Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Improving Tesseract OCR Accuracy Using SymSpell Algorithm on Passport Data

Iqbaluddin Syam Had[1)*], Wiga Maulana Baihaqi[2)], Dwi Putriana Nuramanah Kinding[3)]
[1,2)]Amikom Purwokerto University, Indonesia, [3)]Jenderal Soedirman University, Indonesia
[1)]iqbaluddinsh@gmail.com, [2)]wiga@amikompurwokerto.ac.id, [3)]dwiputriana.kinding@unsoed.ac.id

Submitted :Dec 30, 2024 | **Accepted** : Jan 25, 2025 | **Published** : Jan 27, 2025

Abstract: Optical Character Recognition (OCR) is a technology used to recognize text from images or digital documents, such as passports. One popular OCR tool is Tesseract as it offers high accuracy. However, OCR accuracy is often affected by various factors, including image noise and/or non-text elements. This article discusses the application of the SymSpell algorithm for post processing to improve OCR accuracy on standard Indonesian passports. OCR will be focused on the Visual Inspection Zone, specifically the Place of Birth and Issuing Office values. Unlike the Machine Readable Zone which is composed of individual codes and a clear background, the Visual Inspection Zone often experiences OCR errors due to holograms blocking the text and spaced layouts. SymSpell is an edit distance based spelling correction algorithm designed to process data quickly and efficiently, even on very huge datasets. In this study, SymSpell is used to detect and correct errors in OCR results that are compared to a corpus word list. Experimental results with 10 tested scans and passport photos showed that the integration of SymSpell with the Research and Development methodology was able to improve the OCR accuracy rate by 21,43% for certain Place of Birth and Issuing Office data from the Visual Inspection Zone. With this approach, OCR systems can provide more reliable results for practical applications.

Keywords: Corpus; Optical Character Recognition; Passport; SymSpell; Tesseract;

INTRODUCTION

The era of digitalization has changed a lot of behavior in daily work (Kurniawan, 2023), without exception in the process of inputting data from file documents. A widely used approach to accelerate data entry from an image is Optical Character Recognition (OCR). OCR is a technology that recognizes text in digital images (Mubeen et al., 2022). One of the most popular OCR tools is Tesseract (Moussaoui et al., 2024), Tesseract came with high accuracy in character recognition (ÇeliK, 2021). However, OCR accuracy is significantly affected by noise, document degradation, and non-text elements (Konanykhin et al., 2023).

Passport is a state-owned document that serves as proof of identity when abroad (*Paspor Biasa – Kantor Imigrasi Kelas I Non TPI Depok*, n.d.). Passports are used when entering the borders of other countries. Then the authorities of the destination country will stamp the visa or attachment sheet attached to the passport page as proof of permission to enter a country. Considering the importance of the passport it has become a mandatory document used for issuing international flight tickets and visas. Ticket booking service and visa providers are required to always accelerate the quality of their services in terms of inputting passport data, this is in line with this research which discusses improving the quality of OCR on passport data.

Passport identity page contains 2 main areas, MRZ (machine-readable zone) and Visual Inspection Zone. The MRZ area, composed by individual codes, met the standard of ICAO 9309, which can be captured and processed using OCR systems (Bessmeltsev et al., n.d.) then extracted to be information needed. But the data contained in the MRZ is limited to name, passport number, gender, date of birth, expiration date. While other data that is no less important like Place of Birth and Issuing Office must be taken from the Visual Inspection Zone. There are several challenges while reading Visual Inspection Zone ranging from holograms that cause noise in the image like shown in Figure 1 below, as well as the spaced layout which may make OCR getting errors.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fig. 1 Example of noise in Indonesian passport

OCR errors can have a significant impact, especially in the context of identity verification. For example, an error in recognizing the letter "O" as "0" or "I" as "1" may cause a failure in the data matching system. To address this issue, a OCR post-processing text error correction approach is needed to improve system reliability (de Oliveira et al., 2023). OCR post-processing has significantly improved during the past few years (Hemmer et al., 2023). Both Place of Birth and Issuing Office values from Visual Inspection Zone are predictable values. OCR systems may experience misreadings but if coupled with OCR post-processing these misreadings can be minimized with spelling correction algorithms.

Wolf Gabre introduced the Symmetric Delete Spelling Correction (SymSpell) algorithm in 2012 as an extension of Peter Norvig's algorithm (Garbe, 2012a). SymSpell is an edit-distance based algorithm designed to quickly correct spelling errors, even in large datasets. It optimizes the search process through a precomputed hash map approach, which enables efficient word matching with a low error rate. Wolf Garbe (2012a, 2012b) disclosed that one of the applications for SymSpell is OCR post-processing. By applying SymSpell to the OCR results of passport data, it is possible to detect and correct text errors automatically, improving data recognition accuracy.

Implementation of the SymSpell algorithm for post-OCR has been done by some researchers before. Citing Bjerring-Hansen (2022), introduced a heuristic OCR correction pipeline for 19th-century Danish Fraktur texts, combining safe error corrections, context-based selective adjustments, and SymSpell spell-checking, achieving a 73% improvement in word accuracy and demonstrating the viability of non-ML approaches for resource-constrained projects (Bjerring-Hansen et al., 2022). The study on myOCR by Thura Aung et al., (2024) introduces a Myanmar Optical Character Recognition system that combines CNNs for feature extraction, BiLSTM for sequence modeling, and CTC for decoding, with post-OCR corrections prominently featuring SymSpell for statistical correction alongside neural and LLM-based methods, achieving a chrF++ score of 99.31 and reducing Word Error Rate to 0.66% (Aung et al., 2024).

SymSpell accuracy and performance proofing has been done by Audah et al., (2023) and Ferdiansyah et al., (2023). Both researchers compared SymSpell with other spelling correction algorithms. Audah et al., (2023) comparing SymSpell with Damerau-Levenshtein Distance with the Trie Data Structure, the result is that SymSpell has better performance on non-context sentences with 99.39% accuracy and 66.79% for accuracy with the best match. Second comparison by Ferdiansyah & Nuryana, (2023) is comparing SymSpell with the Burkhard Keller Tree for Indonesian spelling correction. The SymSpell algorithm leads in computation time which records the highest time at 0.05 seconds, while BK Tree is only able to record the highest time at 56 seconds.

This research explores the application of OCR post-processing with the SymSpell algorithm to correct text errors based on the corpus of Place of Birth and Issuing Office on passport data to improve accuracy of OCR results. We also discuss some optimization strategies such as image preprocessing in the OCR process to improve text reading accuracy. With the good performance of the SymSpell algorithm and the close accuracy of tesseract-OCR library as mentioned above are expected to be a practical solution to improve the quality of OCR data on passport documents.

METHOD

According to the working principle of SymSpell, it's as follows: SymSpell is an edit distance algorithm that generates words without modifying transposition, replacement, or insertion operations on the word. The only operation used is deletion, then relating it to the original element. Speed of SymSpell depends on the Symmetric Delete spelling algorithm, and memory requirement is managed by prefix indexing (Tolegenova, 2022). Example of a SymSpell operation can be seen in table 1 which tests the error word KATYA.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1Example operation of SymSpell edit-distance

KATYA	KARYA	Term
ATYA	ARYA	Edit-distance: 1
KTYA	KRYA	
KATA	**KAYA**	
KAYA	KARA	
KATY	KARY	
TYA	RYA	Edit-distance: 2
AYA	**AYA**	
etc..	etc..	

As seen from table 1, there is an example of similarity in the word KAYA at edit-distance 1 and the word AYA at edit-distance 2. With the similarity of words found from the combination of letter deletion, the word KARYA will be included in the right word recommendation for searching the word KATYA.

Symspell is applied to the existing passport OCR system to correct word errors from the OCR results. According to Sukmadinata (2012) in Wynarti (2018) research expresses that the Research and Development method is either used to make new products or improve existing products. Therefore, this research will use the Research and Development (RnD) method framework to apply the SymSpell algorithm to the optical character recognition system to reduce the system reading error rate. The steps of Research and Development for this research are as follows.

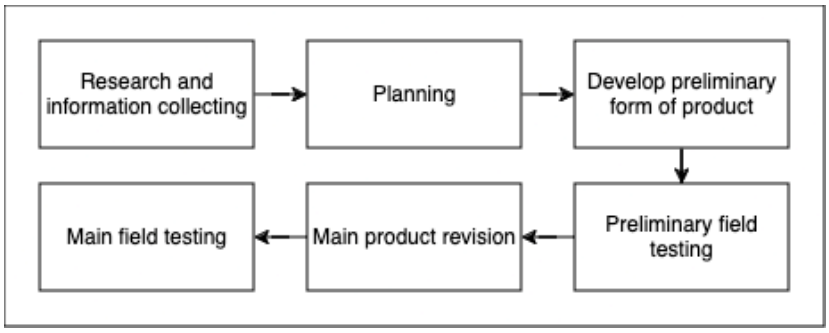

Fig. 2 Research and Development (RnD) steps used in this article

1. Research and Information Collecting
   The first step of Research and Development in this research is Research and information collecting. We collect any information about the SymSpell algorithm, and supporting data that will become the corpus for the symSpell algorithm. This research uses 2 corpuses, the first is a corpus of a list of regencies and cities in Indonesia to predict the value of place of birth, then the second corpus is a list of immigration branch offices in Indonesia to predict the value of issuing offices on passports.
2. Planning
   The planning stage will make a plan to develop a problem solution based on the data and knowledge obtained from the Research and Information collecting stage. Planning involves preprocessing strategies such as changing the color scheme to grayscale and adjusting the brightness to reduce noise in the image. Followed by making all words in the corpus and OCR results into one word by removing spaces.
3. Develop Preliminary Form of Product
   The plan that has been made will be implemented at this stage into a prototype to be tested.
4. Preliminary Field Testing
   After the prototype is built, limited testing is carried out to determine the performance of the tesseract-OCR and SymSpell algorithms in identifying place of birth and issuing office data on passports. Testing is performed by OCR processing on 10 different passport images. The measurement of success is seen by comparing before and after processing using the SymSpell algorithm.
5. Main Product Revision
   The results of limited testing will provide input for the system to work better, the existing input or evaluation will be improved at this stage.
6. Main Field Testing

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

After the system is made, tested with limited testing, then has been improved from the results of limited testing, the last stage is the final testing of the product that has been made. Testing at this stage will be the final result of the implementation of the system in this study.

RESULT

Pre-processing

Preprocessing contains the preparation stage of passport image data before the OCR process. The passport image is being split into two parts, which are MRZ (Machine Readable Zone) and Visual Inspection Zone areas, then the color is changed to gray and the brightness is set at 120 percent using a sharp javascript library. This part is separated because each area will be processed using a different data train by tesseract. MRZ areas will be processed using the mrz.traineddata model, this model has been trained by DoubangoTelecom with the dataset of more than 7 thousand images (.tif) with ground truth (.gt.txt) from Google images augmented by several synthetic data (DoubangoTelecom, n.d.).

Fig. 3 Visual Inspection Zone

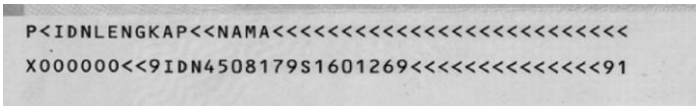

Fig. 4 Machine-Readable Zone

To ensure that the ratio of images to be processed is consistent, we apply a cropping process after the user selects an image on the image upload form.

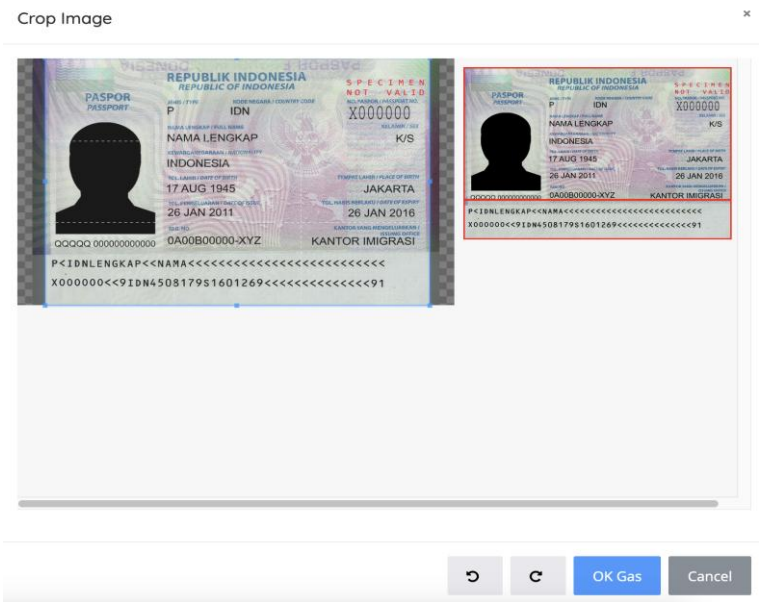

Fig. 5 Cropping passport image

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

OCR Processing

Our OCR process uses the tesseract library with two different data trains. For the MRZ part we used mrz.traineddata and for the visual inspection zone we used ind.traineddata. The reading process in the visual inspection zone is specifically done in the area of the place of birth and the issuing office column.

OCR Post-processing

From the results of the OCR process, the values obtained are then processed using SymSpell. The values of birthplace and issuing office obtained previously were normalized first by removing spaces and special characters to reduce noise in the data. The values are then processed using the SymSpell algorithm to find whether there is a word in the word list in the corpus or a word that is considered similar. If SymSpell returns a recommendation word then the system will take the recommendation and display the results to the user. However, if SymSpell also cannot find a word recommendation from the corpus then the system will give a null or empty return. The OCR post-processing flow can be seen as follows.

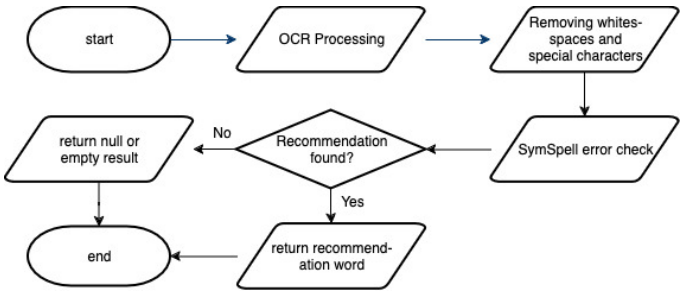

Fig. 6 OCR post-processing flowchart

Preliminary Field Testing

We performed limited testing of the planning stage implementation with 10 different passport images with maximum edit-distances value of 3 for the SymSpell algorithm. These tests were conducted using the default parameters of tesseract OCR with 3 is values for PSM, 3 for OEM and ind.traineddata for train data. The results can be found in table 2.

Table 2. Preliminary field testing results

No	Place of Birth		Issuing Office	
	Raw OCR	SymSpell Result	Raw OCR	SymSpell Result
1	BALI	BALI		
2			BANDUNG	BANDUNG
3	BANDUNG	BANDUNG		
4	sakarta D	JAKARTA	DEPOK	DEPOK
5			BEKASI	BEKASI
6	JAKARTA.	JAKARTA	JAKARTA UTARA	JAKARTA UTARA
7	JAVARTA l	JAKARTA	N ergait A BARAT	
8	JAKARTA:	JAKARTA	BEKASI	BEKASI
9	..JAKARTA	JAKARTA	TANJUNG PRIOK	TANJUNG PRIOK
10			JAKARTA TIMUR.	JAKARTA TIMUR

Main Product Revision

From the preliminary test, the SymSpell algorithm successfully corrected 2 words from the OCR system reading error. Then SymSpell also failed to find a word recommendation in the issuing office column number 7 with the corpus. The error occurred because the issuing office value number 7 had too much noise so it could not find a match with the corpus at the maximum edit-distance value of 3. In other cases, the OCR system also failed to read 3 data in the birthplace column and 2 data in the issuing office column.

In order to handle this particular issue, we used a different setting for tesseract on the PSM (Page segmentation modes) value. Previously we used the default setting value of 3 (Fully automatic page segmentation, but no OSD). In Main field testing we changed the PSM value to 12 which is Sparse text with OSD. PSM 12 is designed to read text from images that have irregular and sparse layouts, OCR will scan the image in its entirety allowing the output of a lot of unnecessary text.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14395
e-ISSN : 2541-2019
p-ISSN : 2541-044X

To overcome the amount of noise in the OCR read text, SymSpell algorithm processing is done line-by-line. When SymSpell succeeds in finding a word recommendation for more than one line, the word from the line that has the lowest edit-distance will be selected and then the recommended word is displayed to the user. In addition, we provide the computation time to determine whether processing multiple lines will make the processing time suffer.

Main Field Testing

Testing at this stage still uses the same data as preliminary field testing to find out the differences after the revision process. The results of main field testing can be seen in table 3.

Table 3
Main field testing results

No	Place of Birth			Issuing Office		
	Raw OCR	SymSpell Result	Computation time (ms)	Raw OCR	SymSpell Result	Computation time (ms)
1	TemPer ane MAA aatu BALI	BALI	0.0047	BANGGA!	BANGGAI	0.0045
2				BANDUNG	BANDUNG	0.0042
3	WieerLann / Pae De t BANDUNG	BANDUNG	0.0050			
4	sakarta D	JAKARTA	0.0051	DEPOK	DEPOK	0.0040
5	JAKARTA	JAKARTA	0.0047	aa BEKASI	BEKASI	0.0055
6	an eee te aenceaan JAKARTA.	JAKARTA	0.0054	| MANDON WA pMa lAAA JAKARTA UTARA	JAKARTA UTARA	0.0051
7	Yiioea aa JAKARTA |	JAKARTA	0.0076	N ergait A BARAT		0.0069
8	“Yebrar LAhm 1 PAI CE ORTA JAKARTA:	JAKARTA	0.0217	AAA n aa an OeRG AA BEKASI	BEKASI	0.0079
9	..JAKARTA	JAKARTA	0.0057	TANJUNG BRIOK	TANJUNG PRIOK	0.0057
10	nbe ie n JAKARTA	JAKARTA	0.0053	JAKARTA TIMUR.	JAKARTA TIMUR	0.0043

From the main field testing column issuing office example number 5, the OCR system generates two lines of text, which are “aa” and “BEKASI”. When the SymSpell algorithm processes the 2 words the algorithm produces word recommendations from both words. For the word “aa” it produces the city “AGAM” with an edit-distance of 2, and the word “BEKASI” produces an edit-distance of 0. The system selects the word from the line that has

*name of corresponding author

the lowest edit-distance, which is the word "BEKASI". With this approach, the algorithm successfully selects the correct word.

DISCUSSIONS

Based on the test, it showed that SymSpell can work properly to correct minor misspellings of ocr readings. For major misspellings, a larger maximum edit-distance is needed in order for SymSpell to recommend more similar words. However, a large edit-distance can lead to worse computation time. Therefore, the maximum edit-distance value that is ideal for a project needs to be considered.

This research still has limitations, including the corpus of place of birth and issuing office used are the names of cities in Indonesia, if there is passport data whose value of place of birth or issuing office outside the country then the algorithm cannot find the right word. In addition, the types of passports used in this study only support ordinary Indonesian passports. Other passports such as official passports or foreign passports may have different visual inspection zone layouts.

CONCLUSION

The symSpell algorithm for post-processing OCR passport data successfully improves accuracy in accordance with the expected output. The increase in accuracy at the final stage was achieved by 21.43% with average computation time at 0.00629 ms. This result was gained from the comparison before and after post-processing using the SymSpell algorithm. Whereas before post-processing only 14 values of place of birth and issuing office were successfully read correctly, then increased into 17 values that match the values in the passport with the help of the SymSpell algorithm. The results of the revision process made the main field testing have better accuracy than the preliminary field testing because three previously unidentified data could be corrected in the main field testing. Changes in PSM settings (page segmentation modes) in tesseract and changes in the SymSpell algorithm approach are the most significant factors in this stage of improvement.

However, even though the improvement has been successful, this research can still be developed further. In this research test found some values that are not being read by the tesseract, this can be addressed by methods such as adjusting the brightness dynamically to adjust the original lighting conditions in the passport image to the point of optimal lighting for the tesseract. Image classification to determine the type of passport can also be adopted, because there are several types of passports and each has a different layout of visual inspection zones, so the reading of the place of birth and issuing office must be adapted to the type of passport being processed. The corpus dataset can also be added so it can provide recommendations for cities outside Indonesia.

REFERENCES

Audah, H. A., Yuliawati, A., & Alfina, I. (2023). A Comparison Between SymSpell and a Combination of Damerau-Levenshtein Distance with the Trie Data Structure. *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA59291.2023.10390399

Aung, T., Thu, Y. K., & Oo, M. N. (2024). myOCR: Optical Character Recognition for Myanmar language with Post-OCR Error Correction. *2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1–6. https://doi.org/10.1109/iSAI-NLP64410.2024.10799448

Bessmeltsev, V., Bulushev, E., & Goloshevsky, N. (n.d.). *High-speed OCR algorithm for portable passport readers*.

Bjerring-Hansen, J., Kristensen-McLachlan, R. D., Diderichsen, P., & Hansen, D. H. (2022). Mending Fractured Texts. A heuristic procedure for correcting OCR data: 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022. *CEUR Workshop Proceedings*, *3232*, 177–186.

ÇeliK, A. (2021). Eğik Karakter Tanıma Başarısını Arttırmak için Yeni Bir Yöntemin Kullanılması. *Harran Üniversitesi Mühendislik Dergisi*, *6*(1), 1–11. https://doi.org/10.46578/humder.720001

de Oliveira, L. L., Vargas, D. S., Alexandre, A. M. A., Cordeiro, F. C., Gomes, D. da S. M., Rodrigues, M. de C., Romeu, R. K., & Moreira, V. P. (2023). Evaluating and mitigating the impact of OCR errors on information retrieval. *International Journal on Digital Libraries*, *24*(1), 45–62. https://doi.org/10.1007/s00799-023-00345-6

DoubangoTelecom. (n.d.). *DoubangoTelecom/tesseractMRZ: Ready-to-use MRZ / MRTD (Machine-readable zone/travel documents) dataset and models for tesseract v4*. Retrieved January 23, 2025, from https://github.com/DoubangoTelecom/tesseractMRZ

Ferdiansyah, M. H., & Nuryana, I. K. D. (2023). Analisis Perbandingan Metode Burkhard Keller Tree dan SymSpell dalam Spell Correction Bahasa Indonesia. *Journal of Informatics and Computer Science (JINACS)*, 305–313. https://doi.org/10.26740/jinacs.v4n03.p305-313

Garbe, W. (2012a, June 7). *1000x Faster Spelling Correction algorithm*. SeekStorm. https://seekstorm.com/blog/1000x-spelling-correction/

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Garbe, W. (2012b, June 7). *SymSpell*. https://github.com/wolfgarbe/SymSpell

Hemmer, A., Brachat, J., Coustaty, M., & Ogier, J.-M. (2023). Estimating Post-OCR Denoising Complexity on Numerical Texts. In N. T. Nguyen, S. Boonsang, H. Fujita, B. Hnatkowska, T.-P. Hong, K. Pasupa, & A. Selamat (Eds.), *Recent Challenges in Intelligent Information and Database Systems* (pp. 67–79). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42430-4_6

Konanykhin, A., Konanykhina, T., & Panishchev, V. (2023). Character Recognition in Images under High Noise Levels. *2023 International Russian Automation Conference (RusAutoCon)*, 930–935. https://doi.org/10.1109/RusAutoCon58002.2023.10272925

Kurniawan, Z. (2023). DAYA SAING SUMBER DAYA MANUSIA DI ERA DIGITALISASI. *Jurnal EBI*, *5*(2), Article 2. https://doi.org/10.52061/ebi.v5i2.182

Moussaoui, H., Akkad, N. E., Benslimane, M., El-Shafai, W., Baihan, A., Hewage, C., & Rathore, R. S. (2024). Enhancing automated vehicle identification by integrating YOLO v8 and OCR techniques for high-precision license plate detection and recognition. *Scientific Reports*, *14*(1), 14389. https://doi.org/10.1038/s41598-024-65272-1

Mubeen, Dr. S., Brahmani, J., Kalyan, D. P., Jagirdar, A., & Kumar, A. P. (2022). Optical Character Recognition Using Tesseract. *International Journal for Research in Applied Science and Engineering Technology*, *10*(11), 672–675. https://doi.org/10.22214/ijraset.2022.47414

Paspor Biasa – Kantor Imigrasi Kelas I Non TPI Depok. (n.d.). Retrieved November 29, 2024, from https://depok.imigrasi.go.id/paspor-biasa/

Sukmadinata, N. S. (2012). *Metode penelitian pendidikan*. Bandung: PT Remaja Rosdakarya.

Tolegenova, A. (2022). AUTOMATIC ERROR CORRECTION: EVALUATINGPERFORMANCE OF SPELL CHECKER TOOLS. *Natural and Technical Sciences*, *58*(1), Article 1. https://doi.org/10.47344/sdubnts.v58i1.690

Wynarti, I. A. (2018). PENGEMBANGAN PERMAINAN CHARADES SEBAGAI MEDIA PEMBELAJARAN MATERI JENIS-JENIS BISNIS RITEL KELAS XI PEMASARAN DI SMK NEGERI 2 BUDURAN. *Jurnal Pendidikan Tata Niaga (JPTN)*, *6*(2). https://doi.org/10.26740/jptn.v6n2.p%p

*name of corresponding author