DOI: https://doi.org/10.52436/1.jutif.2023.4.3.956
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON OF IMAGE SEGMENTATION METHOD IN IMAGE CHARACTER EXTRACTION PREPROCESSING USING OPTICAL CHARACTER RECOGINITON

Condro Wibawa*[1], Dessy Tri Anggraeni[2]

[1,2]Faculty of Computer Science, Universitas Gunadarma, Indonesia
Email: [1]condro_wibawa@staff.gunadarma.ac.id, [2]dessytri@staff.gunadarma.ac.id

(Article received: March 17, 2023; Revision: March 31, 2023; published: June 26, 2023)

Abstract

Today, there are many documents in the form of digital images obtained from various sources which must be able to be processed by a computer automatically. One of the document image processing is text feature extraction using OCR (Optical Character Recognition) technology. However, in many cases OCR technology are unable to read text characters in digital images accurately. This could be due to several factor such as poor image quality or noise. In order to get accurate result, the image must be in a good quality, so that digital image need to be preprocessed. The image preprocessing method used in this study are Otsu Thressholding Binarization, Niblack, and Sauvola methods. While the OCR technology used to extract the character is Tesseract library in Python. The test results show that direct text extraction from the original image gives better results with a character match rate average of 77.27%. Meanwhile, the match rate using the Otsu Thressholding method was 70.27%, the Sauvola method was 69.67%, and the Niblack method was only 35.72%. However, in some cases in this research the Sauvola and Otsu methods give better results.

Keywords: *Image Processing, Niblack, Optical Character Recoginition, Otsu Thressholding, Sauvola*.

1. INTRODUCTION

In this digital era there are many digital images around us especially in social media. These digital images comes from camera photos, scanned documents/images, or those that are produced using image processing applications. Meanwhile, Munir explained that digital images are images obtained from the digitization process of two-dimensional images [1]. Putra explained that a digital image is an array containing values represented by a certain row of bits [2]. Today, many form of digital image used in our daily activities.

In the era of the industrial revolution 4.0, images are not only used as an aesthetic products, but also able to be processed using a computer. One form of image processing is image feature extraction. Image feature extraction is a technique used to obtain features in images for the process of classification and image recognition [3]. One of the features is text character extraction. Text character extraction from images can be defined as the work of extracting text objects from a set of images. Text extraction is a challenging task because there are variations of text size, font, style, orientation and alignment to a complex background [4]. Text character extraction were separated into Handwriten Character Recognition (HCR) and Optical Character Recognition (OCR) [5]. OCR itself is an electronic conversion method of digital text in the form of images into computerized text [6]. Image character extraction can be implemented in many fields, such as for converting scanned document images/photos into text form [7], automatic reading of license plates [3], automatic reading of business cards [8], automatic reading of receipt and invoice documents [9], [10], character captcha recoginition [11], and others. Text character extraction seem an easy task for humans, but nevertheless it was complex problem for computers [12]. Today, there are many OCR technologies can be used like Google Vission, PhotoScan in Windows Operating System, Tesseract, etc.

However, OCR technology is often unable to read text characters accurately. The inability of OCR technology to extract characters from digital images can be caused by several factors, such as low image quality, noises, different font size and types, etc [13]. An image that has too much noise (interference), low lighting quality, etc will make it difficult for OCR technology to read the digital image. Thus, it is necessary to carry out a digital image preprocessing process before applying OCR technology to the image. In his research, Anh concluded that image preprocessing improves text extraction on OCR technology [14]. By using Tesseract, Ahn succeed to increase text character extraction significantly. In other hand, Brisinello in his research claim that image preprocessing used increase OCR regonition using Tesseract by 33,3% [15].

In this study, a comparison of several preprocessing methods on digital images will be discussed. The method used are the Otsu Thressholding, Niblack, and Sauvola methods. The choice of this method is based on the previous studies, where these three methods are widely used to improve document images quality. In research [6] and [16] Otsu method used to improves document quality with a better output result. In another research [7] and [17], Niblack and Sauvola method used to improve old document to get better image quality. Furthermore, in the testing process, the preprocessed image will be extracted using the Tesseract library. Tesseract is an open source library used for text character recognition in digital images and can be used in the Python programming language. Tesseract itself claims that this library using Long Short Term Memory (LSTM) for the image preprocessing [18].

Through this research it is expected to be able to provide an overview of digital image preprocessing methods with the best match rate level when used in OCR technology. In this study, tests were also carried out on several types of image such as vehicle number plates, identity cards, document fragments, full documents, and handwriting documents. These is done to provide variations in the type and quality of the images to be extracted. So that it can provide real result.

2. METHOD

The research method is the sequence of steps carried out in the research process. The stages process in this study can be seen in Figure 1.

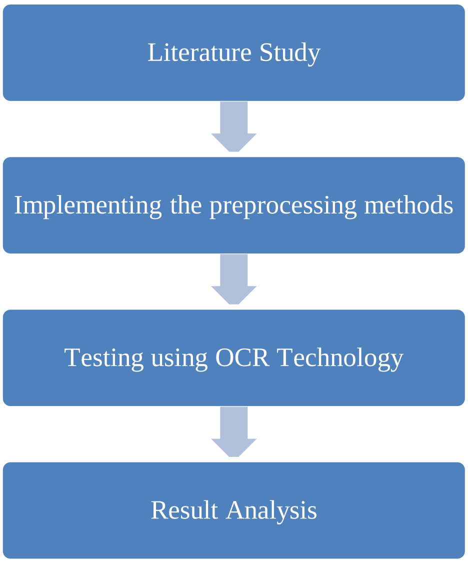

Literature Study
Implementing the preprocessing methods
Testing using OCR Technology
Result Analysis

Fig. 1 Methodology Process.

2.1. Literature Study

A literature study was conducted to explore the methods used in this research.

2.2. Implementing the Preprocessing Methods

The image preprocessing process, especially the segmentation method, has the same steps for any method. It can be seen in Figure 2.

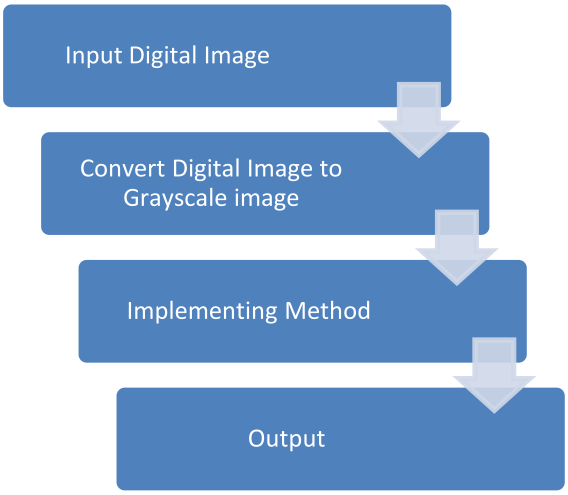


Fig. 2 Flowchart Program for Image Preprocessing Process

The first step in this program is to input the image to be processed. In this study, the image used is a digital image scanned or taken by digital camera in the format of .JPG or .JPEG. The file size is limited to a maximum of 100 Kb so that the image processing does not take too long. The next step is to convert the image from a color image (RGB) to a grayscale image. If the image is already in grayscale format, no conversion is needed. Next, the image is ready to be processed using the Otsu Thresholding, Niblack, and Sauvola methods. The result of image processing is another digital image with the format of .JPG. All this step will be done using Python programming language and the OpenCV library.

2.3. Testing Using OCR Technology

The tests were carried out by extracting the original digital image, the output image from the Otsu Thresholding, Niblack, and Sauvola methods with OCR technology. To carry out this test, a simple application was created with the Python programming language and the Tesseract library. The Tesseract library itself already has a preprocessing process using Adaptive Binarization Long Short-Term Memory (LSTM) [18]. The number of characters that were successfully read using OCR technology is then compared with the original characters. The percentage match rate can be calculated using the following formula.

$$Match\ Rate = \left(\frac{n}{s}\right) * 100\% \qquad (1)$$

Description :
n : Number of Match Extracted Character
s : Number of Real Character

The number of images tested were 15 images obtained from scanned documents/digital camera photos. The test results and the calculation of the match rate value will be stored in a table so that it can be analyzed further.

2.4. Result Analysis

The final step in this study is the results analysis. The results of research that has been done before are analyzed and determined which method has the highest match rate. The final match rate is averaged for each method. The method with the highest average match rate value is the method with the best match rate.

3. RESULT

3.1. Literature Study

Literature review was conducted to explore the methods used in this research, including the Otsu Thresholding, Niblack, and Sauvola methods.

Otsu Thressholding Method

The Otsu thressholding method is a method that automatixally divide the gray image histogram into two different areas without entering a threshold value. The otsu thressholding method works is by applying discriminant analysis which determines a variable so that it can distinguish between two or more naturally occurring groups. The discriminant analysis carried out was able to maximize the separation of objects and backgrounds [19]. The Otsu method can detect a digital image even though it has a high noise level [20].

Niblack Method

The Niblack method is an image segmentation method by setting a threshold value based on the average of the neighboring values, summed by the size of the neighboring area (window) multiplied by the standard deviation of the neighboring values. This method is widely used to process scanned text documents. The equation for Niblack's method can be seen in the following formula [17], [21].

$$Th(x,y) = \mu(x,y) + k.\sigma(x,y) \qquad (2)$$

Description:
$Th(x,y)$: pixel threshold value at f(x,y)
$\mu(x,y)$: average pixel of neighboring area at f(x,y)
k : size of neighboring area
σ(x, y) : standart deviation of neighboring area at f(x,y)

Sauvola Method

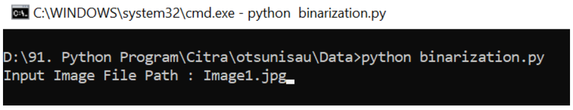
The Sauvola method is a segmentation method which is a development of the Niblack method. Both of these methods use the concept of local adaptive threshold. The difference is that in the Sauvola method there is an adjustment to the local thresholding value [22]. This method also widely used to process scanned text documents. Some of the research use this method to get better result of old document. The equation for Sauvola’s method can be seen in the following formula.

$$Th(x,y) = \mu(x,y)\left[1 + k\left(\frac{\sigma(x,y)}{R}\right) - 1\right] \qquad (3)$$

Description:
$Th(x,y)$: pixel threshold value at f(x,y)
$\mu(x,y)$: average pixel of neighboring area at f(x,y)
k : size of neighboring area
σ(x, y) : standart deviation of neighboring area at f(x,y)
R : constant value for grayscale image (128)

3.2. Implementing Method

The results of this study are in the form of a program that can be used to perform image segmentation and image character extraction. The program is made with the Python programming language. On the main program, the user will be asked to enter the path image. The program display is shown in Figure 3.



```
D:\91. Python Program\Citra\otsunisau\Data>python binarization.py
Input Image File Path : Image1.jpg_
```

Fig. 3 Python Program for Segmentation and Character Extraction from Image

Then, the program will carry out the image segmentation process using the Otsu Thresholding, Niblack, and Sauvola methods. Some examples of the results can be seen in Figures 4, 5, and 6.

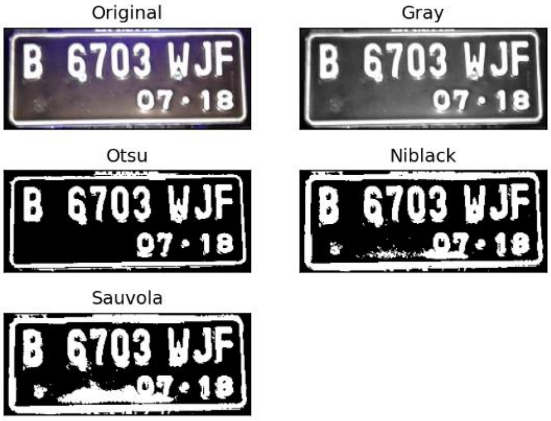


Fig. 4 Image Segmentation of Vehicle License Plates

In this result, Otsu's method seems to have a better result. All characters are readable. Whereas in the Niblack and Sauvola methods there is a white area at the bottom side which interferes reading character of 07.

In line with the results in figure 4, the result in figure 5 also produces a similar output. All characters in the Otsu method are clearly readable and have no noise. Whereas in the Niblack method, even though the characters are readable,but there is some noise. In the Sauvola method there is no noise, but the characters appear lighter than the Otsu method.

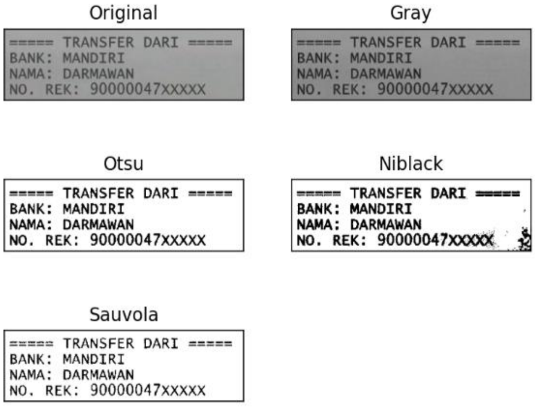


Fig. 5 Image Segmentation of Bank Transfer Receipts

In contrast to the results in figure 4 and figure 5, in figure 6, the Sauvola method appears to give better results. Characters are readable and there is little noise. In Otsu's method, the background is too dark and make characters difficult to read. Likewise in the Niblack method, besides the background being too dark, there is also a lot of noise.

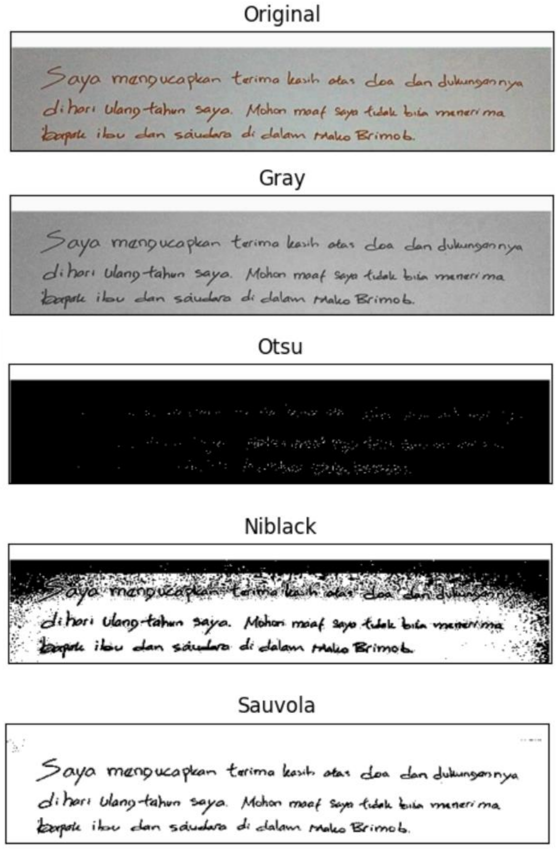

Original
Gray
Otsu
Niblack
Sauvola
Saya mengucapkan terima kasih atas doa dan dukungannya
dihari Ulang-tahun saya. Mohon maaf saya tidak bisa menerima
bapak ibu dan saudara di dalam Mako Brimob.

Fig. 6 Image Segmentation of Handwriting Scan Documents.

3.3. Testing Using OCR Technology

The program set to extract the character from the omage using Tesseract library. The program then will display the results of text extraction. As an example, the extraction results from Figure 4 are summarized in table 1.

Table 1. Character Extraction Result for Figure 4

Process	Description	Result
Original Text	Text Read Manually	B 6703 WJF 07.18
OCR 1	Character Extraction of Original Image	B 4703 WUF 9718
OCR 2	Character Extraction of Image After Processing with The Otsu Method	
OCR 3	Character Extraction of Image After Processing with The Niblack Method	-
OCR 4	Character Extraction of Image After Processing with The Sauvola Method	-

These results indicate that using OCR directly on the original image gives better results than after the image is processed, with a character match rate of: (9/13)*100% = 69%. Whereas in the OCR 2, 3, and 4 processes where the images were processed first using the Otsu, Niblack, and Sauvola methods, they did not give any results, so the match rate was 0%.

Different results are shown from the results of using OCR in Figure 6. The extraction results in Figure 6 can be seen in table 2.

Table 2. Character Extraction Result for Figure 6

Process	Description	Result
Original Text	Text Read Manually	walaupun banyak negeri kujalani yang mahsyur permai dikata orang Tetapi kampung dan rumahku Disanalah kurasa senang Tanaku tak kulupakan engkau kubanggakan
OCR 1	Character Extraction of Original Image	a 2 i Relish pee oe
OCR 2	Character Extraction of Image After Processing with The Otsu Method	-
OCR 3	Character Extraction of Image After Processing with The Niblack Method	-
OCR 4	Character Extraction of Image After Processing with The Sauvola Method	Talavpon+ banal regent Kyjdani gang Mahsyor perma diketa ofang Tetagi Kampung don Mmeabke Disandich tuasa Senang Taneiyu:. tok: owpatean @rskou. htubanggaken:

These results indicate that the use of OCR on images that have been processed using the Sauvola method gives better results compared to the use of the otsu method, the niblack method, or the original image. The Sauvola method provides a character match rate of: (82/135)*100% = 62%. In the OCR 1 process there are extraction results but do not represent the original text. Whereas the OCR 2 and 3 processes did not give any results.

The match rate of characters extracted from text from 15 test images can be seen in table 3.

Table 3. Character Extraction Result for All Images Tested

Image No	Count of Real Text Character	Match Text Count using OCR							
		Original		**Otsu**		**Niblack**		**Sauvola**	
		Match	**%**	**Match**	**%**	**Match**	**%**	**Match**	**%**
Image 1	13	9	**69**	0	-	0	-	0	-
Image 2	12	0	-	4	33	0	-	0	-
Image 3	56	48	**86**	39	70	7	13	46	82
Image 4	63	0	-	0	-	0	-	0	-
Image 5	108	103	**95**	97	90	9	8	99	92
Image 6	120	27	23	43	**36**	42	35	37	31
Image 7	57	45	**79**	44	77	51	**79**	45	**79**
Image 8	51	50	**98**	50	**98**	44	86	49	96
Image 9	69	66	**96**	65	94	51	74	63	91
Image 10	87	56	**64**	51	59	15	17	49	56
Image 11	93	78	**84**	65	70	6	6	58	62
Image 12	77	57	74	61	**79**	13	17	61	**79**
Image 13	118	0	-	0	-	17	14	45	**38**
Image 14	119	98	**82**	80	67	41	34	78	66
Image 15	135	0	-	0	-	0		86	64
Average Match Rate		**77,27 %**		**70,27 %**		**35,72 %**		**69,67 %**	

3.4. Analysis

Based on the data in table 3, it can be concluded that the direct extraction process on the original image without image preprocessing gives the best results with a character match rate of 77.27%. Meanwhile, if the preprocessing process is carried out, then the Otsu Thresholding method gives the best results of 70.27%. This result is slightly different from the Sauvola method which has a character match rate of 69.67%. The Niblack method provides a low compatibility rate of 35.72%.

However, even though the overall extraction of the original images gave the best average results, in some cases the use of the Otsu and sauvola methods gave better results. In image 6 the Otsu method gives the best results with match rate of 36% (compare to 23% for original image, 35% for Niblack, and 31% for Sauvola). Whereas in image 13 the sauvola method gives the best results with match rate of 38% (compare to 0% for original image, 0% for Otsu, and 14% for Niblack). Meanwhile in image 12, the Otsu and Sauvola methods also give the best results with the same level of match rate which is 79% (compare to 74% for original image and 14% for Niblack)..

4. DISCUSSION

The use of OCR technology to extract text characters from digital images is not optimal enough. Toha and Triayudi (2022), in their research entitled "Application of Reading Text in Images Using the Website-Based OCR Method on e-KTP" concluded that the use of OCR is still not optimal [23]. Abdullah and Muhammad (2020) implemented OCR technology to read characters on identity card. The research give the text match rate of 85% [24]. The similar research was done by Afifah, Sujono, and Brilliant (2020) which produced the same level of accuracy, that is 85% [25].

The not optimal extraction of text characters using OCR technology can be caused by several factors, one of which is the quality of the digital image. So before applying OCR technology, image quality needs to be improved through preprocessing. This research discusses several preprocessing methods such as Otsu Thressholding, Niblack, and Sauvola methods which are widely used to enhance image quality specially for scanned document.

In previous studies, the characters were extracted using OCR is to small and limited, such as the characters on vehicle license plates [6], [20], [26] and residence identification numbers [21], [23], [24]. So that the match rate obtained cannot represent the characters in a document. In this study, several sample images with a more varied number of characters were used like vehicle license plates, residence identification numbers, paragraph in a book, bank transfer receipt, book page, etc. So that the match rate obtained can be more accurate.

The results of this study indicate that the best image character extraction process results is extraction on the original image without image preprocessing. However, it should be noted that in some cases, as in the tests in Figures 6, 8, and 12, the Otsu method gives better results. Whereas in Figures 7, 12 and 13 the Sauvola method gives better results. The author assumes that this is due to the different characteristics of the images, especially from different lighting levels. Therefore, the authors hope that there will be research that is able to explain this in the future.

5. CONCLUSION

The best image character extraction process using Tesseract OCR technology is extraction on the original image without image preprocessing with the match rate of 77.27%. Meanwhile, if the preprocessing process is carried out, then the Otsu Thresholding method gives the best results of

70.27%, followed by the Saovula method with a slightly different value of 69.67%. While the use of the Niblack method give the lowest match rate of 35.72%.

However, it should be noted that in some cases, the Otsu method and the Sauvola method give the best results compared to other methods.

REFERENCES

[1] R. Munir, "Pengolahan Citra Digital dengan Pendekatan Algoritmik," Bandung, Informatika, 2004.

[2] D. Putra, "Pengolahan Citra Digital," Yogyakarta, Penerbit Andi, 2010.

[3] G. Kumar and P. K. Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems.," in *Fourth International Conference on Advanced Computing & Communication Technologies*, Rohtak, India, 2020.

[4] A. Rifiana , M. B. Achmad and T. Maulana, "Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment," *International Journal of Advanced Computer Science and Applications,* vol. 9, no. 11, 2018.

[5] T. Somashekar, "A Survey on Handwritten Character Recognition using Deep Learning Technique," *Journal of University of Shanghai for Science and Technology,* vol. 23, no. 6, 2021.

[6] K. I. Mail and M. G. Suryanata, "Ekstraksi Karakter Citra Menggunakan Optical Character Recognition Untuk Pencetakan Nomor Kendaraan Pada Struk Parkir.," *Jurnal Media Informatika Budidarma,* vol. 4, no. 4, 2020.

[7] M. D. Azis, S. A. Syakri and Z. K. Simbolon, "Rancang Bangun Aplikasi Perbaikan Citra Hasil Scan Dokumen Lama Dengan Metode Filtering," *Jurnal Teknologi Rekayasa Informasi dan Komputer,* vol. 1, no. 2, 2018.

[8] D. Z. Putri, D. Puspitaningrum and Y. Setiawan, "Konversi Citra Kartu Nama ke Teks Menggunakan Teknik OCR dan Jaro-WInkler Distance," *Jurnal TEKNOINFO,* vol. 2, no. 1, 2018.

[9] Z. Huang, K. Chen, X. B. He and S. S. Karatzas, "Competition on Scanned Receipt OCR and Information Extraction," in *International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2019.

[10] H. T. Ha, Z. Nevˇeˇrilová and A. Horák, "Recognition of OCR Invoice Metadata Block Types," in *International Conference on Text, Speech, and Dialogue*, Springer, 2018.

[11] D. Lin, F. Lin, F. Cai and D. Cao, "Chinese Character Captcha Recognition and Performance Estimation Via Deep Neural Network," *Neurocomputing,* vol. 288, 2018.

[12] B. G. Weinstein, "A computer vision for animal ecology," *Journal of Animal Ecology,* vol. 87, no. 3, 2018.

[13] K. A. Hamad and M. Kaya, "A Detailed Analysis of Optical Character Recognition Technology," *International Journal of Applied Mathematics, Electronics and Computers,* vol. 4, no. 1, 2016.

[14] P. V. Anh, N. D. T. Khan and T. Manh , "Improved OCR Quality for Smart Scanned Document Management System, Journal of Science and Technique," *Le Quy Don Technical University,* vol. 210, no. 9, 2020.

[15] M. Brisinello, R. Grbi, M. Pul and T. Anđeli, "Improving Optical Character Recognition Performance for Low Quality Images," in *International Symposium ELMAR*, IEEE, 2017.

[16] D. T. Anggraeni, "Perbaikan Citra Dokumen Hasil Pindai Menggunakan Metode Simple, Adaptive-Gaussian, dan Otsu Binarization Thresholding," *Jurnal Manajemen Sistem Informasi dan Teknologi,* vol. 11, no. 2, pp. 71-77, 2021.

[17] F. Kiki, Segmentasi Teks Naskah Kuno yang Lapuk Menggunakan Adaptive Local Thressholding, Surabaya: Departemen Teknik Komputer, Institut Teknologi Sepuluh November, 2018.

[18] Nanonets, "How to OCR with Tesseract, OpenCV and Python," Nanonets, 2023. [Online]. Available: https://nanonets.com/blog/ocr-with-tesseract. [Accessed 2023].

[19] B. Baso, D. Nababan, R. Risald and R. Y. Kolloh, "Segmentasi Citra Tenun Menggunakan Metode Otsu Thresholding dengan Median Filter," *Jurnal Teknologi dan Ilmu Komputer Prima,* vol. 5, no. 1, 2022.

[20] D. R. Medinah and S. Sinurat, "Analisa dan Perbandingan Algoritma Otsu Thresholding dengan Algoritma Region Growing Pada Segmentasi Citra Digital," *Journal of Computer System and Informatics (JoSYC),* vol. 2, no. 1, pp. 9-16, 2020.

[21] N. I. Santikasari, R. D. Atmaja and E. Susatio, "Analisis Dan Implementasi Metode Niblack Pada Sistem Pengenalan Identitas Berbasis Palm Vein," *e-Proceeding of Engineering,* vol. 3, no. 1, 2016.

[22] M. Rofi'i and D. R. Ningtias, "Local Adaptive

Thresholding Menggunakan Metode Sauvola sebagai Tahapan Pra Pengolahan pada Data Citra Isyarat EKG (Elektrokardiogram)," *Jurnal Teori dan Aplikasi Fisika*, vol. 10, no. 1, 2022.

[23] M. R. Toha and A. Triayudi, "Penerapan Membaca Tulisan di dalam Gambar Menggunakan Metode OCR Berbasis Website pada e-KTP," *Jurnal Sains dan Teknologi*, vol. 11, no. 1, 2022.

[24] S. S. Abdullah and F. D. Muhammad, "Penggunaan e-KTP untuk Registrasi Otomatis Memanfaatkan Sistem OCR Dengan Metode Template Matching Correlation," *Media Jurnal Informatika*, vol. 12, no. 2, 2020.

[25] Y. Afifah, A. Sujono and C. H. Brilliant, "The Line Segmentation Algorithm of Indonesian Electronic Identity Card (e-KTP) for Data Digitization," in *THE 5TH INTERNATIONAL CONFERENCE ON INDUSTRIAL, MECHANICAL, ELECTRICAL, AND CHEMICAL ENGINEERING*, Surakarta, 2020.

[26] T. I. Cahyani, M. Zakiyamani, R. Riana and Hardi, "Perbandingan Akurasi Pengenalan Karakter Plat Nomor Menggunakan Tesseract Dan Data Latih Emnist," *Journal of Information Technology and Computer Science*, vol. 5, no. 2, 2022.