Jurnal Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v13i1.865
Vol. 13, No. 1, March 2023, pp. 1-7

Image Processing Technology in Book Metadata Extraction System Using Optical Character Recognition (OCR)

Andi Emil Multazam[1], Akhmad Qashlim[2] & Muhammad Sarjan [3]

[1,2,3] Al Asyariah Mandar University, Polewali Mandar, Indonesia, 91311
E-mail: [1]andiemilmultazam2022@gmail.com, [2]muhsarjan2018@gmail.com, [3]qashlim@main.unasman.ac.id

ARTICLE HISTORY

Received : February 4th, 2023
Revised : March 21st, 2023
Accepted : March 30th, 2023

KEYWORDS

Optical Character Recognition
Book Metadata
Image Processing
Extraction System

ABSTRACT

Extracting book metadata by retyping the identity of the book, such as the author's name, book title, publisher, and several other identities, is a routine that is carried out repeatedly at the Polewali Mandar district library, this activity takes much of time, using several staff and it turns out that this activity has much potential for input errors. Errors in extracting book metadata will result in errors in the book repository system database, resulting in difficulty finding and using books or book data information. This problem can be solved by creating a book metadata extraction system using image processing technology and OCR. This study aims to design a scanner technology to extract book metadata. Accuracy is carried out in 2 stages, the first validation of image extraction results using the ROC method and the second validation by directly matching the result of extracting the book's metadata with the actual book. The results of this study indicate that the system has worked with an accuracy of 98.78% with an average detection time of 1.49 seconds and has succeeded in presenting the extraction results on the website page. Thus the metadata extraction system with the OCR method can be applied to libraries to input book data.

1. Introduction

Book metadata is structured information that can be used or reorganized into useful knowledge for readers and library managers, including as a representation of a book, so that sources of book information are easy to find, classifying book data based on the same source or the same category [1]. Extracting book metadata can be done in many ways, such as scanning documents and saving them as images. However, problems then arise, namely these images cannot be edited and it is very difficult to find out what information is in the images [2] Using layoutLM to convert electronic files into text information without changing the document layout and this method successfully incorporates visual information into the word layoutLM [3]. The most conventional thing is to do conventional data entry and this work is done repeatedly until finally another problem arises, namely an error in entering data that can occur from human negligence *(human error)* including when transferring the image to text form [4] besides that it requires a lot of time and human effort. Errors in extracting book metadata will result in errors in the book repository system database and will ultimately result in difficulties finding and using books or book data information.

Book metadata input activities can be seen in the routines of the Regional Library of Polewali Mandar Regency. Based on table 1 data, it can be seen the number of books that must be inputted from 2019, 2020, 2021 and the large number of data entry officers involved. To minimize errors when entering book metadata such as book title, author, book publisher, one of the efforts made is to use image processing technology and the OCR method. [5]. OCR can help convert handwritten, typed, or printed text in images into text that can later be edited, searched, and used for further processing [6] Thus the aim of this research is to design a book metadata extraction system using the OCR method and automatically become important and must be implemented. The system built has image processing capabilities to be used to process the original image and produce other images that have information according to temporary OCR needs in which there is an OCR tesseract architecture that functions as a character database [7] will be used to read the image form into letter characters [8]and understand the written text contained in the image [9].

2. General Instructions

2.1 Polewali Mandar District Library

Regional Library of Polewali Mandar Regency which is a reading center for both students and the general public, various interests such as finding reference sources, research from lecturers or

academics or just reading for employees of the Regional Library Service office. The large number of books procured every year or donations from various groups means that many of these books are just piled up and not recorded, they don't even have a repository system. The data on the number of books in the Polewali Mandar district library are presented in table 1.

Table 1. Data Processing and Preservation of Library Materials of the Regional Library and Archives Office of Polewali Mandar Regency.

Sub Elements	Number of Books in Year		
	2019	2020	2021
mobile library	2,667	2,687	2,695
Regional Public Library	22,841	23,571	23,909
Total Collection of Book Titles	8,057	8,416	8,454

Source: Polman Satu Data
http://satudata.polmankab.go.id/dataku

Metadata in books as in table 1 is needed and will be difficult to find if it is not properly present and grouped in an orderly manner, image processing technology and algorithms in OCR can be used to complete and assist the process of inputting book metadata into the repository system.

2.2. Image Processing and OCR

Image document analysis is part of an important component of an intelligent system based on the application of pattern recognition techniques with three conventional channels in images, namely Red Green Blue (RGB) [10]. In addition, the document analysis method is also based on digital image processing, where the processing stage begins by converting the document image into a binary image to facilitate letter recognition (Yamin et al., 2022) . A study to improve document analysis performance using Hyperspectral Imaging systems (HSI) which is not only capable of identifying text in images but also capable of extracting signatures and distinguishing ink or scanner use. Object images are taken using a camera device then identify images using RGB or grayscale scales and perform image processing to obtain patterns with the stages of extraction, classification, analysis, and feature recognition [10]as well as in research processing braille document images by doing grayscaling to turn a 3 channel (RGB) image into a 1 channel image with an intensity value that represents the gray level with the crop image stages, image conversion to grayscale, threshold, dilation and erosion, this process is carried out to remove unnecessary parts of the image then discarded [11]. To deal with the problem of irregular documents, layout analysis is needed.

As an important step in extracting information contained in image documents using the OCR algorithm. Extraction of information from scanned documents is highly dependent on the position of fields in the document, OCR algorithms can handle administrative documents with complex layouts [12].

Metadata extraction of books in the regional library of Polewali Mandar Regency can be done by scanning the cover to produce an image which is then processed by image processing to increase the RGB histogram value then extracted into text form using the OCR algorithm, the extract results are entered in the database to serve as a book repository system.

2.3 Extraction System

The process and stages in performing image detection and extraction for conversion to text are as follows:

1. Take pictures from the camera or upload pictures to the system.
2. Transform RGB image into grayscale image. Grayscaling is the process of converting a color image into an image that only has degrees of gray. The program will repeat the per-pixel grayscaling of the image length and width, convert RGB images to grayscale images [13].

$$S = \frac{R+G+B}{3} \qquad (1)$$

Where :
S= new pixel value in grayscale image
R = the red value of the pixel in the RGB image.
G = green value at the pixel in the RGB image.
B = the blue value of the pixel in the RGB image.

3. The value generated from the equation above will be inputted into each element of the basic color of the image *grayscale* .
4. For example, a pixel has RGB components as follows:

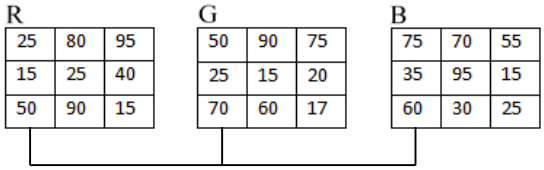

Figure 1. Conversion of RGB images to Grayscale images

R = 50
G = 70
B = 60
S = 50+70+60/3 = 60

To change the pixel to grayscale is to find the average value above so that the pixel value becomes

S = 50 + 70 + 60/3 = 60.

5. Applying edge detection to grayscale images using the Threshold method Thresholding is a process of separating the background from the object you want to observe by changing the image to black and white. At this stage thresholding can convert color or grayscale images into binary images by changing each pixel within a certain range. Meanwhile, a binary image is a digital image with two possible pixel values, namely object pixels worth 1 (white) and background pixels worth 0 (black). Carry out the thresholding process to get a binary image.

$$dst(x,y) = \begin{cases} maxValue & if scr(x,y) > T(x,y) \\ 0 & otherwise \end{cases} \quad (2)$$

Where :
Dst(x,y) = new pixel value in binary image.
maxValue = maximum value in the image if the pixel is greater than the threshold value.
Src(x,y) = pixel value in a grayscale image

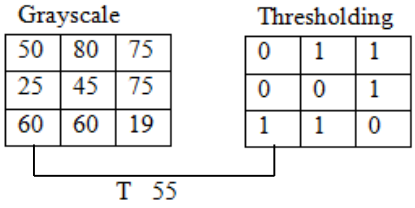

Figure 2. Grayscale image conversion to thresholding

If the grayscale pixel value >= threshold then the binary pixel value = 1
If the grayscale pixel value <threshold then the binary pixel value = 0
Below is an image showing the Image Preprocessing stage on a text image

Figure 3. Image processing stage

6. Segmentation After the image preprocessing stage is complete, the next step is segmentation. Process for detecting connected character components. In this process, Tesseract searches along the image and then identifies the foreground or outline pixels, the segmentation process will continue until all the outermost pixels are marked as character outlines. Outline of each character will be collected into blobs. The outline search process can be seen in Figure 2.2.

Figure 4. The online search process

Carry out the image segmentation process using the Connected component labeling method to obtain connected character components. Initial labeling. Scan the image pixel by pixel from left to right and top to bottom. Let p denote the current pixel in the scanning process and 4-nbr denote the four neighboring pixels in the N, NW, NE and W directions of p. If p is 0, proceed to the next scanning position. If p is 1 and all values in 4-nbrs are 0, assign a new label to the page. If only one value in 4-nbrs is not 0, assign the value to page. If two or more values in 4-nbrs are not 0, assign one of the labels to p and mark the labels in 4-nbrs as equivalent. This process will continue until all the outermost pixels are marked as character outlines.

7. Word Recognition. After the segmentation stage has been completed, the next stage is word recognition. Word recognition is a character recognition process. Tesseract performs topological feature extraction as is done by comparing the features obtained with the features that already exist in the OCR database in Tesseract [6]

Figure 5. Character recognition results

This process is carried out to find characters in an image that has gone through the preprocessing and segmentation stages. The word recognition process is the same as the line search process, it's just that the pixel points found will be compared with the OCR database, so if the pixel points have the same pattern as the pixels in the database, then the object will be taken as a character. According to research conducted by

8. Character Introduction. Tesseract OCR Engine is an open source character recognition engine that can be used by various operating systems. The Tesseract OCR Engine gives good results for each character with an average error rate of only 3.77% or in other words 96.33% is able to recognize characters properly (Hanny Rindiani, 2015.
9. Displays the results of converting images to text on the interface that has been provided.

3. Method

3.1. Research Stage

The book metadata extraction system uses image segmentation techniques using the connected component labeling method to recognize parts of the letters [14], after which topological feature extraction is performed using the OCR method to convert them into text form. In Figure 3.1, there are research stages showing the activities of each stage. It can be seen

that in the first stage, data identification is carried out. This is done by identifying problems and solving methods that will be used in this study. After that, the data requirements and types of data that will be managed will be described. The first stage of the achievement is that a metadata table design will be made using the MySQL database software.

Furthermore, in the second stage, a system block will be designed by identifying the process of extracting images into text form, after which the process will be tested to see the accuracy of the results and validate the process carried out. The results of this stage are a system block diagram which can be seen in 3.2. .

The third stage is the design of the application system, namely the mobile application which will be built according to the block diagram and the website using the database according to the metadata table in the first stage. The extraction results will be presented on a web page and at this stage the application program will be implemented.

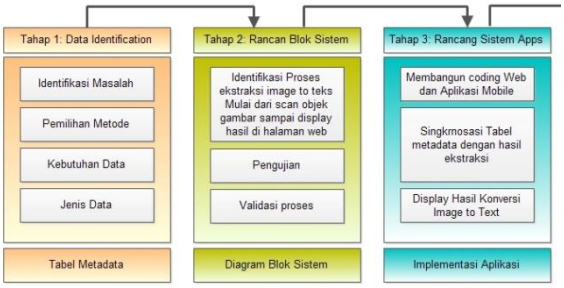

Figure 6. Research Stages

3.2. Research data

The need for research data is adjusted to the design of the application system to be made. Namely the book metadata structure that contains the name of the author, publisher, book title, year of publication. This structure is also adapted to the design of mySQL database tables.

Image files in JPEG format are identified data which are then extracted and used as input data to be entered into the website system.

3.3. System block diagram design

The workings of the application system that was created as a whole can be seen in Figure 3.2. A block diagram of the system that is designed starts working when the image object is successfully obtained, in this case, the book cover is scanned after that preprocessing will occur, namely identifying whether the image can be read or not after that the image processing process is image extraction into text form and sending the text to a database to be presented on a web page. The image capture process is completed with a smartphone device and ultimately presented on a web repository page.

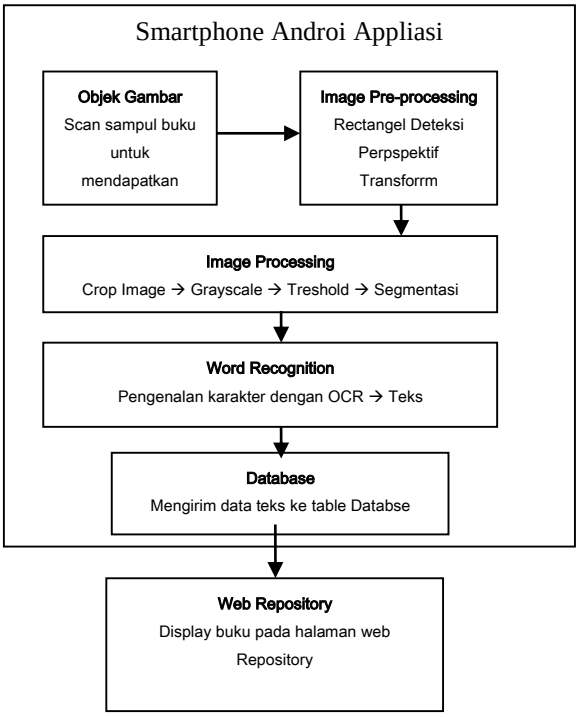

Figure 7. System Block Diagram

4. Result and Discussion

4.1. Result

A metadata extraction system that uses *image processing* technology to identify book images obtained through smartphone cameras and then extracted using *Optical Character Recognition* (OCR) to obtain results in text form has been completed and tested at the Library and Archives Service. Polewali Mandar Regency. The designed application system interface can be seen in Figure 4.1.

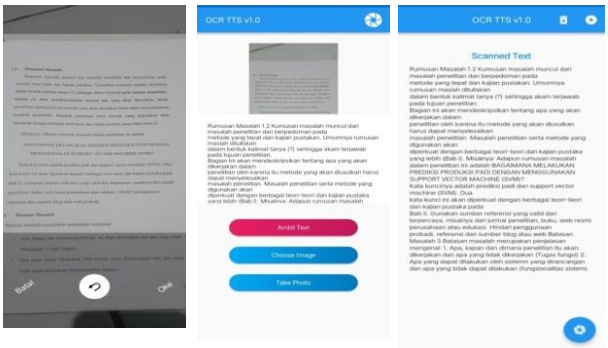

Figure 8. Image to Text Conversion Application System

Part (a) in figure 4.1. is an image file in JPG format which will be extracted and converted into text form using an application designed, the conversion process can be seen in part (b), which also has a text capture feature to extract text in images, or the *choose*

feature image to retrieve image files from the gallery and take photos to retrieve image files from the smartphone camera. while the results of the extract can be seen in part (c). The quality of the extract results will be determined from the quality of the detected image file. The clearer the text in the image file, the clearer, orderly and structured the resulting extract will be. The accuracy or compatibility of the extract results with the image file can reach a value of 80%.

4.2. Discussion

Taking image objects through a smartphone camera is the first process carried out to obtain input data, the image object taken is a book cover image. Images that have been successfully identified at the *image pre-processing stage will be processed up to image* segmentation which is intended to eliminate *noise* in the image so that the *word recognition process* can provide maximum text conversion results. The testing process at this stage is carried out by taking pictures at various angles, namely 90º, 80º, and 60º, this process is carried out to see the best shooting position as presented in Figure 4.2.

4.2.1. Testing Based on Camera Position

The first test was carried out at a camera angle of 90 º. In this process the sample using the letter format is taken right above the letter with good lighting as can be seen in Figure 4.2. while the results of the conversion to text can be seen in Figure 4.3.

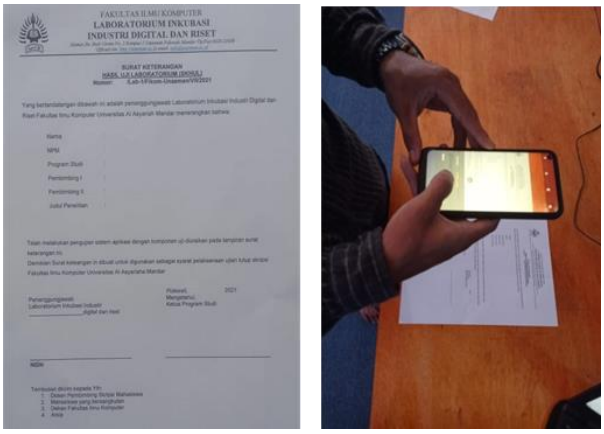

Figure 9. Image Capture Process at 90º position

Figure 10. Image to text conversion results

Figure 4.3. shows the results of converting image objects to the maximum text format. It can be seen that the text reads well even though it is not arranged according to the format but the purpose of the application for text conversion has been achieved. While shooting at an angle of 80º can be seen in Figure 4.4. and at an angle of 60º can be seen in Figure 4.5

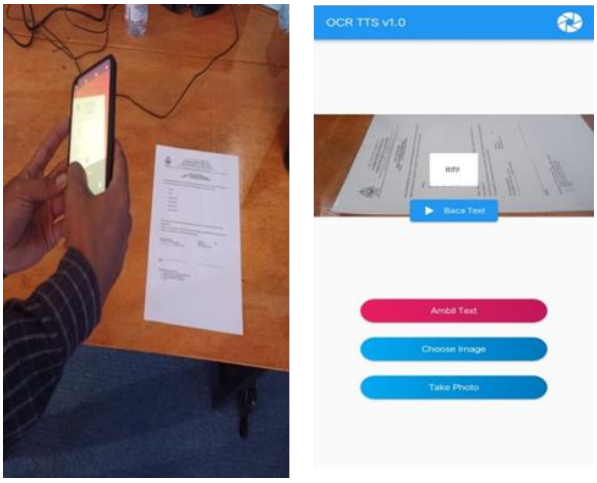

Figure 11. Shooting at 80º Position

Taking pictures at an angle of 80° as shown in Figure 4.3. provides inaccurate detection results, even only able to recognize a few letters, this indicates that the application has not been able to detect images to the fullest. While at an angle of 60º can be seen in Figure 4.4.

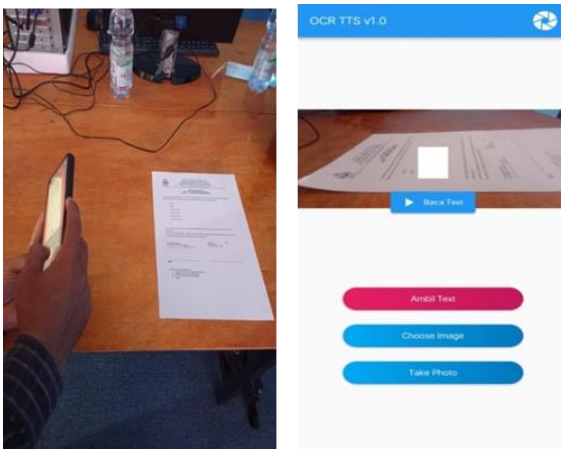

Figure 12. Image Capture Process
At Position 60º

At this capture angle, the app cannot recognize the text and returns no conversion results.

4.2.2. Character Recognition Accuracy

Measurement of the accuracy of character recognition is carried out with the aim of knowing the level of application capability and methods for recognizing letter characters. To measure accuracy, use the following formula (3) [15]

$$Accuracy\ (\%) = \frac{Jumlah\ Karakter\ Benar}{Jumlah\ Karakter\ Keseluruhan} x\ 100\% \quad (3)$$

The number of characters in the letter sample used for accuracy testing in this study is 526, which is the total number of characters. To find out the correct number of characters is to look at the number of characters that were successfully recognized according to the letter format. Accuracy level as follows:

$$Accuracy\ (\%) = \frac{518}{526} x\ 100\%$$
$$= 98.74\%$$

While the average processing time starting from pre-processing, image processing and word recognition = 7.464 $detik$ / 5 = 1.49 seconds.

5. Conclusion

Based on the test results of taking pictures at each different angle, the best position to get the maximum image to text conversion results is at a position of 90º with an accuracy of 98.78%. With a detection time of 1.49 seconds. Given the importance of book metadata for readers and library managers, among others, as a representation of a book, which can help find sources of information easily by identifying book sources. Then book data extraction errors can be solved using the OCR method. Even though the system has succeeded in detecting images using *image processing technology* and converting them to text using the OCR method, the designed application system still has limitations such as not storing detected objects in the database as history, the application system has not been able to tidy up text, image quality that can be converted to Text is very dependent on lighting and camera position, at least the operating system that can run this application is the Android 05 operating system. Further research can be carried out with better methods and accuracy

6. Acknowledgement

Thanks are given to the Ministry of Education and Culture for the 2022 research grant support, thanks to the regional library of Polewali Mandar district as the location for testing the system, thanks also to the Research Laboratory of the Faculty of Computer Science and the research team for their assistance while this research was being carried out.

References

[1] D. D. Prayoga, "Automasi Ekstraksi Metadata Buku," 2020. [Online]. Available: https://repository.uir.ac.id/3731/

[2] Y. Huang, H. Xue, B. Liu, and Y. Lu, *Unifying Multimodal Transformer for Bi-directional Image and Text Generation*, vol. 1, no. 1. Association for Computing Machinery, 2021. doi: 10.1145/3474085.3481540.

[3] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1192–1200, 2020, doi: 10.1145/3394486.3403172.

[4] M. Levy and A. Subburaj, "Emerging Trends in Data Center Management Automation," *2021 IEEE 11th Annu. Comput. Commun. Work. Conf. CCWC 2021*, pp. 480–485, 2021, doi: 10.1109/CCWC51732.2021.9375837.

[5] N. Pawar, Z. Shaikh, P. Shinde, and Y. P. Warke, "Image to Text Conversion Using Tesseract," *Int. Res. J. Eng. Technol.*, vol. 6, no. 2, pp. 516–519, 2019.

[6] A. Sophia, "A Deep Learning Technique For Optical Word Recognition For Indic Script," *Int. J. Eng. Innov. Adv. Technol.*, vol. 3, no. 1, pp. 34–38, 2021, [Online]. Available: https://www.ijeiat.com/images/sliders/72d4c21cc02cc98c08f3c2d556cd70af.pdf

[7] R. R. Santoso, "Prototype Sistem Deteksi Plat Nomor Kendaraan Pada Sistem Perparkiran Berbasis Image Processing Menggunakan Metode Optical Character Recognition (Ocr)," *Digit. Repos. Univ. Jember Digit. Repos. Univ. Jember*, vol. 1, 2020.

[8] S. Hartanto, A. Sugiharto, and S. N. Endah, “Optical Character Recognition Menggunakan Algoritma Template Matching Correlation,” *J. Informatics Technol.*, vol. 1, no. 1, pp. 11–20, 2012, doi: 10.3810/pgm.1998.03.410.

[9] J. Wang, J. Tang, and J. Luo, “Multimodal Attention with Image Text Spatial Relationship for OCR-Based Image Captioning,” *MM 2020 - Proc. 28th ACM Int. Conf. Multimed.*, pp. 4337–4345, 2020, doi: 10.1145/3394171.3413753.

[10] R. Qureshi, M. Uzair, K. Khurshid, and H. Yan, “Hyperspectral document image processing: Applications, challenges and future prospects,” *Pattern Recognit.*, vol. 90, pp. 12–22, 2019, doi: 10.1016/j.patcog.2019.01.026.

[11] R. Bin Yamin, H. Kusuma, and Tasripan, “Image processing-based braille document copying system as additional feature for ITS’s braille printer,” *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 230–237, 2022, doi: 10.1016/j.procs.2021.12.136.

[12] S. Tomovic, K. Pavlovic, and M. Bajceta, “Aligning document layouts extracted with different OCR engines with clustering approach,” *Egypt. Informatics J.*, vol. 22, no. 3, pp. 329–338, 2021, doi: 10.1016/j.eij.2020.12.004.

[13] J. F. Fauzi, H. Tolle, and R. K. Dewi, “Implementasi Metode RGB To HSV pada Aplikasi Pengenalan Mata Uang Kertas Berbasis Android untuk Tuna Netra,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 6. pp. 2319–2325, 2018.

[14] L. He, Y. Chao, K. Suzuki, and K. Wu, “Fast connected-component labeling,” *Pattern Recognit.*, vol. 42, no. 9, pp. 1977–1987, 2009, doi: 10.1016/j.patcog.2008.10.013.

[15] S. S. Nurhaliza and L. ETP, “Sistem Pengenalan Karakter Dokumen Secara Otomatis Menggunakan Metode Optical Character Recognition,” *Petir*, vol. 15, no. 1, pp. 166–175, 2022, doi: 10.33322/petir.v15i1.1610.