Text Extraction of Esports Summary Score Image in the Thai Language Using OCR Technology

Napaphat Wannatrong[1*], Zagon Bussabong[2]

[1,2]*Computer Science Program, Faculty of Science, Buriram Rajabhat University; E-mail: napaphat.wn@bru.ac.th*

Abstracts: The esports industry in Thailand has gained widespread attention, with several organizations starting to organize RoV tournaments to enhance the excitement of the competitions. Consequently, statistics of each match are collected and utilized as data for promoting each round of the competition. Conventionally, the method of data collection involves capturing images of the match results and manually inputting the information for further analysis. However, this process often leads to errors or delays, particularly when dealing with a large volume of data. To address these issues, the researchers explore the use of Optical Character Recognition (OCR) for data extraction from images, aiming to reduce errors associated with manual data entry and improve the convenience and efficiency of data collection. A comparative analysis of image data extraction performance between pytesseract and easyOCR reveals that pytesseract provides superior data extraction results and requires less time for the extraction process.

Keywords: Esports, RoV, Optical Character Recognition, Pytesseract, EasyOCR.

1. INTRODUCTION

Esports is a form of sport where the primary sporting elements are facilitated by electronic systems; the input of players and teams as well as the output of the e-sports system are mediated by human-computer interfaces (Hamari, 2017, p. 211).

The global esports market has been experiencing continuous growth. In 2020, the esports industry generated global revenue of $1.1 billion, representing a 15% growth compared to 2019, when the revenue was $950 million. The industry is showing a consistent upward trend, including an increase in global viewership, which reached 495 million in 2021, up from 443 million in 2020 (an increase of 11.7%). Furthermore, there is a projected upward trend to reach 646 million viewers by 2024. Currently, the largest market for esports is China, followed by the United States, while Thailand ranks among the top 20 globally (Digital Economy Promotion Agency, 2021).

The esports industry in Thailand has gained significant attention and recognition on a broad scale. Numerous companies are increasingly penetrating the esports market, and even the government has started providing greater support to this business sector. This is evident through the establishment of the Thai Esports Association, which has recognized esports as a competitive sport. Additionally, there is growing support for innovation and business development within the esports industry and gaming sector, with a budget allocation of over 400 million baht in the fiscal year 2022. These efforts have contributed to the rapid growth of the esports industry in Thailand (Digital Economy Promotion Agency, 2021). Furthermore, with a total prize of $22,282,988.80 US dollars in 2023, Thailand ranked 16th in the world in terms of esports earnings in 2023 (Esports Earnings, 2023).

Arena of Valor is a mobile-only MOBA esports game. The game was created by TiMi Studio Group, a division of the Chinese company Tencent Games. Honor of Kings, which was also created by Tencent, was its original name (Anunpattana et al., 2018). RoV has gained significant popularity as an esport in Thailand. The Garena World 2019: Unlock Your Passion event held in Thailand in 2022 attracted a massive attendance of up to 269,500 participants (Lerksirinukul, 2019).

Several organizations, including educational institutions, have started organizing RoV tournaments. In order to make the competition more interesting, match statistics are collected for each match and utilized as data for promoting each round of the competition. However, the conventional method of data collection involves capturing

images of the game results and manually entering the information for analysis. This approach sometimes leads to errors or delays, particularly when dealing with a large volume of data. To address these issues, researchers have shown interest in studying the extraction of data from images using Optical Character Recognition (OCR) technology to reduce errors caused by manual data entry and significantly improve the efficiency of data collection.

Based on research studies in OCR, the application of OCR technology for data extraction from esports match result images, particularly in the game Realm of Valor (RoV) in Thailand, has not been found. The objective of this research is to investigate the data extraction from RoV match result images in Thailand using OCR technology. Additionally, the research aims to compare the data extraction performance between Pytesseract and EasyOCR, as both OCR engines share similar characteristics, such as supporting the Thai language and being built-in OCR engines, meaning they come with pre-trained models that are ready for users to utilize. Users can simply adjust the parameters according to their specific needs.

1.1. Related Works

1.1.1. Realm of Valor

According to Garena Thailand (n.d.), RoV stands for Realm of Valor or Arena of Valor, and it is abbreviated as RoV. The game originated from the game "Honor of Kings" or "King of Glory," which is a mobile-based game developed and published by Tencent Games for iOS, Android, and Nintendo Switch platforms. In September 2018, the game generated revenue of $890 million US dollars outside mainland China. As for the Thai server, it was launched on December 26, 2016, marking its first service introduction in Thailand.

The overall gameplay can be summarized as follows: Each player controls one hero, and their objective is to kill creeps, monsters, heroes, or attack towers to gain experience and money. Players can use the money they earn to purchase items, and having more items than their opponents gives them an advantage in combat. When players are strong enough, they can kill enemies and then proceed to attack the opposing base. If they successfully destroy the enemy base, they win the game. After the players finish playing the game, the following information will be displayed to them:

The statistics of the competition results are as follows: 1) Kills: The number of enemy characters killed 2) Deaths: The number of times the player's character was killed by the enemy. 3) Assists: The number of times the player helped teammates on the team. 4) Money: The in-game currency 5) MVP: the player with the highest score, as shown in Figure 1.

Figure 1. Example of Competition Statistics in RoV Game.

In addition, the details of the competition statistics are as follows: 1) Damage dealt 2) Damage dealt as a percentage 3) Damage received 4) Damage received as a percentage 5) Teamfight, and 6) Teamfight as a percentage; an example is illustrated in Figure 2.

Figure 2. Example of Competition Statistics Details in RoV Game.

1.1.2. Optical Character Recognition

The electronic conversion of images of typed, handwritten, or printed text into machine-encoded text is known as optical Character Recognition (OCR). OCR has been utilized in many different industries and for a wide range of purposes, such as data entry from printed records, forms, and papers and the identification of text on road signs, among others. The key benefits of employing OCR technology are that it converts an image into a searchable, editable, and convenient electronic storage format (Malkadi et al., 2020, p. 2).

Optical Character Recognition can be divided broadly into two categories: online and offline. While users enter the characters, the online system can carry out its operation. Additionally, they can aid in capturing the characters' pace of writing. They are hence simpler. The offline system produces complicated pattern recognition because it bases its operation on constant data. Online systems are far more in demand than offline systems because they deliver more precise data, are simpler to create, and may be integrated (Chandra et al., 2020, p. 3038).

1.1.3. PyTesseract

Python-tesseract is an optical character recognition (OCR) tool for Python. That is, it will recognize and “read” the text embedded in images. Python-tesseract is a wrapper for Google’s Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to Tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file (PyPI, 2022).

The optical character recognition (OCR) tool PyTesseract for Python is utilized. This means that it will be able to recognize, read, and extract text in the form of a string from a digital image. The string is then saved, and other actions are carried out on the extracted string (Chadha et al., 2020, p. 33).

1.1.4. EasyOCR

EasyOCR is a ready-to-use OCR tool that supports over 80 languages and all popular writing scripts, including Latin, Chinese, Arabic, Devanagari, Cyrillic, and more (PyPI, 2022). Automatic data entry is done using Easy Optical Character Recognition (EasyOCR). To create data that can be altered on a computer from handwritten, typed, or printed text, simple OCR software is utilized (Kochale et al., 2021, p. 13).

1.1.5. Character Error Rate (CER)

The concept of Levenshtein distance serves as the foundation for CER calculation, which counts the fewest character-level operations necessary to convert the ground truth text (also known as the reference text) into the OCR output. This formula serves as its representation: (Leung, 2021).

$$CER = \frac{S + D + I}{N}$$

S = Number of Substitutions

D = Number of Deletions

I = Number of Insertions

N = Number of characters in reference text (aka ground truth)

The result of this equation is the proportion of characters in the reference text that the OCR output inaccurately predicted. The OCR model performs better the lower the CER value is (with 0 representing a perfect score).

2. LITERATURE REVIEW

The OCR model performs better the lower the CER value is (with 0 representing a perfect score). Character identification was negatively impacted by blurred characters, distinctive number plate designs, and noise from dirt on the plates. These factors significantly impacted plate segmentation (Sawalkar et al., 2022, p. 5862). This paper was reviewed and analyzed different methods for text recognition from images and summarized the steps which are required to detect and recognize the text from any images Scanning the image, Removal of noise, Normalization, Segmentation, Feature Extraction, Classification, and Final text is ready (Saoji, 2021, p. 1674). The proposed License Plate Recognition system using OpenCV & Pytesseract in this paper shows promising and effective results (Chadha et al., 2020, p. 31). Runtime was sacrificed so that the Pytesseract library could excel in the precision metric. If time is not a concern, it would be the best place to go to the library (Ribeiro et al., 2019).

OCR software is applied to the detected plate area to return the license plate number in text format. In order to integrate the transformed number with other IT systems, it is typically stored in a database. The text was extracted from the regions of interest using EasyOCR (Burkpalli et al., 2022, p. 496). This paper uses Easyocr for Character recognition in Automated Vehicle Recognition Identification (AVRI) System. The result showed a letters recognition rate of 86.0% recognition times (ms) of 47.8, a numbers recognition rate of 97%, recognition times (MS) of 50, characters (letters & numbers) recognition rate of 90.7% recognition times (MS) of 53.1. The fundamental difficulty in character identification is dealing with unidentified text arrangement, various font sizes, various lighting situations, reflections, shadowing, and aliasing (Kochale et al., 2021, p. 13).

Extracting patient data using document file conversion techniques has an average processing accuracy rate of approximately 74.62% for attribute extraction and 68.46% for value extraction from documents (Chumwatana et al., 2022, p. 22). This automated pharmaceutical information labeling system demonstrates efficiency in converting image data into text using OCR technology, achieving an accuracy rate of 96.61%, and it can help reduce the time required for medication data collection in databases (Sriborrirux, 2018).

3. METHODOLOGY

3.1. Dataset

The statistical results of the competitions consist of 17 images, along with the details of the RoV tournament's statistics in 2022, totaling 34 images.

3.2. Research Tools

- 1. Python library includes openCV library for handling image data, pandas library for managing data in report format, and re library (regular expression) for formatting string data.

- 2. OCR engines include Pytesseract and EasyOCR, as both OCR engines share similar characteristics, such as supporting the Thai language and being built-in OCR engines, meaning they come with pre-trained models that are ready for users to utilize. Users can simply adjust the parameters according to their specific needs.

3.3. Steps

From the study on data extraction from RoV competition results in images using OCR technology, a comparison was made between Pytesseract and EasyOCR. The research process consisted of the steps shown in the figure.

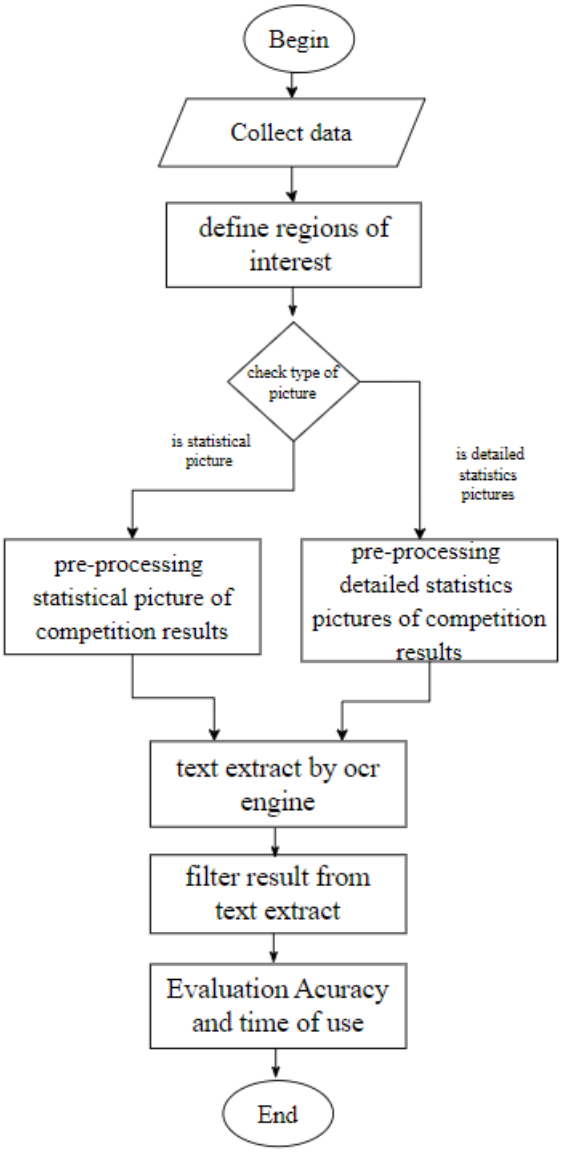

Figure 3. Process of Text Extraction of Esports Summary Score Image

3.3.1. The research process included the following steps:

1. Collecting images of RoV tournament results, totaling 34 images, divided into:

1.1 Images of the statistical results of the matches, including the following data: 1) Player names; 2) Number of kills; 3) Number of deaths; 4) Number of assists; 5) Money; and 6) Most Valuable Player (MVP). There are 17 images in this category.

1.2 Images of the detailed statistical results of the matches, including the following data: 1) Damage dealt; 2) Damage percentage; 3) Damage received; 4) Damage received percentage; 5) Teamfight; and 6) Teamfight percentage. There are 17 images in this category.

2. Determine the desired data positions in the images.

2.1 Convert the images to binary format (3-dimensional array format).

2.2 Rotate the images to a horizontal orientation.

2.3 Identify regions of interest (ROIs) in the images, which are specific points where OCR needs to extract data.

2.4 Obtain the results as ROIs, containing the desired extracted data.

3. Once the ROIs are obtained, the image data is transformed into the desired format before extraction using the OCR engine through a pre-processing step, dividing the color extraction into 2 groups: the group for statistical result images and the group for detailed statistical result images. The process is as follows:

3.1 In the statistical result images group, the pre-processing of the images will be performed as follows:

A. Extract the color from the images, specifically the blue color that highlights the desired text in the images.

B. Extract the color from the images, specifically the red color that highlights the desired text in the images.

C. Combine the results obtained from both filtering rounds of the two images.

D. Convert the image to its complementary color to prepare the image for the color thresholding step.

E. Specify a kernel to be used for image color equalization to prepare for increasing the font size.

F. Increase the font size by reducing the noise. Then, amplify the noise.

G. Perform image denoising to enhance the clarity of the text using a global thresholding technique.

3.2 For the group of detailed statistics images, the filtering process is performed as follows:

A. Extract the color from the images, specifically the blue color that highlights the desired text in the images.

B. Extract the color from the images, specifically the red color that highlights the desired text in the images.

C. Combine the results obtained from both filtering steps.

D. Convert the image to its complementary color to prepare the image for the color thresholding step.

E. Perform image denoising to enhance the clarity of the text, an adaptive thresholding technique is applied, adjusting each pixel individually. This is because the image contains color overlaps with different patterns from the original. The parameters used to calculate the mean are as follows:

1) ADAPTIVE_THRESH_GAUSSIAN_C: to calculate the mean using the Gaussian-weighted method.The mean is computed in the surrounding neighborhood of size blockSize×blockSize, weighted using the Gaussian function, and then subtracted by the C value.

2) blockSize: a pixel neighborhood size variable of the integer type that is used to determine the threshold value (tutorialspoint, n.d.).

3) C: a variable of double type representing the constant used in both methods (subtracted from the mean or weighted mean) (tutorialspoint, n.d.).

6. Once the desired image is obtained from the pre-processing steps in Step 5, an OCR engine is used to extract data from the image.

7. The extracted text is then filtered to match the desired format, including the following:

7.1 Filtering merely names in English, Thai, and Japanese languages. Any other text that does not meet the criteria is set to empty values.

7.2 Filtering numerical scores. Any other text that does not meet the criteria is set to empty values.

7.3 Filtering numbers at specified positions that include the % symbol. Remove the original % symbol, and append a % symbol at the end of the number. If the number does not have a %, add a % symbol at the end. Any other text that does not meet the criteria is set to empty values.

8. The data extracted by OCR is then evaluated for efficiency from the existing actual data using the formula.

$$CER_{normalized} = \frac{S + D + I}{S + D + I + C}$$

C stands for the number of correctly extracted words compared to the actual data.

S stands for the number of changed words compared to the actual data.

D stands for the number of deleted words compared to the actual data.

I stands for the number of inserted words compared to the actual data.

And convert the obtained values to a percentage format by multiplying them by 100.

4. RESULTS

The results of extracting data from RoV competition images using OCR technology, comparing Pytesseract and EasyOCR, can be presented in a graphical format as shown in the figure.

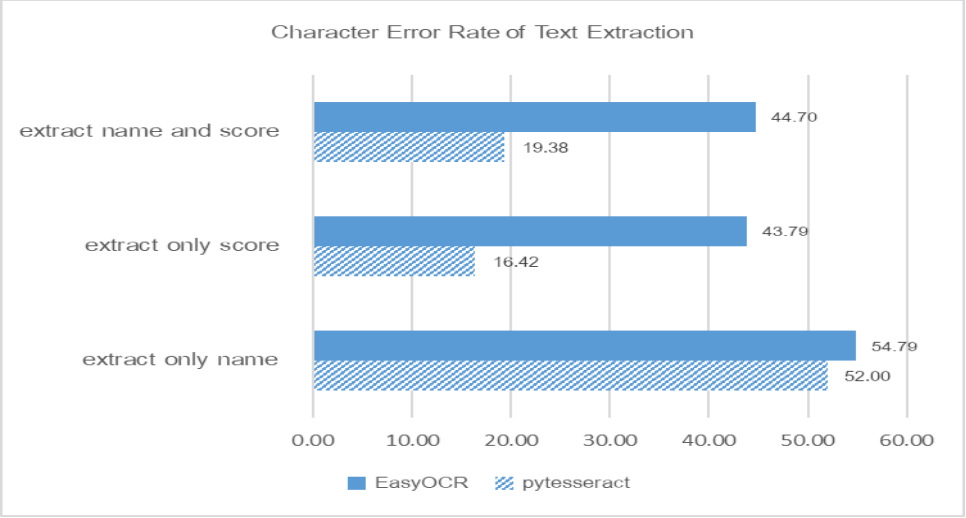

Figure 4. Character Error Rate of Text Extraction (%)

Based on the CER results for text extraction, it was found that Pytesseract had a CER of 52.00% for extracting player names, while EasyOCR had a CER of 54.79%. For extracting scores, Pytesseract had a CER of 16.42%, while EasyOCR had a CER of 43.79%. When extracting both player names and scores together, Pytesseract had a CER of 19.38%, and EasyOCR had a CER of 44.70%.

And the results of the processing time for text extraction using Pytesseract and EasyOCR can be presented in a graphical format, as shown below.

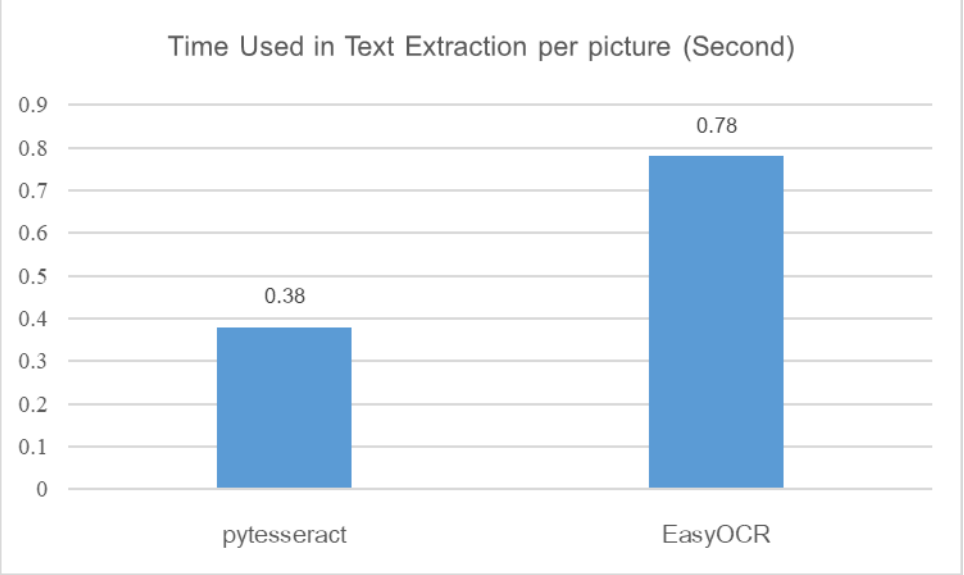

Figure 5. Time Used in Text Extraction per Picture (Second)

Based on the results of the processing time for text extraction comparing Pytesseract and EasyOCR, it was found that Pytesseract takes approximately 0.38 seconds to extract text from one image, while EasyOCR takes approximately 0.78 seconds to extract text from one image.

5. DISCUSSION

According to the CER results for text extraction, it is evident that Pytesseract has a lower error rate compared to EasyOCR. This could be attributed to the fact that the images of RoV game results consist of more than two colors and are not simply black text on a white background. This allows Pytesseract to extract text more accurately from such image characteristics (Liao, 2021). Additionally, from the experimental results, it was found that the CER for extracting player names is higher than that for extracting scores. This is because player names may contain symbols that are not present in the OCR engine's database, resulting in higher error rates. Furthermore, the efficiency of text extraction can depend on various image characteristics such as shadows, background color, text color, and lighting conditions (Kochale et al., 2021, p. 14). Furthermore, the image characteristics also influence the choice of pre-processing methods for image preparation. In addition to that, another factor that affects the performance of text extraction is the appropriate configuration of each OCR engine.

CONCLUSION

Based on the research findings, it can be concluded that Pytesseract performs better in terms of character error rate for extracting text from images of RoV game results compared to EasyOCR. Additionally, when comparing the processing speed for text extraction from images, it was found that Pytesseract is faster, taking approximately 0.38 seconds, while EasyOCR takes approximately 0.78 seconds.

Recommendation

The findings of this research can be applied as a guideline for selecting OCR engines for extracting text from other types of images. It can also be used in developing programs for managing RoV tournaments by extracting data from statistical images to summarize tournament results or present them in various formats according to specific needs. For example, it can be utilized to generate promotional content for each round of the competition by showcasing comparative statistics from previous matches of each team to motivate viewers and increase their interest in watching the matches. Additionally, the findings can be applied to extract statistical data from other esports competitions that have similar image characteristics to RoV matches to leverage the extracted data for future competition benefits.

For further study, it may be beneficial to explore other technologies to assist in extracting text from various symbols. In the case of player names in the statistics of RoV competitions, players often use symbols to differentiate themselves from others. These symbols may not be present in the OCR engine's database, so incorporating other technologies such as Natural Language Processing (NLP) can help extract text from symbols. Additionally, the format of statistical images from RoV competitions may vary each year, requiring the adjustment of the pre-processing methods to suit the specific format of each image. This ensures that the text extraction is performed with the highest efficiency and minimal errors. Furthermore, in the future, it may be worth experimenting with other OCR engines that support the Thai language to further improve the extraction of text from statistical images of RoV competitions, aiming for even greater efficiency and accuracy.

Acknowledgment

The researcher would like to express gratitude to Buriram Rajabhat University for supporting the funding for publishing the article in an international journal.

REFERENCES

[1] Anunpattana, P., Khalid, M. N. A., Yusof, U. K., & Iid, H. (2018). Analysis of Realm of Valor and its business model on PC and mobile platform comparison. Asia-Pacific Journal of Information Technology and Multimedia/Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik, 7(2-2), 1-11.

[2] Chumwatana, T., Rattana-amnuaychai, W., Chauychu, P. (2022). Patient information extraction using optical character. Journal of the Thai Medical Informatics Association, 1, 22-27.

[3] Burkpalli, V., Joshi, A., Warad, A. B., & Patil, A. (2022). Automatic Number Plate Recognition Using TensorFlow and EasyOCR. International

Research Journal of Modernization in Engineering Technology and Science, 4(9), 493-501.

[4] Chadha, A., Kashyap, S., Gupta, M., & Kumar, V. (2020). License Plate Recognition System using OpenCV & PyTesseract. CSI Journal of Computing, 3(3), 31-35.

[5] Chandra, S., Sisodia, S., & Gupta, P. (2020). Optical Character Recognition – A Review. International Research Journal of Engineering and Technology (IRJET), 07(04), 3037-3041.

[6] Digital Economy promotion Agency. (2021). Opportunities and challenges of the Thai esports industry. https://www.depa.or.th/th/article-view/challenges-of-the-Thai-esports-industry

[7] Esports Earnings. (2023). Highest Earnings By Country. https://www.esportsearnings.com/countries

[8] Garena Thailand. (n.d.). Arena of Valor. https://rov.in.th/Hamari, J. (2017). What is eSports and why do people watch it? Internet Research, 27(2), 211-232. https://doi.org/10.1108/IntR-04-2016-0085.

[9] Kochale, A., Khemariya, A., & Tiwari, A. (2021). Real Time Automatic Vehicle (License) Recognition Identification System with the Help of Opencv & Easyocr Model. International Journal of Research, Science, Technology & Management, 24(3), September 2021. 10-15.

[10] Lerksirinukul, P. (2019, April 12). Esports industry worth 34 billion baht! Driving Thailand towards the "eSports Hub" of ASEAN. https://www.salika.co/2019/04/12/garena-world-2019-in-thailand-and-esports-hub-of-asian/

[11] Leung, K. (2021). Evaluate OCR output quality with Character Error Rate (CER) and Word Error Rate (WER). https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510

[12] Liao, C. (2021, December 15). OCR Engine Comparison — Tesseract vs. EasyOCR – The Startup - Medium. https://medium.com/swlh/ocr-engine-comparison-tesseract-vs-easyocr-729be893d3ae

[13] Malkadi, A., Alahmadi, M. & Haiduc, S. (2020). A Study on the Accuracy of OCR Engines for Source Code Transcription from Programming Screencasts. MSR '20, October 5–6, 2020, Seoul, Republic of Korea. https://doi.org/10.1145/3379597.3387468

[14] PyPI. (2022, August 16). pytesseract. https://pypi.org/project/pytesseract/PyPI. (2022, September 16). EasyOCR. https://pypi.org/project/easyocr/Ribeiro, M. R. M. , Julio, D., Abelha, V., Abelha, A. & Machado, J. (2019). A Comparative Study of Optical Character Recognition in Health Information System.

[15] 2019 International Conference in Engineering Applications (ICEA), https://doi.org/10.1109/CEAP.2019.8883448.

[16] Saoji, S., Arora, A., Singh, R., Mangal, A., & Eqbal, A. (2021). Text recognition and detection from images using PyTesseract. Journal of Interdisciplinary Cycle Research, 13(7), 1674-1679.

[17] Sawalkar, A., Pathan, A., Kakade, A., Telvekar, P. & Chandne, B. (2022). Number Plate Recognition System Using Pytesseract & OpenCV. International Research Journal of Modernization in Engineering Technology and Science, 04(06), June, 5861-5863.

[18] Sriborrirux, W. (2018). Automatic Pharmacy Information Leaflet Identification of Organize Antibiotics Drug Safety. http://dspace.lib.buu.ac.th/xmlui/handle/1234567890/3738.

Tutorialspoint. (n.d.). OpenCV - Adaptive. Thresholdhttps://www.tutorialspoint.com/opencv/opencv_adaptive_threshold.htm

DOI: https://doi.org/10.15379/ijmst.v10i4.2272