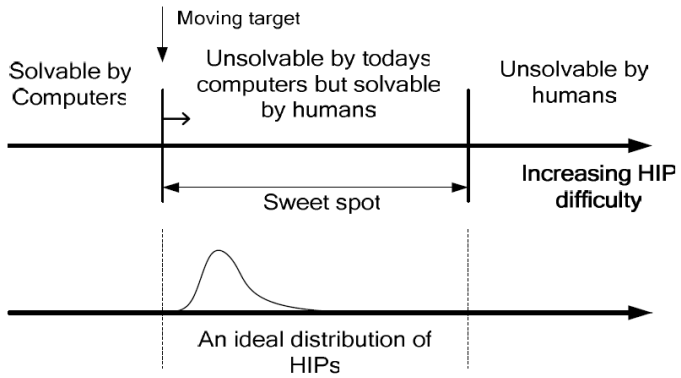

Fig. 1: Regions of feasibility as a function of HIP difficulty for humans and computer algorithms [4]

Development of technology challenges the Sweet spot area shrink day to day. Left side boundary in Sweet spot shifted towards the right side because of efficient algorithms of Machine Learning to recognize text. So CAPTCHA developer introduces more complex techniques to confuse the computer bot program, but still possible to be identified by a human.

Initially CAPTCHA security techniques start with Text-Based CAPTCHA and as time gone many different techniques are developed by researcher to secure the web resources and web services from the computer bot program.

III. TYPES OF CAPTCHA:

1. Text-Based CAPTCHA:

In this technique, an image of secure code is included in Web pages, which is required to be identified by human and enter the same security code in giving the text box to submit the web page successfully. Otherwise, the web request is rejected by server process and protect from malicious attack. Normally Text-Based CAPTCHA uses English letters and digits to generate secure code. It is better to avoid local letters and local digits in secure code for wide usability of an application or service over the world for different kinds of users. Security of Text-Based CAPTCHA is again a challenge for developers to protect from computer bot programs because of the enhancement in Machine Learning algorithms for character recognition. Security of Text-Based CAPCTHA is more rely on segmentation and recognition. If secure code introduces the complex distortions like rotation, scaling, transformation, then it becomes more difficult for a computer bot program to detect secure code. Using the Machine Learning algorithm, like K-Nearest Neighbors, Neural Network, it is possible to break CAPCTHA efficiently with good success ratio. To secure CAPTCHAs from segmentation, it is recommended to use multilayer text and overlapped letters preferred with variations in character size and rotation angle in secure code. Text-based CAPTCHAs are used by many websites such as Yahoo, Gmail, YouTube, PayPal, Gimpy, Ex-Gimpy, Baffle-Text, MSN-CAPTCHA etc. Fig. 1 shows the simple example of Text-Based CAPTCHA, but it is possible to design more complex to confuse the computer bot program.

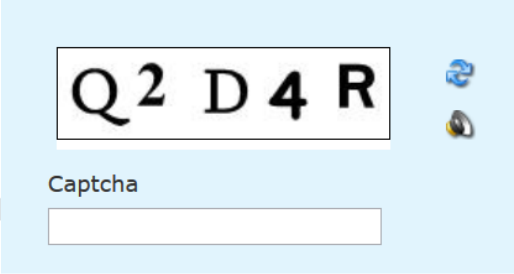

Figure 1: Text-Based CAPTCHA

2. Image-Based CAPTCHA:

In Image-Based CAPTCHA, many images are visualized to the user and asking to select some images based on the question. Pix and Bongo CAPCTHAs are examples of Image-Based CAPTCHA. In Image-Based CAPTCHA, the challenge is to keep lots of images in the database which requires more storage and also not secure compared to Text-Based CAPTCHA because of finite questions related to images. Using Object Detection algorithm in Machine Learning, it is easy to break the Image-Based CAPTCHA for targeting any pattern. Fig. 2 shows the no of images and asked to user, select image of a flower. If the user selects the correct image, then the web page request is processed by web application otherwise rejected to protect web resources and web services from malicious computer bot program.

Figure 2: Image-Based CAPTCHA

3. Audio-Based CAPTCHA:

In Audio-Based CAPTCHA, user can play audio and accordingly it is required to enter the word in the given textbox. It is audio device based systems, so application required rights to access audio device which may be not available in a public place like cybercafé. So it creates difficulty for public web application users to submit a web page successfully. Also web application required to generate and keep thousands of sound clips to avoid the brute force attack. Because of Text to Speech & Speech to Text conversion algorithm evaluated efficiently in the last few years, it is easy to break the audio-based CAPTCHA efficiently with good success ratio. In fig. 3 the Audio-based CAPTCHA is shown.

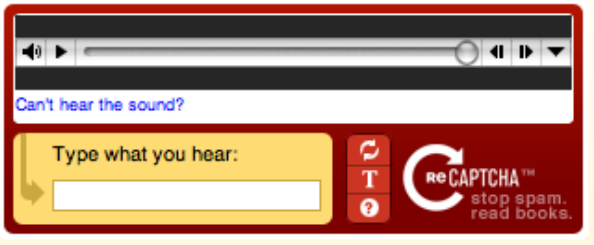

Figure 3: Audio-Based CAPTCHA

4. Video-Based CAPTCHA:

In Video-Based CAPTCHA, a video contains different words, user play and watch those words and accordingly it is required to enter them in the given textbox. Difficulty of video-based CAPCTHA is to generate and keep a large set of videos which requires more storage compared to all other methods. If the user's browser is not supported for video plug-in then it's difficult for the user to submit the web pages. In fig. 4 Video-Based CAPTCHA is shown.

Figure 4: Video-Based CAPTCHA

5. Puzzle-Based CAPTCHA:

In Puzzle-Based CAPTCHA, chunks of single image displayed on a web page and it is asked to the user for rearrangement of all chunks to create a complete picture. In some applications a numeric or text related puzzle is also asked to user. User can solve the puzzle and submit the answer. In this technique, images and chunks of images occupies the large size of memory. It is very easy to solve the numerical or text puzzle using advanced technologies such as Googles OneBox (instance answer). Old online LIC and UGVCL sites used the text puzzle-based CAPTCHA. Fig. 5 shows the Puzzle-Based CAPTCHA.

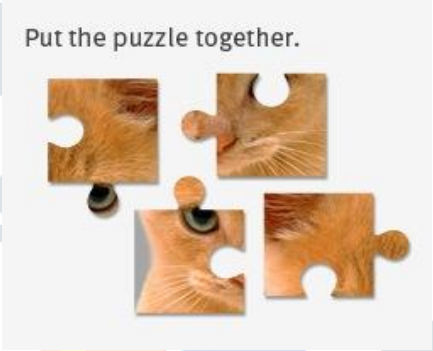

Figure 5: Puzzle-Based CAPTCHA

6. noCAPTCHA - reCAPTCHA:

noCAPTCHA – reCAPTCHA is a new method of Google to block unwanted requests from visitors. This method is a behavior analysis method of human activities on the web page. When user clicks on "I am not a robot", Java script submits the user movement, cookies, and event information to the Google analysis server. Behavior analysis server collects all data provided by client script for analysis and returns encoded value depending on user and time, it's indicate, the user is verified or not. In case of confusion, Google server asks for image-check to verify the user further. As a limitation for other web applications, excluding Google, all the web traffic detail will be submitted to Google analytic server which may not be preferred as privacy terms. In following fig. 6 Google's noCAPTCHA – reCAPTCHA is shown.

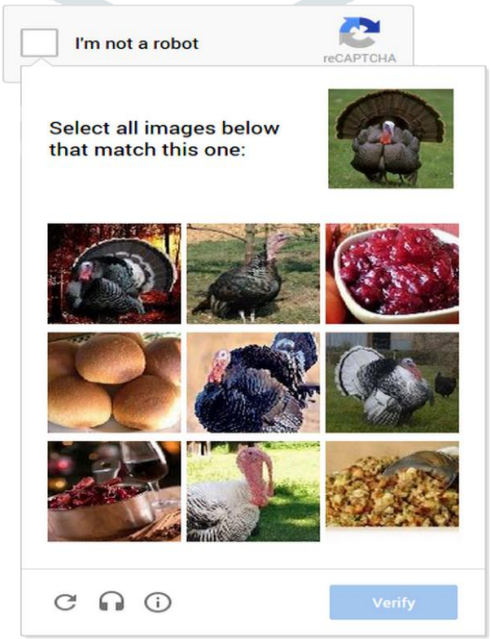

Figure – 6: noCAPTCHA reCAPTCHA

IV. SUMMARY FOR TYPES OF CAPTCHAS:

Based on a study of different methods of CAPTCHA, important points are identified in Table - I to be considered before using that method.

Table - I: Summary of CAPTCHA methods

Type of CAPTCHA	Drawback	Security	Usability	Implementation
Text-Based	Confusing Character, Identified by OCR	Good	Average	Easy
Image-Based	Image identification	Good	Easy	Difficult
Audio-Based	Language	Good	Difficult	Difficult
Video-Based	High bandwidth	Good	Difficult	Difficult
Puzzle-Based	More time to solve	Average	Average	Average
noCAPTCHA-reCAPTCHA	Third party	Good	Easy	Easy

Implementation and security point of view Text-Based CAPTCHA is widely used compared to other CAPTCHA schemes [1]. Following list represents some reasons for widely usage of Text-Based CAPTCHA

1. Do not require large memory to store the sets of CAPTCHAs like others.
2. Easily possible to generate different CAPTCHAs on demands.
3. Possible to make it complex to secure from bot program.

Following Table - II shows the CAPTCHA solving time by users of different web services.

Table- II: Expected solving time by human [13]

Service	Time	Accuracy	Expected Time
Authorize	6.8	0.98	6.9
Baidu	7.1	0.93	7.6
Captchas.net	8.2	0.84	9.8
Digg	8.2	0.92	8.9
eBay	7.3	0.93	7.8
Google	9.7	0.86	11.3
mail.ru	12.8	0.7	18.3
Microsoft	13	0.8	16.3
Recaptcha	11.9	0.75	15.8
Skyrock	7.9	0.95	8.3
Slashdot	7.7	0.87	8.8
Blizzard	9.3	0.95	9.8
Yahoo	10.6	0.88	12
Authorize audio	11.9	0.59	20.2
Digg audio	14.8	0.38	39
eBay audio	11.8	0.63	18.8
Google audio	35.0	0.35	100.6
Microsoft audio	16.6	0.38	43.8
Recaptcha audio	30.1	0.47	64.1
Slashdot audio	11.7	0.68	17.2
Yahoo audio	25	0.68	36.8

V. APPLICATIONS OF CAPTCHA:

Day to Day, the use of CAPTCHA in websites increases and becomes necessary for the security of web resources. Mostly usage of CAPTCHA is categorized as follows.

1. **Secure website registration application:**
   Many websites offer free services like email, doc storage, calendar etc. For daily necessary task related to the normal real life of thousands of users. To attack on website resources, the computer bot program tries to register as a fake user with dummy data and create a lot of accounts unnecessarily to engage the web resources and also degrade the service performance like response time of real user's request. So it is really necessary to stop those malicious users or bot program using advance techniques like CAPTCHA.
2. **E-Pooling:**
   Public opinion is found useful for development of future product. If products are designed or improves, according to user's expectation, then it's really more useful, for users and profit making for developers like win-win strategy. To collect public opinion for different survey is normally possible using the web application in large scale of users. But it is also risky if malicious opinion submitted using computer program. To protect auto polling from the computer bot program, CAPTCHA plays a vital role to stop bogus voting in most web applications.
3. **Anti-Crawling:**
   Many search engines crawl the web pages of websites without any permission or inform to the owner. It is also used to analyze the data from different websites. In case of protecting web pages from indexed by search engine, web pages introduced with CAPTCHA to restrict the web pages display only to the human.

VI. CAPTCHA BREAKING ISSUES:

1. **Segmentation:**
   The most difficult task of breaking CAPTCHA is to divide the CAPTCHA in different letters or digits to recognize correctly. Using edge detection and color variation techniques, it can be possible to divide the CAPTCHA. Because of overlapping, rotation, extra noise and distortion of the text, it becomes more difficult for a computer to decode the correct CAPTCHA.
2. **CAPTCHA usability:**
   To secure CAPTCHA, some distortion methods are applied to text to make it more difficult for a computer to break the CAPTCHA automatically. It is also true that distorted text also difficult for the end user to read and decode. Based on text readability, performance is measured by the response time to submit the CAPTCHA. Accuracy of the method is measured based on how many times users try to submit CAPTCHA successfully.

VII. CAPTCHA BREAKING PROCEDURE:

Figure 6 represents the five operations, that suggested by researcher to break the CAPTCHA in a research article [2]. 1) Pre-processing – remove the background pattern or noise 2) segmentation – divide the image in small parts as a single character 3) post-segmentation – cleanup each segment by normalizing 4) recognition – identifying individual character and 5) post-processing – order of characters or spell checking.

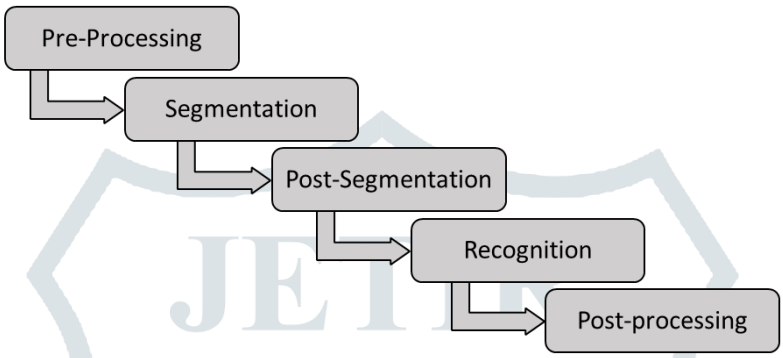

Figure 6: CAPTCHA solving process

VIII. CAPTCHA'S SECURITY FEATURES:

1. **Anti-Segmentation Features:**
   Widely used features to protect segmentation are:
   Background Complexity – Use of Complex image as a background, add lines and curves randomly as background design like a net or color combination to confuse the character recognition.
   Lines & Curves – adding lines and curve randomly to misguide the character recognition methods, such as large lines with similar width of character, line over the some characters of CAPTCHA, change slope of lines, use the same color for line as letter color, randomize the length of lines etc.
   Character Overlapping – Collapsing character on each other are most effective technique in anti-segmentation, but more collapsing letter also creates confusion for human.
   Confusing Character Set - Confusing character like 'i' and 'j', 'O' and digit '0', create more confusion for bot program but it is also creating difficulty for human.
   Following Table – III represents CAPCTCHA's features use of different services.

Table – III: Image CAPTCHA Features [13]

Service	Min. Length	Max. Length	Character Set	Word as a Code
Authorize	5	5	a0	No
Baidu	4	4	0A	No
Captchas.net	6	6	a	No
Digg	5	5	A	No
eBay	6	6	0	No
Google	5	10	a	Pseudo
mail.ru	6	6	0aA	No
Microsoft	8	8	0A	No
Recaptcha	5	20	0aA-_!	Yes
Skyrock	5	6	a0	No
Slashdot	6	8	a	Yes
Blizzard	6	8	a0	No
Yahoo	5	8	0aA	No

2. **Anti-Recognition Features:**
   Widely used features to protect recognition are:
   Length – Instead of fixed length, dynamic length of CAPTCHA is more effective.
   Character Set– Usage of an international alphabet and digits are better to avoid localization problems.
   Font Family & Size – Variation in font family or font size is better to create a more difficulty to recognition by computer.
   Character Distortion & Blurring - Spreading character boundary and the blurring will create more confusion to identify the character classification.
   Character Rotation – Each character rotated with different angle will create a more confusion for text recognition.

Waving the character – Waving the each character in a different way will create a more confusion for a bot program to find the boundary.

3. **Recognition/Classification:**
Widely used classifiers are Support Vector Machine (SVM) which is better performed in distorted character, K - Neural Network (KNN) which is better performed in a mixture of different fonts, Convolutional Neural Network (CNN) which is better performed for rotated character [2].

IX. SUMMARY OF REVIEW PAPERS FOR CAPTCHA BREAKING RATIO:

Based on the different research articles, it observed the following results to break the CAPTCHA using various techniques.

1. Microsoft two layer CAPTCHA successfully broken up to 44.6% within 9.05s using LeNet-5 technic [1].
2. Achieved more than 50% broken ratio for various sites CAPTCHA using the DECAPTCHA tool with SVM & KNN algorithm [2].
3. Up to 77% success broken ratio is achieved within the 15s using 2D Log-Gabor filters [3].
4. Microsoft CAPTCHA achieved broken ratio up to average 90% in 80ms using CFS algorithm [6].
5. Hollow CAPTCHA was broken successfully up to 89% using CNN and CFS algorithm [7].
6. Achieved satisfactory broken ratio using Faster R-CNN algorithm [8].
7. Identified the word in an EZGimpy image with a success rate of 92%, and the requisite words in a Gimpy image 33% of the time [11].

X. CONCLUSION:

Based on the review of different research papers on security of CAPTCHA, it is observed that the Text-Based CAPTCHA is easy and secure method compared to others. Because of advances Text Recognition techniques, security of Text-Based CAPTCHA is a current issue. Some of the techniques can successfully break the Text-Based CAPTCHA with very high success rate. But still many text based CAPTCHAs like Google, Bing etc. are difficult to break with good success rate. SVN, KNN and CNN are the most efficient methods to break Text-Based CAPTCHAs. CNN is more complex than other methods, but also have a good success rate to break any text based CAPTCHAs compared to other methods. Still more efficient techniques are in demand to design a CAPCTHA to protect the web resources and performance of web services from the computer bot program.

REFERENCES

[1] Haichang Gao, Mengyun Tang, Yi Liu, Ping Zhang, and Xiyang Liu, "Research on the Security of Microsoft's Two-Layer Captcha", IEEE Transactions On Information Forensics And Security, vol. 12, no. 7, pp. 1671-1685, 2017

[2] E. Bursztein, M. Martin, and J. Mitchell, "Text-based CAPTCHA strengths and weaknesses," in Proc. 18th ACM Conf. Comput. Commun. Secur., pp. 125–138, 2011

[3] K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Building segmentation based human-friendly human interaction proofs (HIPs)," in Human Interactive Proofs. Heidelberg, Germany: Springer, pp. 1–26, 2005

[4] J. Yan and A. S. El Ahmad, "Usability of CAPTCHAs or usability issues in CAPTCHA design," in Proc. 4th Symp. Usable Privacy Secur., pp. 44–52, 2008

[5] H. Gao et al., "A simple generic attack on text CAPTCHAs," in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS), San Diego, CA, USA, pp. 1–14 , 2016

[6] J. Yan and A. S. El Ahmad, "A low-cost attack on a Microsoft CAPTCHA," in Proc. 15th ACM Conf. Comput. Commun. Secur.,pp. 543–554, 2008

[7] H. Gao, W.Wang, J. Qi, X. Wang, X. Liu, and J. Yan, "The robustness of hollow CAPTCHAs," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 1075–1086, 2013

[8] Feng-Lin Du, Jia-Xing Li, Zhi Yang, Peng Chen, Bing Wang, and Jun Zhang, "CAPTCHA Recognition Based on Faster R-CNN" in Springer International Publishing ICIC, Part II, LNCS 10362, pp. 597–605, 2017

[9] Ye Wang, Yuanjiang Huang, Wu Zheng, Zhi Zhou, Debin Liu, Mi Lu, "Combining convolutional neural network and self-adaptive algorithm to defeat synthetic multi-digit text-based CAPTCHA" in IEEE International Conference on Industrial Technology (ICIT), pp. 980- 985, 2017

[10] Gaihuan An, Wanjun Yu, "CAPTCHA Recognition Algorithm Based on the Relative Shape Context and Point Pattern Matching" in International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. -168-172, , 2017

[11] G. Mori, J. Malik. "Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA", Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pp. 134-141,2003

[12] Baird HS and Luk M, "Protecting Websites with Reading-Based CAPTCHAs", Second International Web Document Analysis Workshop (WDA'03), Edinburgh, Scotland, pp. 53-56, 2003

[13] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell, Dan Jurafsky, "How Good are Humans at Solving CAPTCHAs- A Large Scale Evaluation", IEEE Symposium on Security and Privacy (SP), Berkeley/Oakland, CA, USA, 2010

[14] Kiranjot Kaur, Sunny Behal, "Captcha and Its Techniques- A Review", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , pp. 6341-6344, 2014

[15] Ved Prakash Singh, Preet Pal, "Survey of Different Types of CAPTCHA", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , pp. 2242-2245, 2014