Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 13, Issue. 6, June 2024, pg.22 – 26

OCR Technology for Detecting Homoglyph Spam

Kostiantyn Shcherbyna

Department of Computer Science and Technologies, National Aviation University, Ukraine
konstantinshcherbina02@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2024.v13i06.004

Abstract— This paper examines the effectiveness of Optical Character Recognition (OCR) technology in detecting spam composed of homoglyphs, visually similar characters from different scripts. The research utilizes a console application capable of generating all possible homoglyphic representations for specified words, which are then visually represented and analyzed using the OCR tool Tesseract. The study assesses the recognition accuracy of homoglyphs in different text cases: only uppercase, only lowercase, random case, and only the first letter uppercase. Findings reveal variable accuracy rates across these text formats, with uppercase letters generally showing higher recognition rates. This differential recognition underscores the challenges and potential for refining OCR applications to better detect and filter homoglyph-based spam, enhancing security across digital communication platforms.

Keywords— Homoglyphs, Text recognition, OCR, Spam, Recognition accuracy

I. INTRODUCTION

In the digital age, the proliferation of spam across various communication platforms has posed significant challenges for information security. Particularly, the use of homoglyphs has become a sophisticated method for bypassing conventional spam filters. These characters complicate the process of spam detection, as they can be used to mimic legitimate text to deceive both users and automated systems. This study focuses on leveraging OCR technology to enhance the detection and filtering of homoglyph spam, thereby improving the reliability of digital communication security.

The primary motivation behind this research is the increasing need for more effective tools to combat spam that employs homoglyphs to evade detection. Traditional spam detection systems often fail to recognize these characters correctly, leading to security vulnerabilities and potential data breaches. By exploring the capabilities of OCR technology, particularly in recognizing text composed of homoglyphs across different cases and formats, this paper aims to provide insights into the strengths and limitations of current OCR applications.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

II. METHODOLOGY

This research employs a comprehensive methodology designed to assess the efficacy of OCR technology in identifying and deciphering homoglyphs, which are visually similar characters [1] from different scripts often used in spam messages. The methodology is structured into several key phases: generation of homoglyphs, creation of image representations, and OCR recognition testing.

A. Homoglyph Generation

The initial phase involves generating a complete set of homoglyphs for specified target words commonly found in spam, such as “Free”, “Winner”, and “Discount”. This is accomplished using a console application sourced from an open-access GitHub repository [2]. The program is capable of producing all potential homoglyphic variations of each letter in these words, both in uppercase and lowercase forms. These variations are stored in individual TXT files for each letter, facilitating systematic analysis.

B. Image Representation

Each homoglyph is then converted into a visual format. A custom-developed C# program reads each line from the TXT files containing the homoglyphs and creates corresponding PNG images. The images are generated using the “Segoe UI Symbol” font at a size of 24 points, chosen for its extensive coverage of the Unicode system [1]. This font selection is crucial to minimize the number of characters displayed as rectangles, which represent unrecognized or unsupported glyphs. Modifications were made to the document to manually remove characters that fail to display correctly even at larger image sizes, ensuring that only valid homoglyphs are included in the testing phase [3].

C. OCR Recognition Testing

The core of the methodology lies in the OCR recognition testing, conducted using Tesseract, a popular OCR tool integrated into the .NET framework [4]. Tesseract is equipped with trained data for the English alphabet, which it uses to attempt recognition of the expected letter from each homoglyph image. The recognition process does not discriminate based on the case of the letters, allowing for an unbiased evaluation of recognition accuracy across different text cases (uppercase, lowercase, random case, and initial uppercase).

Each homoglyph image is processed by Tesseract, and the recognition results are compiled and analyzed to determine the percentage of successful identifications for each letter and word. The success rates are then aggregated to evaluate the overall performance of the OCR technology in recognizing homoglyphs used in spam [5].

D. Statistical Analysis

The data collected from the OCR tests are subjected to statistical analysis to identify trends and patterns in recognition success rates across different text formatting conditions. This analysis helps in understanding which homoglyphs are more or less likely to be correctly identified and why, providing insights into potential improvements in OCR technologies for spam detection.

III. RESULTS

A. Total and Removed Homoglyphs

During the homoglyph generation phase, a number of symbols were initially displayed as rectangles due to limitations in font support, necessitating their removal from the analysis (Figure 1). After adjusting for these unsupported symbols, the number of usable homoglyphs for testing reduced (Figure 2) but provided a clearer, more focused dataset for accurate OCR testing.

Fig. 1 Number of homoglyphs removed for each letter

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

Fig. 2 Number of homoglyphs before and after removal of non-displayable characters

The removal of unsupported homoglyphs affected the overall dataset size but ensured that the subsequent OCR testing was performed only on valid, displayable characters, thus enhancing the reliability of the results.

B. Recognition Rates of Individual Letters

The OCR technology tested showed varying levels of success in recognizing individual homoglyphs, as detailed in the results presented in Figure 3. Notably, certain letters such as “B”, “D”, “M”, “N”, “T”, “W”, and “X” exhibited high recognition rates, on average above 90%. These letters are typically well-defined and distinct, which contributes to their higher recognition accuracy. Conversely, letters like “I”, “L”, and “O” demonstrated significantly lower success rates, often below 50%. These characters have simpler shapes that are easily confused with similar glyphs, leading to frequent misrecognitions.

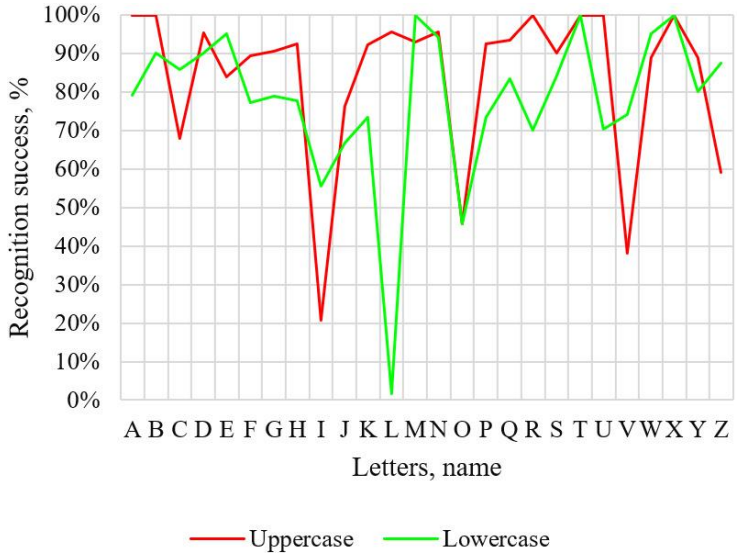

Fig. 3 Recognition success for each letter

C. Effects of Different Text Cases

The experiment also explored how different text cases affect the recognition accuracy. As depicted in Figures 4 and 5, the OCR system performed better when processing words in all-uppercase compared to other cases. The non-strict recognition method (Figure 4) showed that all-uppercase text had the highest average recognition rate of 86.46%, closely followed by random case text at 85.72%. The strict recognition method (Figure 5) confirmed these trends, with all-uppercase also leading in recognition success. These findings suggest that uppercase homoglyphs are generally more distinguishable to OCR systems, potentially due to their larger and more distinct character shapes.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

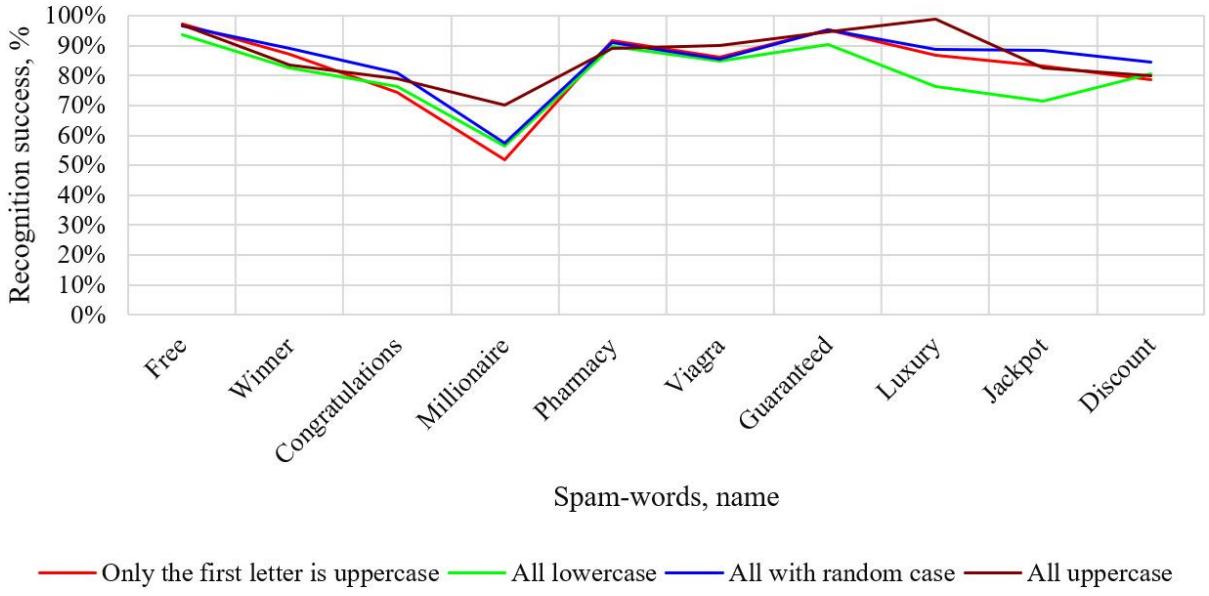

Fig. 4 Success of word recognition using a non-strict method

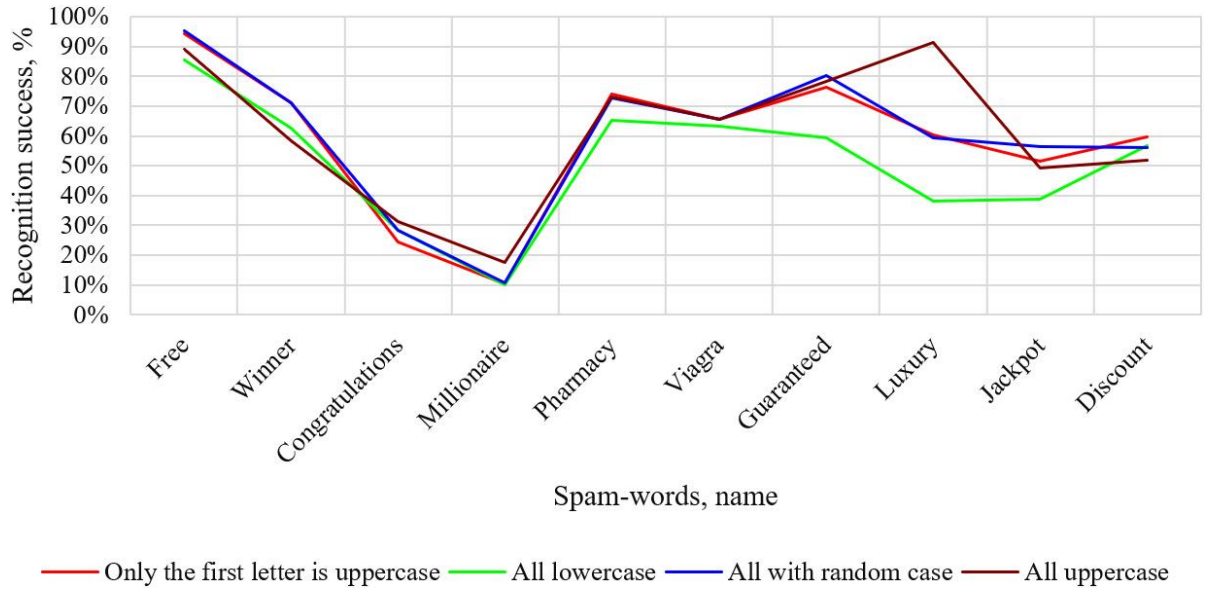

Fig. 5 Success of word recognition using the strict method

D. Word-Level Recognition Performance

When examining whole words composed of homoglyphs, the recognition performance varied significantly depending on the chosen method and text case. Table 1 summarizes these findings, illustrating that the non-strict method yielded better overall results across all text cases. For instance, the non-strict method resulted in an average success rate of 83.92% for words with only the first letter uppercase, and 57.42% for all lowercase. In contrast, the strict method, which requires perfect recognition of every letter in a word, showed lower success rates, highlighting the challenges in achieving flawless recognition of homoglyphs.

TABLE I
Recognition Success under Different Conditions

Non-strict method				Strict method			
Only the first letter is uppercase	All lowercase	All with random case	All uppercase	Only the first letter is uppercase	All lowercase	All with random case	All uppercase
83.25%	80.27%	85.72%	86.46%	58.74%	50.76%	59.56%	60.60%
83.92%				57.42%			
Average				Average			

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

E. Challenges and Limitations

Despite some success, the OCR system still faced challenges, particularly with letters and words that have homoglyphs easily mistaken for other symbols. This issue was exacerbated in cases where homoglyphs did not render correctly due to limitations in the chosen font or the image size, as initially suspected but later disproved through further testing.

IV. CONCLUSIONS

This study has thoroughly investigated the efficacy of OCR technology in recognizing homoglyphs. The research primarily focused on evaluating the OCR's performance across different text cases and formats, including uppercase, lowercase, random case, and initial uppercase scenarios, which were assessed using both strict and non-strict recognition methods.

The results confirmed that uppercase letters are recognized with greater accuracy by OCR systems compared to lowercase or mixed-case formats. This finding suggests that the distinct and larger forms of uppercase letters enhance their recognizability, which could be particularly beneficial for improving spam detection systems focused on homoglyph identification.

Lowercase letters and words with complex character compositions, including random cases, presented significant challenges in terms of recognition accuracy. Certain characters demonstrated particularly low recognition rates, indicating potential areas for technology improvement.

The non-strict recognition method, which averages the recognition rates across characters within words, generally yielded higher success rates. This approach could be advantageous in applications where partial recognition of text is sufficient for effective operation.

The strict method, which requires perfect recognition of all characters in a word, showed lower success rates. This highlights the challenges in achieving comprehensive accuracy in OCR technology when dealing with homoglyphs, especially in spam detection where exact character recognition is crucial.

The insights gained from this study underscore the need for ongoing advancements in OCR technologies, particularly in enhancing their capability to recognize and differentiate between homoglyphs accurately.

In conclusion, while OCR technology has shown promising results in recognizing homoglyphs, particularly in simpler text cases, there remains a substantial scope for improvement to meet the challenges presented by sophisticated spam techniques. Future research should focus on refining OCR algorithms to handle a broader array of character types and complexities, thereby strengthening the security mechanisms against homoglyph-based spam attacks.

REFERENCES

[1]. M. Y. J. A. Azmi and S. C. Navaneethakrishnan, "Lightweight Language Agnostic Data Sanitization Pipeline for Dealing with Homoglyphs in Code-Mixed Languages," in *Speech and Language Technologies for Low-Resource Languages*, Second International Conference, SPELL 2023, Perundurai, Erode, India, December 6-8, 2023, pp. 148-160.

[2]. (2022) Shadawch/homoglyph, "Generate all homoglyphs for a given input sentence". [Online]. Available: https://github.com/shadawck/homoglyph

[3]. L. Struppek, D. Hintersdorf, F. Friedrich, M. Brack, P. Schramowski, K. Kersting, "Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis," *Journal of Artificial Intelligence Research*, vol. 78, pp. 1017-1068, 2023.

[4]. (2024) Tesseract-ocr/tesseract, "Tesseract Open Source OCR Engine (main repository)". [Online]. Available: https://github.com/tesseract-ocr/tesseract

[5]. N. H. Imam, V. G. Vassilakis, D. Kolovos, "OCR post-correction for detecting adversarial text images," *Journal of Information Security and Applications*, vol. 66, issue C, May 2022.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden