Recognition of CAPTCHA Characters Using Machine Learning Algorithms

Dipika Malhotra, Satinder Kaur

Guru Nanak Dev University, Amritsar, India

Corresponding author: Dipika Malhotra, Email: dipika10061998@gmail.com

The CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is an authentication test used to distinguish humans from bots in a variety of online applications. Recognition of characters is incorporated with the help of the CAPTCHA method to make web applications safe and trust worthy. CAPTCHA contains some complicated pictures in some cases and owing to noisy values, it might be difficult to distinguish the characters from these images. Several researchers have sought to solve this problem using machine-learning techniques. As a result, the focus of this work is on a comparison of classification algorithms such as k-NN, SVM, and CNN for recognizing CAPTCHA characters in the literature. Following a thorough analysis of previous research, it has been established that CNN, rather than k-NN or SVM, is the most accurate classification approach. In the future, CNN might be used to improve the process of character recognition.

Keywords: CAPTCHA, k-NN, SVM, CNN.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 101–106. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-10

1 Introduction

CAPTCHA ensures that the authentication procedure in web-based applications is secure. In recent research, the multimedia security mechanisms of CAPTCHAs have been characterized as Human Interactive Proofs (HIPs), which may be utilized to protectmultimedia privacy [1]. CAPTCHA has a wide range of applications. CAPTCHA authenticity is now verified using Google, Yahoo, and other well-known websites. Image processing, character recognition, AI-based approaches, and a range of disciplines all play a part [2]. CAPTCHA preserves you from spam and password decryption by forcing you to complete a simple test to prove that you are a human, not a computer, accessing a password-protected account. A CAPTCHA exam consists of two parts: a distorted picture made comprised of a randomly generated sequence of characters and/or numbers, and a textbox. To pass the test and validate your human identity, simply type the characters you see in the image into the text space. Text-based CAPTCHA, image-based CAPTCHA, audio-based CAPTCHA, and sound-based CAPTCHA are all examples of CAPTCHA. These CAPTCHAs are commonly used to safeguard user passwords and data on websites and mobile applications. Machine learning-based human-machine interaction includes a variety of strategies to aid human recognition in the task. There have been a lot of studies done on the mechanisms that have led to the creation of CAPTCHA. However, Azad [3]'s attack against CAPTCHA gets more serious as technology advances. Thobhani et al.[4] suggested ways to improve CAPTCHAs usability and reliability. The k-nearest neighbour (k-NN), Supportvector machine (SVM), and Convolutional neural network have all been the subject ofcurrent research in this field (CNN). Clustering is the foundation of the first approach, k-nearest neighbour (k-NN). On the notion of hyperplanes, the second approach SVM is utilized [5].CNN, on the other hand, employed a multi-tiered approach [6]. The goal of this research is to look into some of the most often used machine-learning algorithms for CAPTCHA character recognition. The rest of the paper is structured as follows: Section 2 is a review of the research onstrategies for recognizing CAPTCHA letters from complicated photos. The gaps in theliterature are presented in Section 3. Section 4 discusses the benefits and drawbacks of k-NN, SVM, and CNN and the comparative analysis. Section 5 leads to discussions and future recommendations. The paper draws to a close in Section 6.

2 Literature Review

This section examines CAPTCHA-breaking technologies. Table 1 lists the procedures connected with various CAPTCHAs.

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

Demonstration	UsedBy	Rate of Success	References	Methods used
	Yahoo, Rediff	66%	Huang et al.[1]	CNN
	Google, Rediff, Yahoo	61%	Bostik[2]	CNN
	Hotmail	40%	Azad[3]	SVM

	Yahoo and MSN	45%	Thobhani etal.[4]	Projectionandk-NN
	UploadMega	79%	Shu-Guangetal.[5]	CNN
	Microsoft	77%	Wang[6]	k-NN

In September 2000, Carnegie Mellon University created a commercial CAPTCHA to resist fraudulent ads using the k-NN technique [7]. In 2002, researchers conducted research on CAPTCHA text recognition using several techniques that yielded varying degrees of accuracy [8]. Several researchers have proposed aneffective approach for identifying patterns by using k-NN [8-9]. Moreover, the k-NN based technique is employed to identify the characters by matching features based on the Euclidean distance [1]. In addition, using this method on Malayalam characters, hand written writing may be detected [9]. The primary drawbacks of this strategy are its high cost and lack precision [10]. One research used a time-domain and dynamic feature extraction technique with an 85% accuracy [11]. The support vector machine (SVM) is an excellent method for identifying CAPTCHA characters. SVM has been used to recognize characters in various research articles [10,12]. Furthermore, a model for optical character recognition (OCR) is designed using machine-learning methodologies [13]. Additionally, SVMs layered method (CNN) minimizes the complexity of the identification procedure [11]. Character recognition in Tamil is effectively detected using SVM with upto 82% classification accuracy [14]. However, the accuracy of this categorization is low and might be improved.The usage of a Convolutional Neural Network (CNN) based technique to improve CAPTCHA detection classification accuracy [12]. For the identification of CAPTCHA characters, however, the layered technique (CNN) used in deep neural networks separates the entire process into parts [15]. The noise from the CAPTCHA picture will be eliminated in the first step. The second layer will be used to extract features once the noise has been removed. The processing layer will extract the feature and choose the effective features from the picture in the second phase. Furthermore, the CNN-based technique has a classification accuracy of above 90% [16]. Binarization, despite this, is a necessary step for accurate feature extraction [17].

This literature is examined to discover the most effective CAPTCHA breaking strategy. All CAPTCHA-breaking algorithms, however, lack binarization, resulting in classification accuracy issues.

3 Gaps in Literature

- The execution time increases when the picture is huge.
- NoisecanhampertheperformanceoftheSVM.
- The accuracy of classification in practically all approaches may be increased.
- Thebinarizationphasemaycausethelossofimportantfeatures.

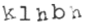
4 Probable Solutions

The issue arises as the training can be accomplished with the high entropy images only. This means complex images are rejected from the training process. The training process thus may not be complete and result in poor classification accuracy. To overcome the issue focal loss function can be accommodated within the training process. By raising the entropy of the complicated pictures, this function is capable of allowing training from them. Moreover, in this manner, the training consequent to the picture and images will be complete. As a result, categorization precision will be great.

5 Comparative Analysis of KNN, SVM & CNN

Machine-learning algorithms are an excellent alternative for understanding CAPTCHA characters. Each approach, however, has advantages and disadvantages, which are listed inTable 2.

Table2: The techniques used to crack CAPTCHA text are compared

Techniques	CAPTCHA	Parameters	Merits	Demerits
KNN [13]		Classification accuracy, Error Rate	Characters were recognized from almost every CAPTCHA	Classification accuracy with different CAPTCHAs is not upto the mark.
Neural Network-based Approach [14]		Recognition Rate	Supervised learning mechanism presents better recognition rate	The only Particular type of 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]		ErrorRate, RecognitionRate	The rate of Recognitio n is improved using CNN, KNN, SVM, Random forest, etc.	The cost and Complexity of the detection process is poor and there is an area for improvement.

6 Discussions & Future Recommendations

CAPTCHA characters are detected using a variety of techniques, ranging from basic to complicate. A layered approach, such as a convolutional neural network (CNN), is favoured in general.Table 2 shows the machine learning methods that were utilized to assess the CAPTCHA images. Significant error rates, on the other hand, are produced by a pattern's failure to identify characters. Moreover, a preprocessing approach can improve classification accuracy. In the future, it's a good idea to think about the order or sequence of processes for better CAPTCHA character recognition, such as data collecting, pre-processing, segmentation, and classification. Besides, employing recommended procedures can boost accuracy and recognition rate.

7 Conclusion

CAPTCHA is a popular multimedia security technique used in web-based applications. It is a typical way to tell the difference between people and robots. The CAPTCHA character cracking algorithms presented in this work are utilized to recognize characters from complicated pictures. Binarization implementation has been shown to be insufficient in earlier studies, resulting in less trustworthy feature extraction results. A segmentation approach is also included to choose only the most significant attributes, decreasing the operation's complexity. Furthermore, if the CAPTCHA security level is high, accurate character recognition may be possible. CNN is also the most successful in recognizing CAPTCHA characters. In the future, binarization may be researched to increase classification accuracy.

References

[1] Huang, S. Y. et al. (2008). A projection-based segmentation algorithm for breaking MSN and YAHOO CAPTCHAs. In *Proceedings of the World Congress on Engineering*.

[2] Bostik, O., and Klecka, J. (2018). Recognition of CAPTCHA characters by supervised machine learning algorithms. *IFAC-Papers online*, *51*(6): 208-213.

[3] Azad, S., and Jain, K. (2013). Captcha: Attacks and weaknesses against OCR technology. Global Journal of Computer Science and Technology.

[4] Thobhani, A. et al. (2020). CAPTCHA Recognition Using Deep Learning with Attached Binary Images. Electronics, 9(9): 1522.

[5] Shu-Guang, H. et al. (2011). ACAPTCHA recognition algorithm based on holistic verification. In *2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control* (525-528). IEEE.

[6] Chen, J. et al. (2017). A survey on breaking technique of text-based CAPTCHA. *Security and Communication Networks*, *2017*.

[7] Sreeraj, M., and Idicula, S.M.(2010). k-NN based On-Line Hand written Character recognition system. In *2010 First International Conference on Integrated Intelligent Computing* (171-176). IEEE.

[8] Xu, X. X. G. X. C., and Fan, C. (2017). Maximizing Reliability of Energy Constrained Parallel. *Power*, 10(21): 7-16.

[9] Yan, J.(2016).A simple generic attack on text captchas.

[10] Wang, X., and Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, 36(10): 2429-2439.

[11] Sharma, S., Sasi, A., and Cheeran, A. N. (2017). A SVM based character recognition system. In *2017 2nd IEEE International Conference on Recent Trends inElectronics, Information & Communication Technology (RTEICT)* (1703-1707). IEEE.

[12] Azizi, N. et al. (2014). A new hybrid method combining genetic algorithm and support vector machine classifier: Application to Ca AD system for mammogram images. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*(415-420).IEEE.

[13] Yang, J. et al. (2018). A novel multimodal biometrics recognition model based on stacked ELM and CCA

methods. *Symmetry*, *10*(4): 96.

[14] Chakraborty, M., Biswas, S. K., and Purkayastha, B. (2020). Data Mining Using Neural Networks in the form of Classification Rules: A Review. In *2020 4th International Conferenceon Computational Intelligence and Networks(CINE)*(pp. 1-6). IEEE.

[15] Pal,K.K., and Sudeep, K. S. (2016). Preprocessing for image classification by convolutional neural networks. In 2016 IEEE International Conference on Recent Trendsin Electronics, Information & Communication Technology (RTEICT)(1778-1781).IEEE.

[16] Bursztein, E. et al. (2014). The end is nigh: Generic solving of text-based captchas. In 8th {USENIX} Workshop on Offensive Technologies ({WOOT}14).

[17] Ghanbari, M., Kinsner, W., and Ferens, K. (2017). Detecting a distributed denial of service attack using a pre-processed convolutional neural network. In 2017 IEEE Electrical Power and Energy Conference (EPEC)(1-6).IEEE.

[18] Sachdev, S. (2020). Breaking CAPTCHA characters using Multi-task Learning CNN and SVM. In 2020 4th International Conference on Computational Intelligence and Networks (CINE)(1-6). IEEE.

[19] Von Ahn, L., Blum, M., and Langford, J. (2002). Telling humans and computers apart automatically or how lazy cryptographers do AI.