Globus An International Journal of Management & IT
A Refereed Research Journal
Vol 10 / No 2 / Jan-Jun 2019 ISSN: 0975-721X

CODE BREAKING APPROACH IN AUTOMATED CHARACTER RECOGNITION

*Marites Dumay-Manganti

Abstract

The Code Breaking Approach in Automated Character Recognition is a system that tests the security of free social networking web site CAPTCHA images. It compares the efficiency and accuracy of two methods, the parts-based and segmentation which serve as a guide and also a basis for further studies in AI community. The researcher has employed both experimental and descriptive design by using online generated data samples which is used for gathering accuracy and efficiency result of parts-based and segmentation method. The findings shows that the over-all result in comparing the accuracy and efficiency of two methods have no significant difference using the file size range of a CAPTCHA image. On the other hand, the study still got efficient and accurate results which show that the system is reliable in testing the accuracy and efficiency in identifying image code using the two methods. Therefore, the actual implementation of the Code Breaking Approach in Automated Character Recognition shall be recommended to be used in improving internet security, especially in free email services and social networking websites that provides security and authentication to its user.

Keywords: Code Breaking, Parts-Based, Segmentation, Algorithms, Automated, Character Recognition, CAPTCHA, Computer Security.

Background of the Study

In computing, there have been a lot of challenges for the Artificial Intelligence (AI) community, programmers, and other researchers. One of the examples is dealing with optical character recognition, the capability of software to read text or alphanumeric characters from a scanned image of printed, handwritten, or type written sources. It is one of the problems that was solved by using Artificial Intelligence concepts and even though it is not easy to make a one hundred percent running program, programmers and researchers are trying to prove that people of today can make a difference and can change the way what Artificial Intelligence should be in this computing age.

There are so many instances that happened in the internet nowadays. Some people create programs for the benefit of the users and some are trying to harm others which probably would not be acceptable to the society. Why would anyone need to create a test that can tell humans and computers apart? That automated program could be part of a larger attempt to send out spam mail to millions of people. That is why programmers create a solution to this problem, which is called C**ompletely Automated Public Turing Test to Tell Computers and Humans Apart** (CAPTCHA) test.

The concept of Code Breaking Approach in Automated Character Recognition, which would test how secure are the internet or social networking websites which offers free email services and also have different authentications have been initially originated in the study of CAPTCHA which was introduced in the study of Luis Von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford (2003), an automated test that humans can pass, but current computer programs can't pass: any program that has high success over a captcha can be used to solve an unsolved Artificial Intelligence (AI) problem. In this study, they have provided several unique captchas and also the different applications in solving problems that can be exploited for security purposes with captchas applications for web security. Also, it is cited that the study is much likely a research in cryptography which has a positive impact on algorithms for factoring and discrete log which allows the advancement in the field of Artificial Intelligence.

*Carlos Hilado Memorial State College, College of Engineering and Information Technology, Alijis Campus, Bacolod City, Philippines

Review of Related Literature

The purpose of this section is to provide the reader with a broader understanding of the study which includes foreign and local related studies and literatures from published books, journals, and other related writings which is significant to the development of this study.

Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction by J. M. White and G. D. Rohrer (1983) concepts using two algorithms in extracting binary images of characters form machine or hand-printed documents. Algorithms were utilized were in thresholding is a critical step in optical recognition (OCR) and also essential in character image extraction (CIE) applications in processing machine printed or handwritten characters. Algorithms like nonlinear, adaptive procedure, is implemented with a minimum of hardware and is intended for many CIE applications in this study. The other is used for more aggressive approach to answer a more complex and specialized applications.

Mauldin, M.(1994) studied "CHATTERBOTS, TINYMUDS, and the Turing Test Entering the Loebner Prize Competition" The Turing Test was proposed by Alan Turing in 1950; he called it the Imitation Game. Hugh Loebner started the Loebner prize competition in 1991, offering a $100,000 prize to the author of the first computer program to pass an unrestricted Turing test. This also describes the development and technical design that was used in the competition with continuous interaction with the internet via Tinymuds (multiuser network communication servers) with the advancement of the Artificial Intelligence.

An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities by Phamab,D.L. and Prince,J.L.(1999) present "a novel algorithm for obtaining fuzzy segmentations of images that are subject to multiplicative intensity in homogeneities, such as magnetic resonance images". The algorithm was formulated in changing the objective function in the fuzzy C-means algorithm in varying characteristics of images and also the iterative algorithm which also minimizes the function and demonstrate the efficacy in testing several images.

In the study of G. Mori and J. Malik, (2003), they have explored object recognition in clutter. They have used techniques in object recognition on visual CAPTCHA such as Gimpy and EZ-Gimpy. A CAPTCHA ("Completely Automated Public Turing test to Tell Computers and Humans Apart") is a program that can generate and grade tests that most humans can pass, yet current computer programs can't pass. Currently, yahoo have used EZ-Gimpy and the other one is based on word recognition in the presence of clutter is Gimpy CAPTCHA. These were utilized because of its adversarial character which can confuse computer programs. They also have developed efficient methods based on shape context matching that has a success rate of 92% in identifying using EZ-Gimpy while 33% was identified in a Gimpy image. Object recognition using this methods gives insights to the researcher to develop a more complex systems.

According to the study of Bursztein, E., et.al (2011), Text-based CAPTCHA Strengths and Weaknesses, they have included a systematic study of existing visual CAPTCHA based on distorted characters that are augmented with anti-segmentation techniques. This study focuses on the vulnerability of automated attacks which they have gathered CAPTCHAs in different web sites which shows that out of 15 current CAPTCHA, 13 are vulnerable to attacks. This study also shows that reliable authentication is needed to distinguish human and computers.

A Study of CAPTCHAs for Securing Web Services by M. Tariq and Banday, N. A. Shah (2011) also discusses different security and usability issues which merely focuses on the review of the existing CAPTCHA schemes which are being utilized. They also cited that they have grouped the different CAPTCHA in terms of its security and usability which general method was used to generate and break text-based and image-based CAPTCHAs.

In the study of Todorov,A. (2014) Securing passwords with CAPTCHA based hash when used over the web -"A password security system, hosted by a server, sends a web page over a network to a client, that includes a CAPTCHA challenge, a request for a CAPTCHA answer, a graphical user interface for receiving a user identifier and a password, and a security script". Security script has been utilized in this study by using CAPTCHA in password authentication. Client hash value is compared with the server to determine the correct password of the client.

The presented literatures gave significant insights and ideas in the development of this study, also with the concepts and algorithms that is comparable to the current and existing system which serves as the bases for the process of the tool to be made.

Statement of the Problem

This study aims to develop a tool that will break an image code through Automated Character

Recognition. Specifically, this seeks to answer the following questions:

1. What is the efficiency of the program in identifying characters of an image code using:
   a. Parts-Based Approach
   b. Segmentation
2. What is the extent of accuracy of the program in identifying characters of an image code using:
   a. Parts-Based Approach
   b. Segmentation
3. Is there a significant difference in the accuracy and efficiency of the program in identifying characters of an image code using Parts-Based and Segmentation Approach?

Hypothesis

There is no significant difference in the accuracy and efficiency of the program in identifying characters of an image code using Parts-Based Approach and Segmentation.

Theoretical and Operational Framework

The selection of the most adequate concepts and efficient tools is imperative in the proposal of the Code Breaking Approach in Automated Character Recognition. Thus, the following are the essential concepts which contribute to the development and implementation of the tool.

Figure 1 shows the relationship among the approaches and the process in the development of the study. Thus, the concept of processes presented in the figure plays a significant role in developing the tool in code breaking. The study shows different methods and concepts from Turing Test, Issues about code breaking, and lastly the Code Design and Breaking Approaches.

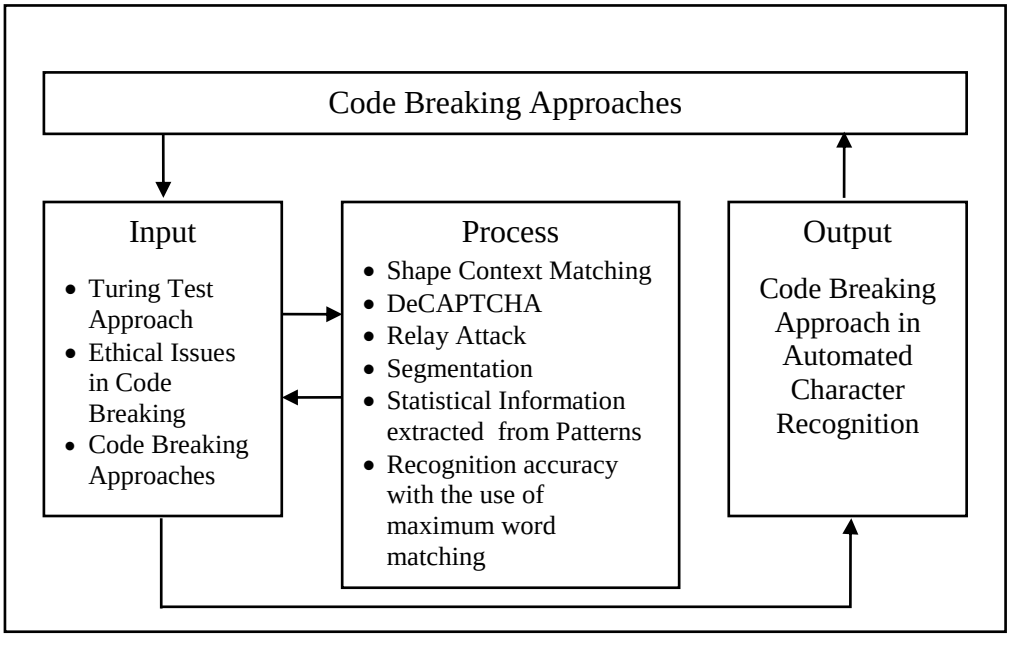


Figure 1.Schematic Diagram of the Theoretical/Conceptual Framework.

Figure 2 shows how the independent variable, dependent or outcome variable, and the system of the study can affect each other. The independent variables consist of the methods use such as Parts-Based Approach and Segmentation. These are the factors that affects the dependent or outcome variable which refers to the efficiency and accuracy of the system in identifying characters in a visual CAPTCHA.

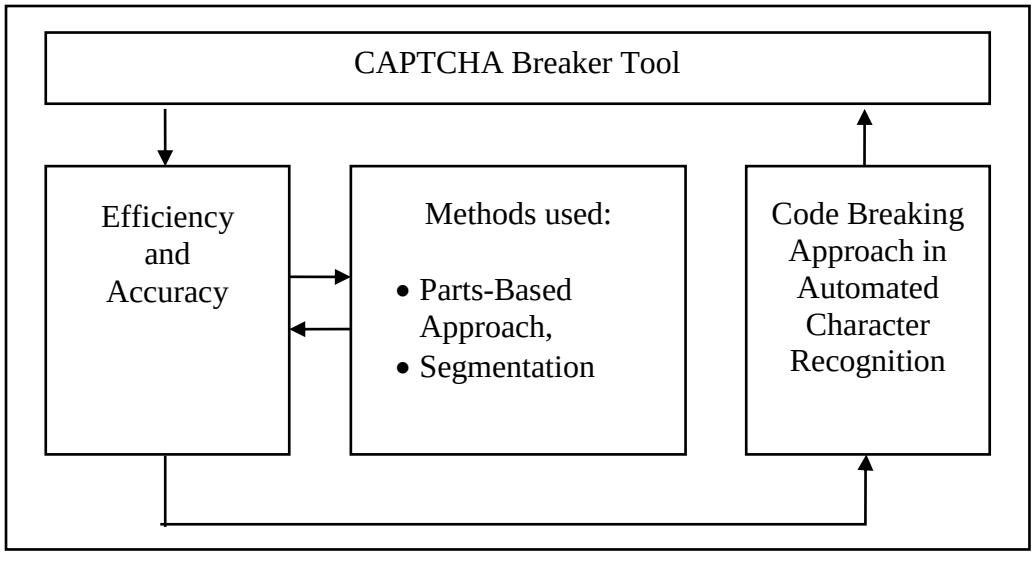


Figure 2. Schematic Diagram of the Operational Framework.

Research Methodology

This chapter provides relevant information about the methods of research utilize in this study. It presents a series of discussion about the research methods, data sets of the study, data-gathering procedure, measures, statistical tools and treatment, and data analysis.

Research Design

The researcher used quantitative research specifically, experimental and descriptive research design. It aims to determine the relationship between an independent variable and dependent or outcome variable in a population. These method were utilized in determining the answers to the problem statements stated. Specifically, several sample CAPTCHA images were generated and tested to determine the efficiency and accuracy of the program using the methods in identifying characters in a CAPTCHA image.

Data Sets/ Subjects of the Study

In the study of code breaking, the researcher defines the different data which used as subjects. Base from the problem statement, the variables that were utilized in this study compares the result of the test on how efficient and accurate is the program using the two methods (a) parts-based approach and (b) segmentation in identifying sets of CAPTCHA images. The results will show the mean of accuracy and efficiency of the algorithms. The CAPTCHA images presented and used in this study refers to the different ez-gimpy CAPTCHA images (composed of one word or line of alphanumeric character) that are used on online websites. Figure 3 shows the sample CAPTCHA image from pligg and word press multi-user website.

Figure 3. Sample of Pligg and WPMU website CAPTCHA.

Measures/ Sample Size

Since the transition of generated CAPTCHA images have almost the same environment in terms of background, character position, and structure, Table 1 presents the summary of number of samples.

Table 1. Number of Samples.

METHOD	Number of Samples	Image File Size Range in Bytes
PARTS-BASED APPROACH	10	2100-2199
	10	2200-2299
	10	2300-2399
	10	2400-2499
	20	3300-3500
SEGMENTATION APPROACH	10	2100-2199
	10	2200-2299
	10	2300-2399
	10	2400-2499
	20	3300-3500
Total	**120**	

In general, CAPTCHA images that will be tested in the experiment are composed of 120 images. 80 images from Pligg website and 40 images from WordPress Multi-User website. It is decomposed

into 10 samples from each file size range in Pligg CAPTCHA images, while 20 images from WPMU.

Sampling Techniques

Since that this study is focused on determining the efficiency and accuracy of images matching and breaking of characters, the researcher used the purposive sampling technique. Purposive or judgmental sampling is a strategy in which particular settings persons or events are selected deliberately in order to provide important information that cannot be obtained from other choices (Maxwell, 1996).

Data Gathering Instrument

As mentioned earlier in data sets, the instruments that were utilized in this study are the CAPTCHA images generated from online sources. It was the bases for comparing the methods or approaches in automated character recognition and these images were utilized in analyzing the results.

Data Gathering Procedures

Data sets were gathered through online server and online websites were CAPTCHA images was stored and used.

First, the researcher had selected two approaches, the Parts-Based Approach and Segmentation which used in developing an Optical Character Recognition Tool in breaking CAPTCHA images.

Second, CAPTCHA images were tested and generated in individual manner using the tool and the approaches: (a) parts-based, and (b) segmentation. The two approaches used one method which combines all the character that has been identified, the holistic method. Online generation of result of how many characters have been identified and the time spent in seconds is done through manual execution or encoding of code to crack each image. To get the original CAPTCHA image, the researcher has provided the URL which connects to Pligg and WPMU websites and manually saves the image to designated file size sample.

Lastly, the data sets that have been generated are separated by its file size and the method used. Then, accuracy and efficiency were computed and serves as a basis for experimentation in gathering the correct output.

Methods on Data Analysis

Statistical treatment of data is an important aspect of all experimentation and a thorough understanding is necessary to conduct the right experiment with the right inferences from the data obtained. For the first and second problem which sought to determine the efficiency and accuracy of the program in identifying characters of an image code using Parts-Based and Segmentation Approach, the weighted arithmetic mean was used. Mean of Accuracy was computed by getting the total number of identified characters divided by the total number of characters tested. Therefore, in each image file size range can have different result which can be compared and use as basis in getting reliable accuracy and efficiency result in each approach.

For the third problem which sought to determine if there is a significant difference in the accuracy and efficiency of the program in identifying characters of an image code using Parts-Based Approach and Segmentation, and assuming that the distribution was normal, the one-way Analysis of Variance (ANOVA) was used, respectively. The significance level of the tests was set to 0.05 levels. ANOVA is a statistical tool that compares two or more instances that occurs in an environment. In this study, this tool was used to compare the importance of the approaches and to determine which of the two is much effective to be used in breaking codes.

Results and Discussion

This part presents findings of the study through the use of statistical tools in the treatment of the experimental and inferential data. The discussions with regards to the respective results are also included.

Results

The generated CAPTCHA image from online website were utilized to determine the accuracy and efficiency of identifying characters using two approaches such as parts-based approach and segmentation, and with the use of holistic approach in both methods at the end of the process. Table 2 shows the results of accuracy and efficiency of parts-based and segmentation approach.

METHOD	Image File Size Range	Accuracy (Percentage)	Efficiency (measured in seconds)
Parts-Based Approach	2100-2199	85.00	2.286
	2200-2299	91.67	1.504
	2300-2399	90.00	2.715
	2400-2499	95.00	6.827
	3300-3500	50.00	7.545
Total Average		**90.42**	**3.333**

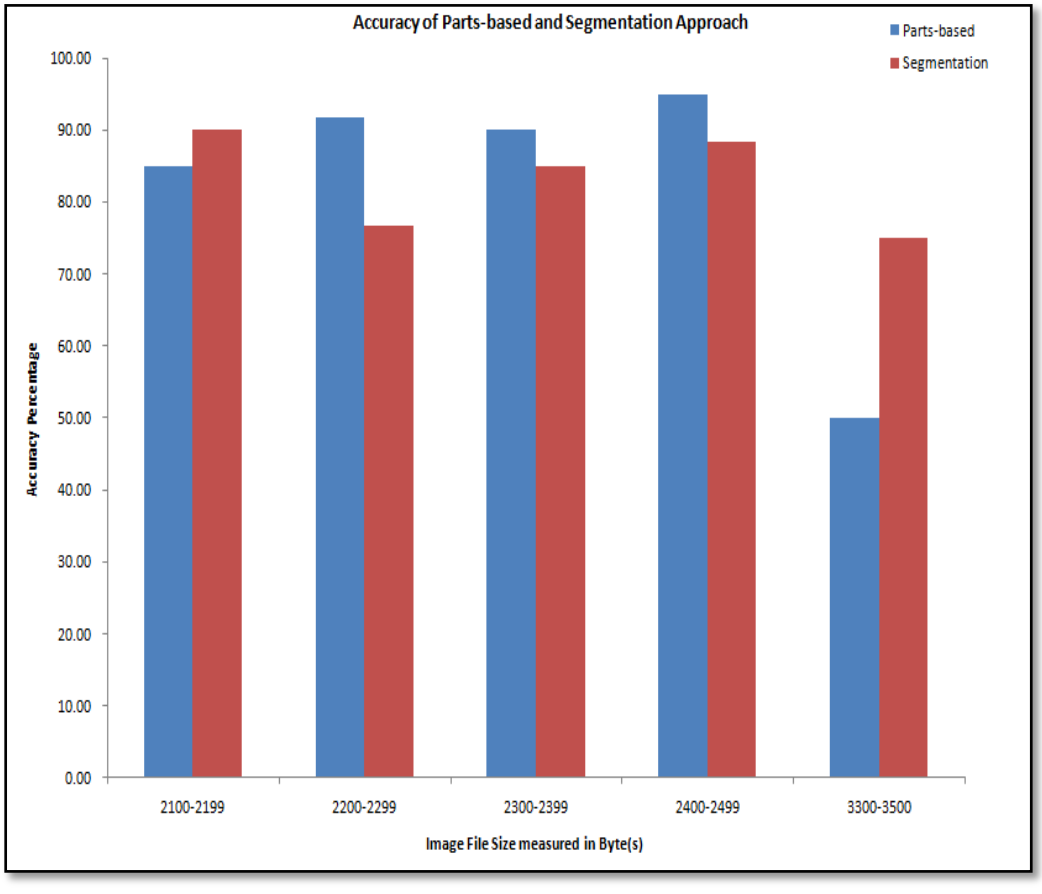
Segmentation Approach	2100-2199	90.00	0.126
	2200-2299	76.67	0.136
	2300-2399	85.00	0.136
	2400-2499	88.33	0.134
	3300-3500	75.00	1.599
Total Average		**83.00**	**0.426**

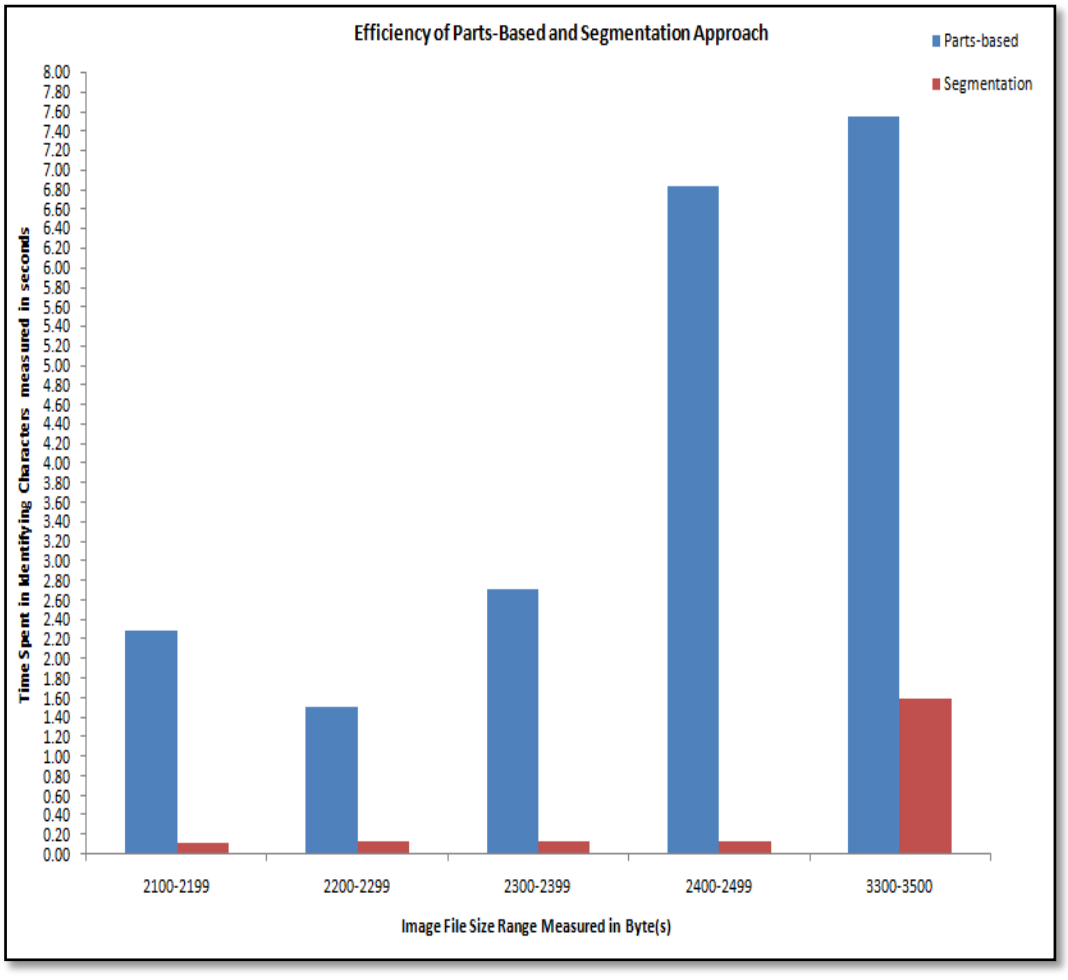
Table 2. Tests Result (Parts-Based and Segmentation Approach).

In testing for the efficiency of parts-based approach in identifying CAPTCHA images using the image file size range of 2,100 to 2,199 bytes, the result is 2.863 seconds; in 2200 to 2299 bytes, the result is 1.504 seconds; in 2300 to 2399 bytes, the result is 2.715 seconds; in 2400 to 2499 bytes, the result is 6.827 seconds;and in 3300 to 3500 bytes, the result is 7.545 seconds. On the other hand, the efficiency of segmentation approach in identifying CAPTCHA images using the image file size range of 2,100 to 2,199 bytes, the result is 0.126 seconds; in 2200 to 2299 bytes, the result is 0.136 seconds; in 2300 to 2399 bytes, the result is 0.136 seconds; in 2400 to 2499 bytes, the result is 0.134 seconds; and in 3300 to 3500 bytes, the result is 1.599 seconds. In general, the parts-based approach obtains the average mean of 3.333 seconds, while segmentation approach obtains the average mean of 0.426 seconds in its efficiency. The comparative graph of the results are shown in APPENDIX T.

In testing for the accuracy of the program using parts-based and segmentation approach in identifying characters, ten generated CAPTCHA images were tested. In parts-based approach, using the image file size range of 2,100 to 2,199 bytes, the accuracy result is 85% out of 100%; in 2200 to 2299 bytes, the accuracy is 91.67% out of 100%; in 2300 to 2399 bytes, the accuracy is 90% out of 100%; in 2400 to 2499 bytes, the accuracy is 95% out of 100; and in 3300 to 3500 bytes, the accuracy is 50% out of 100%. On the other hand, the segmentation result using the image file size range of 2,100 to 2,199 bytes is 90% out of 100%; in 2200 to 2299 bytes, the accuracy is 76.67% out of 100%; in 2300 to 2399 bytes, the accuracy is 85% out of 100%; in 2400 to 2499 bytes, the accuracy is 88.33% out of 100; and in 3300 to 3500 bytes, the accuracy is 75% out of 100%. In general, the parts-based approach obtains the accuracy average of 90.42% while segmentation approach obtains the accuracy average of 83%. The comparative graph of the results are shown below.

Comparative Chart in Accuracy Result





Comparative Chart in Efficiency Result.

Table 3 presents the parts-based and segmentation approach difference and it shows that parts-based is more accurate than the segmentation approach with the difference of 7.42%. Thus, in testing for the efficiency of the program, segmentation got the least time in identifying characters with the difference of 2.907 seconds which shows that segmentation approach is more efficient than parts-based. It shows that both methods can be used as a tool in developing an efficient and reliable system.

Method	Parts-Based	Segmentation	Difference
Accuracy	90.42	83.00	7.42
Efficiency	3.333	0.426	2.907

Table 3. Comparison of Average Results.

Hypothesis Test Result

A one-way Analysis of Variance (ANOVA) was employed to compare the significant difference of the accuracy and efficiency of the program using parts-based and segmentation in identifying characters of an image code called CAPTCHA. The result come up with no significant difference in the accuracy and efficiency of the program in identifying characters of an image code using parts-based approach and segmentation. In testing for the accuracy of the program resulted to $p>.05$ level for the parts-based and segmentation approach [$F= 2.841176$, $p = 0.141187$]; and the efficiency of the program resulted to $p>.05$ level for the parts-based and segmentation approach [$F= 0.573563$, $p = 0.694917$] and the means' are the same. The result implies that two methods used in identifying CAPTCHA images do not have a significant

difference in terms of its accuracy and efficiency in recognizing characters.

Discussion

The problem of identifying characters in an image code using different method provides valuable insight into the study of Code Breaking Approach in Automated Character Recognition however, in this study the file size of an image serves as a factor in determining the relationship of two methods which resulted in the acceptance of the null hypothesis of the study. To optimize system performance, recognizers have been used such as stroke and word segmentation which precedent to recognition depends primarily on the geometric relationship between components. The commonly adopted computational approach to solve such issues is to extract the distance between pairs of adjacent components and to generate an array of primitives. Parts-based used the lexical information to decide which characters can be formed into words. In comparing the distinction of two methods, the researcher used another method called holistic method after using the two methods to combine identified characters. Extraction of holistic features from a word image and match the features directly against the entries of a lexicon. The process of constructing a holistic feature representation for all lexicon entries are accomplished before matching to the extracted features. Thus, the study generally accepted for testing the efficiency and accuracy of identifying CAPTCHA characters.

Breaking CAPTCHA by Mori and Malik (2003) studied that efficient methods based on shape context matching can identify the word with an EZGimpy image with a success rate of 92%, and the requisite 3 words in a Gimpy image 33% of the time which used different methods and factors in reading the characters. While in this study, the major sources of variation in identifying characters are not in file size factor only. The observed factors that affect in segmentation and parts-based accuracy and efficiency recognition performance are its different characteristics such as image and character color, size (height), slant, skew (slope), stroke width, font and rotation of characters in a CAPTCHA image.

Conclusions

This study which set out to determine the efficiency and accuracy of the program in identifying characters of an image code using parts-based and segmentation approach which shows that segmentation is more efficient than parts-based but on the other hand, parts-based is more accurate than segmentation approach. One of the more significant findings to emerge from this study is that there is no significant difference in the accuracy and efficiency of the program in identifying characters of an image code using parts-based and segmentation approach. Taken together, these findings suggest that both methods can be combined to produce more accurate and efficient algorithm in identifying characters using optical character recognition and both have its own important contribution.

This study confirms previous findings and contributes additional evidence that the file size of the image does not affect the differences of the accuracy and efficiency of the program in identifying the characters. Thus, the CAPTCHA images are identified to be with the same characteristics in terms of color combination and structure that is why, the accuracy and efficiency of the program using parts-based and segmentation approach needs more factor that can differentiate its significance in character recognition in internet technology.

The results of this paper conclude that the system is reliable in testing the accuracy and efficiency using two approaches in identifying characters of an image code for the reason of having high result for the accuracy and efficiency in testing the data sets.

Recommendations

This study has thrown up many questions in need of further investigation and understanding of the readers. In view of the foregoing conclusions, the researcher formulated the following recommendations to help other researchers develop a study and produce a new efficient and accurate algorithm.

First, that the actual implementation of the Code Breaking Approach in Automated Character Recognition shall be recommended to be used in improving internet security in free email services and social networking websites.

Second, that other factors such as image and character color, size(height), slant, skew(slope), stroke width, font and rotation will be considered in further studies of future researchers.

Lastly, that the future researchers will use this study as a basis in improving internet services and not to be used in hacking and phishing of personal information of other people that could devastate the society.

Bibliography

1. Pham, D. L., & Prince, J. L. (1999). An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity

inhomogeneities. Pattern recognition letters, 20(1), 57-68.
2. Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003, May). CAPTCHA: Using hard AI problems for security. In International Conference on the Theory and Applications of Cryptographic Techniques (pp. 294-311). Springer, Berlin, Heidelberg.
3. Mori, G., & Malik, J. (2003, June). Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. (Vol. 1, pp. I-I). IEEE.
4. Bursztein, E., Martin, M., & Mitchell, J. (2011, October). Text-based CAPTCHA strengths and weaknesses. In Proceedings of the 18th ACM conference on Computer and communications security (pp. 125-138). ACM.
5. Banday, M. T., & Shah, N. A. (2011). A study of captchas for securing web services. arXiv preprint arXiv:1112.5605.
6. Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical character recognition. International journal of recent technology and engineering (IJRTE), 2(1), 72-75.
7. Todorov, A. (2014). U.S. Patent No. 8,640,212. Washington, DC: U.S. Patent and Trademark Office.
8. Mauldin, M. L. (1994, August). Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In AAAI (Vol. 94, pp. 16-21).
9. White, J. M., & Rohrer, G. D. (1983). Image thresholding for optical character recognition and other applications requiring character image extraction. IBM Journal of research and development, 27(4), 400-411.