JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Recognition Of Captcha Characters Using Machine Learning Algorithms

[1]Riya
Mtech Scholar
Sri Sai college of Engineering and Technology

[2]Sorab Kumar
Assistant professor
Sri Sai college of Engineering and Technology

ABSTRACT

CAPTCHA provides way to secure the authentication process by differentiating humans from bots. CAPTCHA includes complex images and identifying characters from the image could be difficult task. This paper presents techniques including KNN, SVM and neural network-based approach commonly used for character recognition. Literature survey presents the detail of models used for character recognition. The results and limitations of KNN, SVM and neural network-based approach were highlighted to determine best possible approach out of these approaches. The results obtained from KNN, SVM and neural network-based approach are compared and SVM based approach yield best possible approach in terms of classification accuracy. The classification accuracy obtained from KNN, SVM and neural network-based approach was 85%, 92% and 90%. The compared techniques are best among character recognition mechanism, but further modifications cold be done in terms of pre-processing and feature extraction. In future, hybridization of pre-processing, optimized feature extraction along with SVM can be used to enhance the process of character recognition.

Keywords: CAPTCHA, KNN, SVM , CNN

1. Introduction

CAPTCHA provides security through authentication process within web-based applications. Multimedia security mechanism CAPTCHA also described as Human Interactive Proofs can be used to ensure multimedia privacy. The field of usage of CAPTCHA is vast. Now CAPTCHA has been accommodated by Google, Yahoo, and other prominent websites. To verify the validity of CAPTCHA, different mechanisms play a part including image processing, character recognition, Artificial based mechanisms along with many distinct disciplines.

The research on CAPTCHA has great applications that includes development of CAPTCHA, human and machine interaction using artificial intelligence and it serves as important prerequisite to actual human machine interaction. With the growth in technology, [] attack on CAPTCHA becomes profound. [1] discussed ideas to make CAPTCHA more usable and robust. Thus, CAPTCHA breaking technologies are required to be researched to make CAPCTHA more secured and robust. The methods associated with different CAPTCHA's is given in table 1.

This paper discussed some of the commonly used method for character recognition within CAPTCHA including KNN, SVM and CNN. KNN is based upon the clustering. This mechanism group together the relevant features based on Euclidean distance. SVM is strictly based upon the concept of hyperplanes. Hyperplanes are labelled with the results of the characters. In [2] CNN is evaluated corresponding to CAPTCHA attacks. CNN is convolution neural network and is a layered approach. Operations such as acquisition, pre-processing, segmentation, and classification are performed by different layers.

The remainder of this paper is organized as under section 2 provides the literature survey of the techniques used for character recognition within CAPTCHA, section 3 provides the limitations of the discussed techniques in section 2, section 4 gives the discussion and section 5 gives the conclusion and future scope.

2. Literature Survey

Demonstration	Used By	Rate of Success	Reference	Method used
	Yahoo Rediff	66%	[1]	CNN
	Google Rediff Yahoo	61%	[2]	CNN
	Hotmail	40%	[3]	SVM
	Yahoo and MSN	45%	[4]	Projection and KNN
	UploadMega	79%	[5]	CNN
	Microsoft	77%	[6]	KNN

Table 1: CAPTCHA and methods for character recognition[7]

In September 2000, the Carnegie Mellon University, designed a commercial CAPTCHA to resist the malicious advertisement. In 2002, research on CAPTCHA text recognition begins. The research with different mechanisms yields different set of accuracies. The KNN based approach for character recognition discussed in [8] provide effective mechanism to identify patterns. The KNN based approach used in [1] identify characters by matching features based on Euclidean distance. the handwritten text can also be identified using this approach. The main problem with this approach is high cost and low accuracy. KNN based approach for character recognition discussed in [9]. This paper discussed an effective handwritten character recognition-based approach for Malayalam characters using KNN mechanism. Time domain and dynamic feature extraction mechanism employed in this literature extract accurate features. Excellent classification accuracy of 98% was achieved using this literature. Support vector machine provides excellent way to identify the characters from the CAPTCHA. Character recognition using SVM was discussed in [10]. SVM and machine learning based approaches were employed to design a model for Optical character recognition. Layered approach used in SVM reduce complexity of identification process. Layered approach of SVM with hyperplane applied within [11]. Tamil character recognition with SVM was used for effectively identifying these characters. A classification accuracy of 82% was achieved. This classification accuracy is significantly low and can be improved further. The Neural network-based approach can be very effective in increasing the classification accuracy of CAPTCHA detection. This approach was discussed in [12]. Layered approach followed in deep neural network divides the entire process of character recognition into phases. In the first phases noise from the CAPTCHA image was removed. After removing the noise, feature extraction was applied using second layer. The processing layer will extract the feature and select the effective features. Optimized iterative based approach applied at the last layer for classification. The CNN based approach provide great classification accuracy of over 90%.

3. Merits and Demerits of KNN, SVM and CNN for CAPTCHA breaking

The techniques like KNN, SVM and CNN offers great mechanisms for decoding CAPTCHA characters however there exists room for improvement. The merit and demerit of these mechanisms are given in table 2

Technique	CAPTCHA	Parameters	Merits	Demerits
KNN[13]		Classification Accuracy Error Rate	Characters were recognized from almost every CAPTCHA	Classification accuracy with different CPATHCA's is not up to the mark
Neural Network based Approach[14]		Recognition Rate	Supervised learning mechanism	Only particular type of

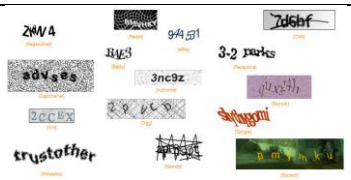
			presents better recognition rate	CAPTCHA having specific patterns can be selected
CNN and SVM[15]		Recognition Rate	Improved rate of recognition	High cost in terms of effort in training and testing mechanism
Other mechanisms for character recognition[12], [16][17], [18]		Error Rate Recognition Rate	Rate of recognition is improved using CNN, KNN, SVM, Random forest etc.	Cost and complexity of detection process is poor and there is a room for improvement

Table 2: Comparison of mechanism used for CAPTCHA text breaking.

4. Discussion

The mechanisms used for the detection of CAPTCHA characters vary from simple to complex. Generally, layered based approach within Neural network is better. In addition, binarization is missing within discussed mechanisms. The binarization is compulsory phase in case features are to be extracted accurately. Pre-processing mechanism can also enhance the classification accuracy. In the future, order or sequence of operations suggested for improving CAPTCHA character recognition includes data acquisition, pre-processing, segmentation, and classification phase. The accuracy and recognition rate can be improved greatly using suggested mechanism.

5. Conclusion and future scope

This paper presents CAPTCHA character breaking mechanisms used for identifying the characters from complex images. The feature extraction from the image using KNN, SVM and CNN does not involve binarization and hence accuracy in feature extraction is missing. The segmentation mechanism could include region of interest mechanism to select only critical features. This will reduce the complexity of operation. The security using CAPTCHA could be high in case breach in character recognition is identified accurately. CAPTCHA character recognition using CNN is observed to the best. In future, binarization with region of interest identification can be used to improve the classification accuracy.

6. References

[1] J. Y. and A. S. E. Ahmad, "A low-cost attack on a microsoft CAPTCHA," *CCS'08*, pp. 543–554, 2008.

[2] F. J.-B. and R. Paucher, "The Captchacker Project," *citereex*, 2009.

[3] and Y. A. R. A. Nachar, E. Inaty, P. J. Bonnin, "Breaking down Captcha using edge corners and fuzzy logic segmentation/recognition technique," *Secur. Commun. Networks*, vol. 8, no. 18, pp. 3995–4012, 2015.

[4] S. Huang, Y. Lee, G. Bell, and Z. Ou, "A Projection-based Segmentation Algorithm for Breaking MSN and YAHOO CAPTCHAs," *ResearchGate*, vol. 2170, no. 1, pp. 727–730, 2008.

[5] B. Madar, G. Kiran, and C. Ramakrishna, "Captcha Breaking using Segmentation and Morphological Operations," *Int. J. Comput. Appl.*, vol. 166, no. 4, pp. 34–38, 2017, doi: 10.5120/ijca2017914013.

[6] N. Yu and K. Darling, "A low-cost approach to crack python CAPTCHAs using AI-based chosen-plaintext attack," *Appl. Sci.*, vol. 9, no. 10, 2019, doi: 10.3390/app9102010.

[7] J. Chen, X. Luo, Y. Guo, Y. Zhang, and D. Gong, "A Survey on Breaking Technique of Text-Based CAPTCHA," *Secur. Commun. Networks*, vol. 2017, no. September 2000, 2017, doi: 10.1155/2017/6898617.

[8] T. K. Hazra, D. P. Singh, and N. Daga, "Optical character recognition using KNN on custom image dataset," *2017 8th Ind. Autom. Electromechanical Eng. Conf. IEMECON 2017*, pp. 110–114, 2017, doi: 10.1109/IEMECON.2017.8079572.

[9] M. Sreeraj and S. M. Idicula, "K-NN based on-line handwritten character recognition system," in

Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010, 2010, pp. 171–176, doi: 10.1109/ICIIC.2010.58.

[10] S. Sharma, A. Sasi, and A. N. Cheeran, "A SVM based character recognition system," in *RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, Jul. 2017, vol. 2018-January, pp. 1703–1707, doi: 10.1109/RTEICT.2017.8256890.

[11] N. Shanthi and K. Duraiswamy, "A novel SVM-based handwritten Tamil character recognition system," *Pattern Anal. Appl.*, vol. 13, no. 2, pp. 173–180, May 2010, doi: 10.1007/s10044-009-0147-0.

[12] A. Thobhani, M. Gao, A. Hawbani, S. T. M. Ali, and A. Abdussalam, "CAPTCHA recognition using deep learning with attached binary images," *Electron.*, vol. 9, no. 9, pp. 1–19, 2020, doi: 10.3390/electronics9091522.

[13] E. Bursztein, J. Aigrain, A. Moscicki, and J. C. Mitchell, "The end is nigh: Generic solving of text-based CAPTCHAs," *8th USENIX Work. Offensive Technol. WOOT 2014*, 2014.

[14] J. OndrejBostik, "Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms - ScienceDirect," *Science Direct*, 2018. https://www.sciencedirect.com/science/article/pii/S2405896318309017 (accessed Apr. 28, 2021).

[15] S. Sachdev, "Breaking CAPTCHA characters using Multi-task Learning CNN and SVM," Feb. 2020, doi: 10.1109/CINE48825.2020.234400.

[16] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-based character recognition via web security measures," *Science (80-.).*, vol. 321, no. 5895, pp. 1465–1468, Sep. 2008, doi: 10.1126/science.1160379.

[17] R. Hussain, H. Gao, R. A. Shaikh, and S. P. Soomro, "Recognition based segmentation of connected characters in text based CAPTCHAs," in *Proceedings of 2016 8th IEEE International Conference on Communication Software and Networks, ICCSN 2016*, Oct. 2016, pp. 673–676, doi: 10.1109/ICCSN.2016.7586608.

[18] H. G. and I. K. R. Hussain, K. Kumar, "Recognition of merged characters in text based CAPTCHAs," *IEEE*, 2016.