P-ISSN: 2715-2448 | E-ISSN: 2715-7199
Vol.5 No.2 June 2024
Buana Information Technology and Computer Sciences (BIT and CS)

Identification of Socio Economic Registration Data Using OCR Based Tesseract and Google Cloud Vision

Lionardi Ursaputra Pratama [1]*, Aviv Yuniar Rahman [2], Rangga Pahlevi Putra [3]
[1 2 3] Department of Informatics Engineering, Faculty of Engineering, Universitas Widyagama Malang, Indonesia
E-mail: ursaputra@cordelia.id*, aviv@widyagama.ac.id, rangga@widyagama.ac.id

Abstract:

The Indonesian government program, called Socio-Economic Registration (Regsosek), aims to measure and monitor the socio-economic conditions of low-income people. One of the relevant data used for research is Regsosek. This method is used to analyze the influence of economic and social infrastructure on economic growth, analyze the socio-economic determinants of ownership of work accident insurance for informal workers, create a women's socio-economic vulnerability index (IKSEP), and study intercultural literacy from a social, economic and political perspective. The success of the government's Socio-Economic Registration program depends on the role of data collection officers or surveyors, who directly interact with the community to obtain information about Socio-Economic Registration (Regsosek) data collection. This method also has other obstacles that significantly affect the overall results of the survey, where the survey results must be entered manually by the surveyor from a form with handwritten data, after which it is entered into the website. This method is vulnerable to human error, where the handwriting is difficult to read, and mistakes are made during the data input. The technology that can be used to handle this problem is implementing the OCR method, where writing that was initially handwritten manually can be identified and converted into digital text that can be edited (editable text) and processed automatically. This research shows that the proposed method has good accuracy, with an Accuracy of 96.45%, CER 0.3%, and WER 4.30%.

Keywords: Hand Writing, Optical Character Recognition, Socioeconomic infrastructure, Surveyor officer.

I. Introduction

The Indonesian government program, called Socio-Economic Registration (Regsosek), aims to measure and monitor the socio-economic conditions of low-income people. According to this program, social protection is the government's way of dealing with poverty in Indonesia [14]. One of the relevant data used for research is Regsosek. This method is used to analyze the influence of economic and social infrastructure on economic growth, analyze the socio-economic determinants of ownership of work accident insurance for informal workers, create a women's socio-economic vulnerability index (IKSEP), and study intercultural literacy from a social, economic and political perspective [11].

The success of the government's Socio-Economic Registration program depends on the role of data collection officers or surveyors, who directly interact with the community to obtain information about Socio-Economic Registration (Regsosek) data collection. So that the public as respondents can provide answers that are appropriate to the conditions, this must be explained well in the field. However, there are obstacles to getting transparent information from respondents. This method also has other obstacles that significantly affect the overall results of the survey, where the survey results must be entered manually by the surveyor from a form with handwritten data, after which it is entered into the website. This method is vulnerable to human error, where the handwriting is difficult to read, and mistakes are made during the data input [12].

The technology that can be used to handle this problem is implementing the OCR method, where writing that was initially handwritten manually can be identified and converted into digital text that can be edited and processed automatically (Memon et al., 2020). Before the system can recognize human handwritten image patterns, information representing the image must be retrieved, known as input data. A scanning process is carried out on the resulting image to obtain digital data, and then a preprocessing process is carried out [7]. This process can complete the creation of an intelligent computer system that can recognize handwriting. Often, these methods are combined with other algorithms to create new applications that can solve more complex problems [5].

Currently, OCR is applied in various fields, including data entry. Previous research utilized OCR as automation for logistics warehouse data input management [2]. The results of the introduction of OCR are used as input for the logistics warehouse management system, which will then be processed by the management system and entered into the database. In detecting handwriting, this research produces an accuracy of 46.9%. It shows that Tesseract can be implemented to convert images into text. However, the recognition results depend on the quality of the image and the variety of text used as test data and taking images requires sufficient light. Other research uses Tesseract OCR for text recognition in social survey data. This research produced a CER of 2.60% and a WER of 25.20% using the Tesseract method [1].

The difference in accuracy and error rate is due to the library's dependence on the test document's quality, so the OCR output is often inappropriate. Both studies show that the quality of the image or document used as a source scanned by an OCR application is the main problem in using OCR on handwritten text.

Tesseract is an open-source OCR application created by HP from 1984 to 1994. Tesseract was first created as a graduate project and released by Hewlett Packard and the University of Nevada, Las Vegas, in 2005. Now, Google partially funds its development, and version 2.0, Tesseract 4. x, was released in December 2019. Tesseract has four main functional modules: static character classifier, word recognition, paragraph and sentence search, linguistic analysis, and adaptive classifier. However, if the image is preprocessed, the results from Tesseract will be much better.

To overcome this problem, research was carried out to automate data input in handwriting from survey forms using the OCR method based on Google Vision and Tesseract. This research aims to create a system that can be used to recognize handwriting so that it can be easily recognized as text using computer vision methods.

II. Methods

This research is divided into several general stages. This stage includes literature study, planning, data collection, implementation, testing, and report preparation. These stages will be carried out sequentially to produce an optimal report. The general stages of research can be seen in Figure 1. The literature study phase involves a comprehensive review of existing research, theories, and relevant publications to establish a solid theoretical framework for the study.

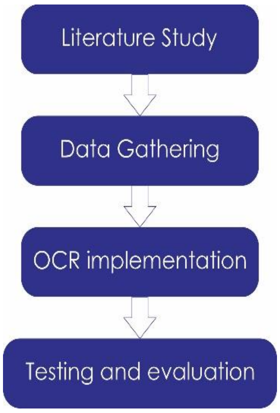


Fig1. Study Method

A. Literature Study

In this research, the references used come from books, journals, ebooks, previous research and other reliable sources. The literature study used Google Scholar as the primary search tool for references related to text detection, Tesseract and Google Vision. Literature research was carried out to obtain information about several main topics of this research, namely:

a) OCR
b) Google Cloud Vision
c) Tesseract
d) Merged Method

B. Data Gathering

The required data collection stage is the most important in this research. The data collected is used for system testing, whether the method used is optimal for the test data. The survey form image data was collected by scanning 100 data.

C. OCR Implementation

Based on Figure 2, this research began by entering the Survey Data image. After that, the image is processed for thresholding and greyscale. This pre-processing is done so OCR can more easily differentiate between objects and backgrounds. Next is detecting text; Tesseract and Google Vision are used to extract images into ASCI characters. After the ASCI characters are obtained, compare the results of the two methods, and which is the best with Ground Truth? Then, the best results will be normalized and entered into the ignore list process to determine which words will be used and which will be ignored. The following process is testing, which will use the CER, WER and Accuracy parameters. These metrics will assess the performance of the OCR methods and provide valuable insights into their effectiveness in accurately converting the images into editable text, thereby ensuring the reliability of the data obtained from the survey images. Calculations of accuracy, Word Error Rate (WER), and Co-efficient of Error Rate (CER) can be made using formulas 1, 2, and 3, respectively.

$$\mathrm{CER} = (\mathrm{Insertion} + \mathrm{Removal} + \mathrm{Substitution})/\mathrm{TotalnumberofcharactersinGroundTruth} \quad (1)$$

$$WER = (Insertion + Deletion + Substitution)/TotalnumberofwordsinGroundTruth \quad (2)$$

$$Accuracy = (1\text{–}\, average\ CER) * 100 \quad (3)$$

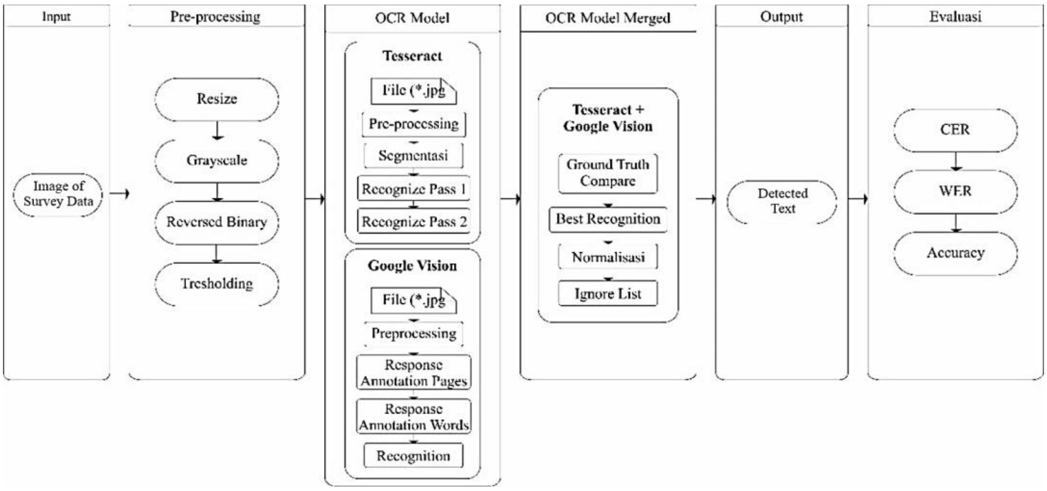


Fig 1. Optical Character Recognition Process

These assessments were fundamental in ensuring the fidelity and precision of the data derived from the survey images, thereby bolstering the overall quality and trustworthiness of the research outcomes.

III. Results

A. Literature Study

a. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is an application that can recognise ASCII characters in digital photographs and turn them into editable text data [6].

As seen in Figure 3, the OCR procedure often comprises multiple steps. The three main processes in this procedure are preprocessing, feature extraction, and recognition. Preprocessing entails improving the image and reducing noise, feature extraction finds essential elements, and recognition converts these elements into text [9].

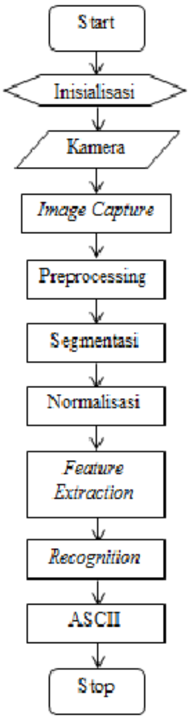


Fig 2. OCR Process Flowchart

b. Tesseract

Tesseract is a widely used optical character recognition (OCR) library for processing scanned images for text recognition. This engine uses neural networks to analyze image patterns and recognize characters with high accuracy. This library can recognize characters in various languages, such as Arabic, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, English, French, German, Greek, Hindi, Indonesian, and Italian. Some studies have suggested using convolution-based preprocessing with specific kernels to improve the accuracy of the Tesseract OCR engine [3].

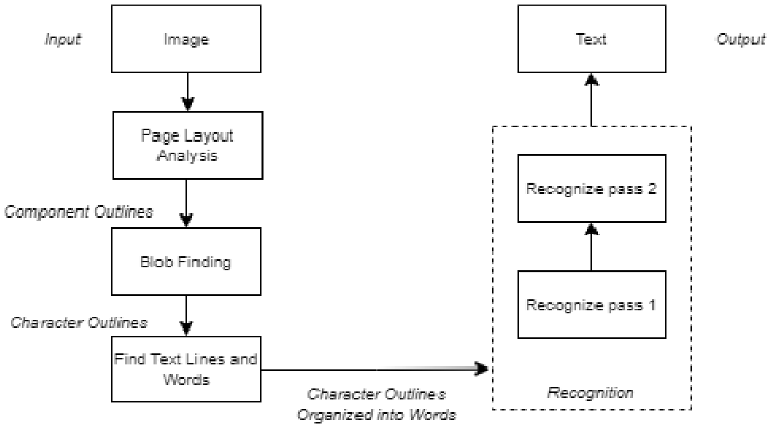


Fig 3. Architecture of Tesseract Library

First, the text area in the image is identified through page layout analysis. Next, the found text areas are divided into several "blobs". Blobs are classifiable units consisting of multiple characters or parts of multiple characters [10]. The third step is to determine the lines of text and combine the blobs into a series of words that fill the space. The first step is to prepare the words for recognition. The next step is to recognize each word in two routes. Tesseract breaks down and merges the single word in each path into blobs, forming a series of identifiable character outlines. Tesseract recognizes the character outline on the first computer with a static classifier based on the feature library. The architecture of Tesseract is depicted in Figure 4.

c. Google Vision

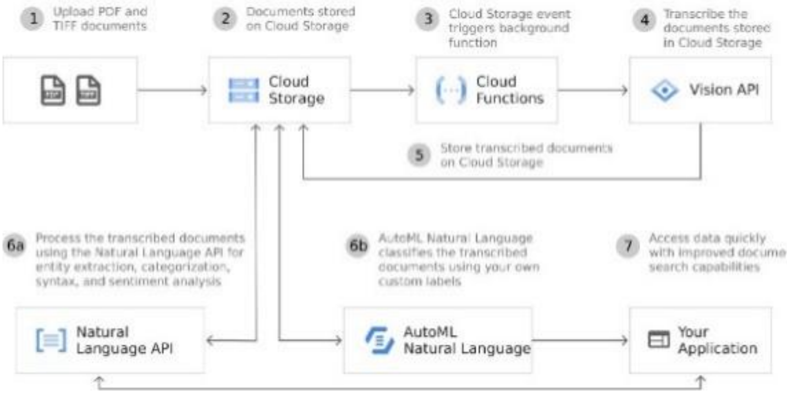
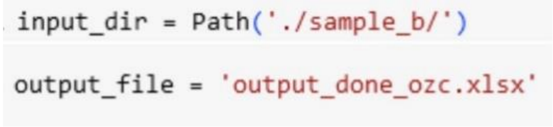

1 Upload PDF and TIFF documents
2 Documents stored on Cloud Storage
3 Cloud Storage event triggers background function
4 Transcribe the documents stored in Cloud Storage
Cloud Storage
Cloud Functions
Vision API
5 Store transcribed documents on Cloud Storage
6a Process the transcribed documents using the Natural Language API for entity extraction, categorization, syntax, and sentiment analysis
6b AutoML Natural Language classifies the transcribed documents using your own custom labels
7 Access data quickly with improved docume search capabilities
Natural Language API
AutoML Natural Language
Your Application

Fig 4. Flowchart of Google Vision

Google Vision is a computer vision technology developed by Google that uses machine learning to analyze and understand visual information in images and videos. This technology offers many features, such as image processing, face processing, object detection, and labelling. The Google Cloud Vision API allows developers to integrate this service into their applications [4]. The utilization of Google Vision can be implemented, as shown in Figure 5.

B. Implementing Method

The end product of this research is an application that can be used to extract text from handwritten text photographs. Python is the programming language used to create the program, and it runs on the Google Colab platform. Google Colab produces an Excel file with the file name and detection results as part of its output. And the outcomes of detection. The Excel file's detection findings are anticipated to make it easier for surveyors to interpret handwritten responses on survey sheets. The user will be prompted to input the image directory and the output file name in the main program, which is an Excel file. Figure 6 displays the program display.



```
input_dir = Path('./sample_b/')

output_file = 'output_done_ozc.xlsx'
```

Fig 5. Input and Output Program

Setting up the Google Cloud Vision API service credentials using the **GOOGLE_APPLICATION_CREDENTIALS** environment variable completes the procedure. Next, the input directory's image file list (img_files) is initialised.

One method of obtaining truth data is using an Excel file, previously mentioned in ground_truth_file. The file's text data is loaded into a Pandas data frame, and the 'text' column is adjusted to change the text to lowercase (lowercase). Then, in identifying text in photos, a list of words (ignore_list) that will be disregarded includes pointless or useless terms. Every picture file in the img_files directory undergoes one iteration. Tesseract OCR with Google Cloud Vision API's OCR (Optical Character Recognition) technique extracts text from each read image.

Additionally, word similarity between the resultant text and the truth text (ground_truth_text) was determined using the Natural Language Toolkit (NLTK) library's CER (Character Error Rate) and WER (Word Error Rate) metrics. The text detection findings are displayed as graphics with distinct colour markers for each found word.

For every word in ground_truth_df, an iteration loop completes the process. The average is then determined by adding up the findings of the CER and WER measurements. The application then uses a Pandas DataFrame-which will be added or built if it does not exist-to store the detection results in an Excel file.

The print programme averages the CER and WER from the Tesseract technique, Google Vision, and a mixture of both after all photos have been analysed. The accuracy of each approach is then printed by the programme as well. This program records and saves the detection findings into an Excel analysis file for additional study. It also gives a comprehensive report on the accuracy of the text detection results from the two OCR algorithms utilised. Figure 7 displays certain outcomes from the programme output.

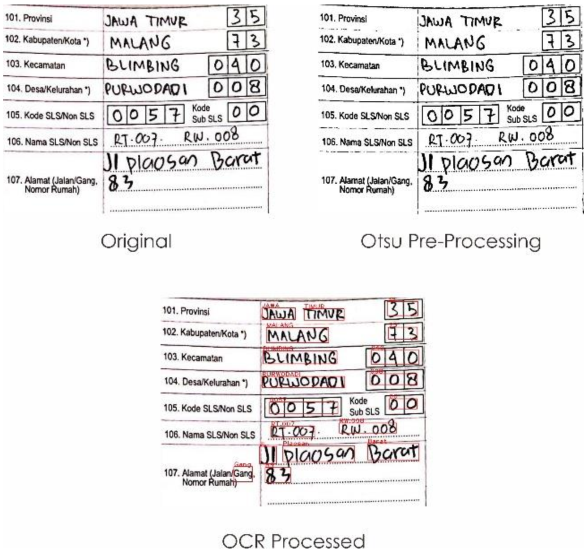


Fig 6. Output results and pre-processing

The ignore list condition is used in the subsequent findings, as shown in Table 1. During the pre-processing step, the Otsu approach gives the image better contrast, eliminates noise, and facilitates program processing. All the text in the image is readable correctly when viewed from the output the programme generated using the suggested way.

Table 1. Text Extraction Result

Ground Truth	Ignore List	Result
-	101. Provinsi	-
-	102. Kabupaten /Kota *)	-
-	103. Kecamatan	-
-	104. Desa/Kelurahan*)	-
-	105. Kode SLS/Non SLS	-
-	106. Nama SLS/Non SLS	-
-	107. Alamat (Jalan/Gang, Nomor Rumah)	-
JAWA TIMUR	-	JAWA TIMUR
35	-	35
MALANG	-	MALANG
73	-	73
BLIMBING	-	BLIMBING
040	-	040
PURWODADI	-	PURWODADI
008	-	008
0057	-	0057
Kode	Kode	-
Sub SLS	Sub SLS	-
00	-	00
RT. 007	-	RT.007
RW. 008	-	RW.008.
JL Plaosan Barat 83	-	JI Plaosan Barat 83

In comparing image processing methods as shown in Table 2, Data 1 uses a more straightforward image processing technique: grayscale on the tested images. Meanwhile, in Data 2, the method applied is more sophisticated, applying the Otsu Thresholding method. *Otsu Thresholding* is a technique used to segment images by determining thresholds automatically based on image histogram analysis. Using grayscale in Data 1 means only using the brightness channel of the image, ignoring colour information. Although simple, grayscale can be helpful in some cases, especially when colour accuracy is not the central aspect required in text or object recognition.

Table 2. Otsu Tresholding Implemented and Grayscale Only

File Name	CER Difference	WER Difference	Accuracy Difference
data-1.jpg	0.23% (Increase)	1.05% (Increase)	-0.23% (Decrease)
data-2.jpg	-0.27% (Decrease)	-0.42% (Decrease)	0.27% (Increase)
data-3.jpg	0.18% (Increase)	0.21% (Increase)	-0.18% (Decrease)
data-4.jpg	-0.08% (Decrease)	0.20% (Increase)	0.08% (Increase)
data-5.jpg	0.13% (Increase)	-0.42% (Decrease)	-0.13% (Decrease)
data-6.jpg	-0.35% (Decrease)	-0.85% (Decrease)	0.35% (Increase)
data-7.jpg	0%	0%	0%
data-8.jpg	0%	0.42% (Increase)	0%
data-9.jpg	0.23% (Increase)	-0.19% (Decrease)	-0.23% (Decrease)
data-10.jpg	0.29% (Increase)	0.42% (Increase)	-0.29% (Decrease)

On the other hand, applying Otsu Thresholding to Data 2 can provide advantages because this method can automatically determine the optimal threshold for separating objects from the background. The table shows the results of a comparative analysis of the performance of Tesseract and Google Vision in terms of Character Error Rate (CER), Word Error Rate (WER), and Accuracy differences between ten different data pictures for Optical Character Recognition (OCR). The two approaches in each photograph show apparent differences in performance metrics. For example, Tesseract and Google Vision both show higher CER and WER in data-1.jpg, which is a factor in Tesseract's decreasing accuracy. On the other hand, Google Vision keeps its CER constant while somewhat increasing its WER. The combined data show a modest loss in overall accuracy along with a minor rise in CER and WER.

Both approaches show a drop in CER and WER in data-2.jpg, with Tesseract showing a more significant improvement over Google Vision. As a result, Tesseract's accuracy increases noticeably, while Google Vision's accuracy slightly increases. The combined results show a drop in WER and CER; however, accuracy has somewhat decreased. From data-3.jpg to data-10.jpg, the remaining photos show consistent patterns of volatility. Specifically, differences in CER, WER, and accuracy point to distinct advantages and disadvantages between Tesseract and Google Vision on various datasets. Using this technique, images may undergo better segmentation, improving the system's ability to better recognize text or objects in the image.

Otsu Method

The Otsu method uses discriminant analysis to find variables that distinguish between two or more naturally occurring groups [15]. Discriminant analysis maximizes these variables to divide objects into foreground and background. The discriminant analysis produces a threshold value that divides a grayscale image into two groups of black and white [13].

C. Analysis

The combination approach proved to be more effective than either one alone. While Tesseract and Google Vision have somewhat higher average CERs-roughly 1.8% to 1.9% for Tesseract and 1.0% to 1.7% for Google Vision-the combined method's average CER is 1.8% to 1.9%.

For the same reason, the combined approach performs better when calculating the Word Error Rate (WER). The combined method's average WER ranges from 1.9% to 2.0%, Tesseract's from 2.0% to 2.5%, and Google Vision's from 1.7% to 2.2%. The error rate in text detection in photos is successfully decreased by combining the output of two OCR algorithms. This demonstrates that using Google Vision in conjunction with Tesseract OCR can yield better accurate results regarding text recognition on photos than either technique alone.

As a result, the program's output results demonstrate that the combined approach can identify text on images with a greater accuracy rate, making it a better option for text recognition applications requiring a low error rate.

Discussion

Reference	Novelty	Results of Previous Research	Results of the Proposed Method
(Arianto et al., 2023)	**Difference:** Previous research produced a CER of 2.60% and a WER of 25.20% in detecting text and did not include the accuracy obtained in the test. We investigated how to reduce the CER and WER values and obtain high accuracy in detecting written text. hand. **Novelty:** This research shows that the combination of OCR using Google Vision and Tesseract is a method with a lower WER level, and there is a decrease in CER and WER compared to before.	CER: 2.60% WER: 25.20% Accuracy: -	CER: 0.3% WER: 4.30% Accuracy: 96.45%
(Berg, So and Seo, 2019b)	**Difference:** This research only produces an accuracy of 46.9% in detecting handwriting, we analyze how accuracy in detecting handwriting can be improved. **Novelty:** This study shows that the most accurate OCR approach and also shows improvements compared to previous methods is the combination of Tesseract and Google Vision.	CER: - WER: - Accuracy: 46.9%	
(Thammarak *et al.*, 2022)	**Difference:** This research only produces accuracy on Tesseract of 47.02% and produces an accuracy of 84.03% in experiments using Google Vision in detecting writing. We analyze how accuracy in detecting handwriting	**Tesseract** CER: - WER: - Accuracy: 47.02% **Google Vision** CER: - WER: - Accuracy: 84.03%	

	can be increased by combining the two methods based on evaluations from previous research. **Novelty:** This study shows that the most accurate OCR approach and also shows improvements compared to previous methods is the combination of Tesseract and Google Vision.	

Table 4.4 illustrates a comparison of the results of this study with previous research. Previous research, without considering accuracy, recorded CER and WER for text detection of 2.60% and 25.20%. In an effort to achieve high levels of accuracy, as well as lower CER and WER, this study shows that utilizing Tesseract and Google Vision together for OCR can reduce WER values and increase accuracy up to 96.45%. On the other hand, previous research only achieved an accuracy of 46.9% in identifying handwriting. However, the results of this research show that combining Tesseract with Google Vision is the most accurate OCR solution and provides significant progress compared to previous methods.

It is important to note that the collaboration of the two OCR methods not only results in a higher level of accuracy, but also provides significant progress compared to previous studies. These final results confirm that the combined approach between Tesseract and Google Vision is not only effective in reducing WER values, but also creates a superior OCR solution in recognizing handwriting, overcoming the obstacles faced by previous methods.

IV. Conclusion

The program uses Tesseract OCR, Google Vision API, and combination approaches to recognise text on images. The study of the program's output results concludes that the combined method better recognises text on photos. While both Google Vision API and Tesseract OCR have reasonable accuracy rates, combining the two results in a far lower mistake rate regarding text recognition.

Compared to individual approaches, the combined method successfully demonstrated lower average CER (Character Error Rate) and WER (Word Error Rate). This demonstrates how text identification accuracy in photos can be increased using a method that combines the outcomes of Tesseract OCR with Google Vision. Therefore, using integrated approaches can be a more efficient and best option for applications that need text recognition with a low mistake rate. This conclusion demonstrates that combining different techniques can yield more dependable and accurate results when it comes to text recognition in photos.

References

[1]. Arianto, R. F., Rahman, A. Y., & Marisa, F. (2023). *Text Recognition For Socioeconomic Data Survey Sheet Using Ocr Tesseract*.

[2]. Berg, S. A., So, R. H. Y., & Seo, S. Y. (2019). Application Of Optical Character Recognition With Tesseract In Logistics Management. *International Journal Of Internet Manufacturing And Services*, *6*(3), Article 3. Https://Doi.Org/10.1504/Ijims.2019.10022461

[3]. Chesley, E., Marcantonio, J., & Pearson, A. (2019). Towards Syriac Digital Corpora: Evaluation Of Tesseract 4.0 For Syriac Ocr. *Hugoye: Journal Of Syriac Studies*, *22*(1), Article 1. Https://Doi.Org/10.31826/Hug-2019-220105

[4]. González, G., & Evans, C. L. (2019). Biomedical Image Processing With Containers And Deep Learning: An Automated Analysis Pipeline: Data Architecture, Artificial Intelligence, Automated Processing, Containerization, And Clusters Orchestration Ease The Transition From Data

Acquisition To Insights In Medium-To-Large Datasets. *Bioessays*, *41*(6), Article 6. Https://Doi.Org/10.1002/Bies.201900004

[5]. Haji, C. M. (2022). Linguistic Analysis On Cursive Characters. *The Journal Of Duhok University*, *25*(2), Article 2. Https://Doi.Org/10.26682/Sjuod.2022.25.2.3

[6]. Huang, J., Pang, G., Kovvuri, R., Toh, M., Liang, K. J., Krishnan, P., Yin, X., & Hassner, T. (2021). A Multiplexed Network For End-To-End, Multilingual Ocr. *2021 Ieee/Cvf Conference On Computer Vision And Pattern Recognition (Cvpr)*, 4545–4555. Https://Doi.Org/10.1109/Cvpr46437.2021.00452

[7]. Hukkeri, G. S., Goudar, R. H., Janagond, P., & Patil, P. S. (2022). Machine Learning In Ocr Technology: Performance Analysis Of Different Ocr Methods For Slide-To-Text Conversion In Lecture Videos. *International Journal Of Advanced Computer Science And Applications*, *13*(8), Article 8. Https://Doi.Org/10.14569/Ijacsa.2022.0130839

[8]. Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (Ocr): A Comprehensive Systematic Literature Review (Slr). *Ieee Access*, *8*, 142642–142668. Https://Doi.Org/10.1109/Access.2020.3012542

[9]. Muharom, S. (2019). Pengenalan Nomor Ruangan Menggunakan Kamera Berbasis Ocr Dan Template Matching. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *4*(1), Article 1. Https://Doi.Org/10.25139/Inform.V4i1.1371

[10]. Mursari, L. R., & Wibowo, A. (2021). The Effectiveness Of Image Preprocessing On Digital Handwritten Scripts Recognition With The Implementation Of Ocr Tesseract. *Computer Engineering And Applications Journal*, *10*(3), Article 3. Https://Doi.Org/10.18495/Comengapp.V10i3.386

[11]. Putri, M. H., & Yuhan, R. J. (2020). Indeks Kerawanan Sosial Ekonomi Perempuan Indonesia Tahun 2017. *Seminar Nasional Official Statistics*, *2019*(1), Article 1. Https://Doi.Org/10.34123/Semnasoffstat.V2019i1.117

[12]. Rohman, M. A. A., & Djasuli, M. (2022). *Penerapan Good Corporate Governance Tranparansi Terhadap Kinerja Surveyor Registrasi Sosial Ekonomi Dalam Mewujudkan Data Akurat*.

[13]. Smith, R., Newton, C., & Cheatle, P. (N.D.). *Adaptive Thresholding For Ocr: A Significant Test*.

[14]. Suharto, E. (2015). Peran Perlindungan Sosial Dalam Mengatasi Kemiskinan Di Indonesia: Studi Kasus Program Keluarga Harapan. *Sosiohumaniora*, *17*(1), Article 1. Https://Doi.Org/10.24198/Sosiohumaniora.V17i1.5668

[15]. Wibawa, C., & Anggraeni, D. T. (2023). Comparison Of Image Segmentation Method In Image Character Extraction Preprocessing Using Optical Character Recoginiton. *Jurnal Teknik Informatika (Jutif)*, *4*(3), 583–589. Https://Doi.Org/10.52436/1.Jutif.2023.4.3.956