IMAGE BACKGROUND PROCESSING FOR COMPARING ACCURACY VALUES OF OCR PERFORMANCE

Desiana Nur Kholifah[1]; Hendri Mahmud Nawawi[2]; Indra Jiwana Thira[3]

[1,2,3] Program Studi Ilmu Komputer
STMIK Nusa Mandiri
www.nusamandiri.ac.id
[1]desiana.dfh@bsi.ac.id; [2]hendri.hiw@bsi.ac.id; [3]indrathira@gmail.com

Abstract—*Optical Character Recognition (OCR) is an application used to process digital text images into text. Many documents that have a background in the form of images in the visual context of the background image increase the security of documents that state authenticity, but the background image causes difficulties with OCR performance because it makes it difficult for OCR to recognize characters overwritten by background images. By removing background images can maximize OCR performance compared to document images that are still background. Using the thresholding method to eliminate background images and look for recall values, precision, and character recognition rates to determine the performance value of OCR that is used as the object of research. From eliminating the background image with thresholding, an increase in performance on the three types of OCR is used as the object of research.*

Keywords: *OCR, document, background image, elimination.*

Intisari—*Optical Character Recognition* (OCR) merupakan aplikasi yang digunakan untuk mengolah citra digital text menjadi text. Banyak dokumen yang memiliki latar belakang berupa gambar pada konteks visual gambar latar belakang tersebut meningkatkan keamanan dokumen yang menyatakan keotentikan, tetapi gambar latar belakang menyebabkan kesulitan pada kinerja OCR karena menyulitkan OCR mengenali karakter yang tertimpa oleh gambar latar belakang. Dengan penghilangan gambar latar belakang dapat memaksimalkan kinerja OCR dibandingkan dengan citra dokumen yang masih berlatar belakang. Menggunakan metode thresholding untuk mengeliminasi gambar latar belakang dan mencari nilai recall, precision, dan character recognition rate untuk mengetahui nilai kinerja OCR yang dijadikan objek penelitian. Dari penghilangan gambar latar belakang dengan thresholding ini didapat peningkatan kinerja pada tiga jenis OCR yang dijadikan sebagai objek penelitian.

Kata Kunci*: OCR, dokumen, gambar latar belakang, eliminasi.*

INTRODUCTION

Optical Character Recognition (OCR) technology has various applications in document processing. Many documents have a background image, while the background image increases document security or visual effects (Shen & Lei, 2015). Usually the background with writing pictures or letter characters such as writing on blank KTP, SIM and other documents that have a background so that the results are more optimal then the background must be removed. Optical Character Recognition (OCR) performance is influenced by the image quality that the characters will recognize. One of the most influential factors on the quality of the image so that it can be read properly OCR is the background/background of the image itself. Image enhancement is a very important part of processing low-level images. The aim is to improve the quality of images that have low contrast values, to increase the difference in intensity between objects and background images and increase the interpretability or perception of information contained in the image (Ahmad & Hadinegoro, 2012).

Therefore it is very important to reprocess the document by deleting the background image before text detection. Image background removal is expected to improve OCR performance compared to background image. The dataset that will be used is an image that has a lot of background including KTP, SIM, and award certificates, and other documents that have a background image.

MATERIALS AND METHODS

Definition of Image

The image is an image in a two-dimensional plane which is produced from a two-dimensional analog image and continues to be a discrete image, through the process of sampling an analog image divided into M rows and N columns

so it becomes a discrete image (Kumaseh, Latumakulita, & Nainggolan, 2013).

Image improvement is one of the simplest and most interesting methods in digital image processing. The idea behind image improvement techniques is to bring out obscured details, or just to highlight certain interesting features in an image (Ahmad & Hadinegoro, 2012). The purpose of image quality improvement techniques is to process the image so that the results have a relatively better quality than the initial image for a particular application. Good words here depend on the type of application and the problem at hand (Priyawati, 2013).

Tresholding

The basic understanding of thresolding states that the left histogram represents the image f (x, y), which is composed of bright objects on a dark background (Yogi, 2016).

Digital Image

In general, digital image processing refers to processing 2-dimensional images using a computer. The main purpose of image processing is that images are easily interpreted by humans and machines. With image processing, an image is transformed into another image (Gusa, 2013).

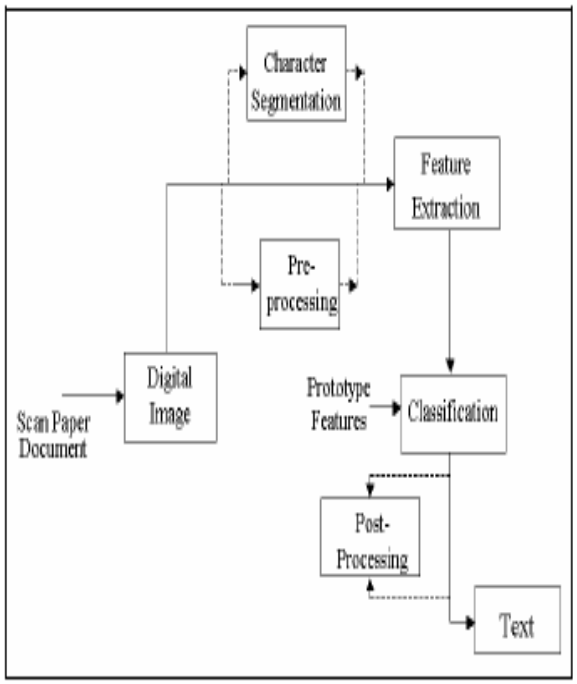
Grayscale image

Grayscale image is a digital image that only has one channel value in each pixel, in other words, the value of the section red = green = blue (Gusa, 2013), (Susanto, 2019). On changing an image to grayscale can be done by taking all the pixels in the image then the color of each pixel will be taken information about 3 basic colors namely red, blue and green (through the warnatoRGB function), these three basic colors will be added up then divided by three so that the values obtained average (Santi, 2011). This value is used to indicate the level of color intensity that a Grayscale image has is a gray color with various levels from black to white (Gusa, 2013).

Optical Character Recognition

Optical Character Recognition (OCR) involves translating images from written text types into an editable text engine. In other words, this involves translating character images into standard encoding representations (Alginahi & Munawarah, 2008). The level of recognition using OCR depends on the image to be recognized, the lighting, and the right image capture (Bahtiar, 2016).

The processing steps of the OCR system are explained in the following figure



Sources : (Alginahi & Munawarah, 2008)

Figure 1. OCR System Steps

First, the document was scanned using the image management tools in this study, the author uses onlineocr.net, convertio.co, and smallseotools.com. The image that I use is an image that has a KTP, SIM and certificate background, and then calculates the accuracy value on the original image. The next step is to change the original image to grayscale, which is the process of changing the RGB image to gray or Grayscale aims to simplify the processing of the image object because in the color image at each pixel there are three layers of color namely Red, Green, and Blue while in grayish images each pixels are only represented by one level of gray. Grayscale process can be done by taking the average value of R, G, and B (Budianita, Jasril, & Handayani, 2015). After the original image is extracted to grayscale then it is tested again with the same tools and the results are a significant increase in the accuracy value before and after extraction.

The image processing starts from image data acquisition, floating, edge detection, image segmentation until the image is ready to be analyzed to calculate the number of image pixels (Gusa, 2013). The purpose of image acquisition is to determine the data needed and choose a method for recording digital images (Hasugian & Zufira, 2018).

The stages of this research are as follows:

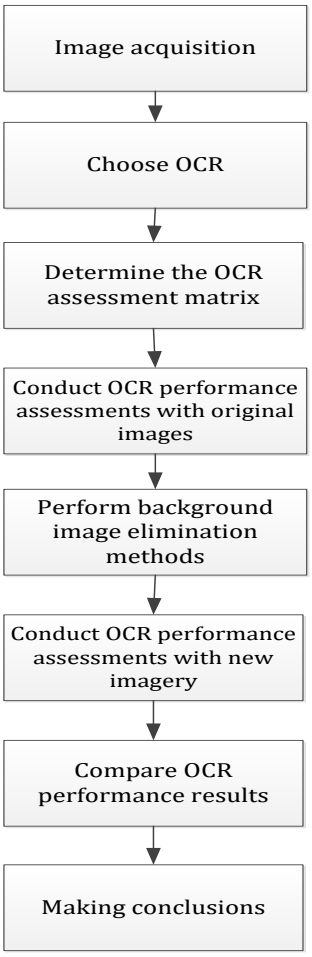
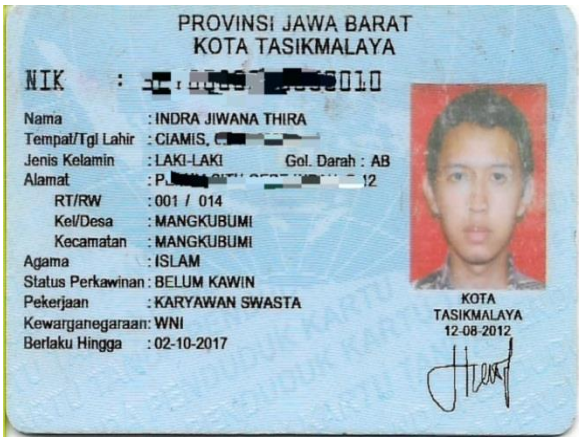


Figure 2. Research Stages

In this study the acquisition of images in the form of scans using the HP Deskjet Ink Advantage 2135 scanner with a resolution of 200dpi.

The following image is used as a dataset :



Sources : (Kholifah, Nawawi, & Thira, 2020)

Figure 3. KTP Image

Sources : (Kholifah et al., 2020)

Figure 4. SIM Image

Sources : (Kholifah et al., 2020)

Figure 5. Certificate Image

Choosing the OCR used in this study is OCR online.net, convertio.co, and smallseotools.com.

Determining OCR assessment metrics, there are several metrics used to measure OCR performance. The general accuracy of OCR is defined as follows :

$$Character\_recognition\_rate = (nm)/n \quad \text{..... (1)}$$

Where n is the correct number of characters, and m is the number of errors. Recall and Precision are usually used in image capture. We can adapt them to OCR size. Recall is defined as follows :

$$Recall = n/all \quad \text{..... (2)}$$

Where n is the correct number of characters, and all is the total character in the image region. Precision is defined as follows :

$$Precision = n/(n+m) \quad \text{..... (3)}$$

Where n is the number of correct characters and m is the number of errors.

But none of the three metrics are perfect in the OCR scenario. Base metrics will return a

negative level when n is less than m. Negative levels are not intuitive to interpret. The recall method is proportional to n but is not related to m. The precision method does not reflect all factors. For example, if all, n, m are 50, 5, 0 each, the level of character recognition is 1 with the precision method, and that doesn't make sense. And obtained metrics to measure the accuracy of OCR as follows (Shen & Lei, 2015):

Character_recognition_rate=n/(all+m) ..... (4)

Conduct OCR performance assessments with original images. Following is the OCR performance evaluation :

Table 1. OCR Performance Values on Original Image

Metric	Optical Character Recognition (OCR)		
	online ocr.net	convertio.co	smallseotools.com
Character Recognition	91,10%	79,30%	89,74%
Word recognition	76,91%	78,33%	86,18%
Area Introduction	53,39%	62,19%	64,89%
Precision	91,99%	83,47%	90,83%

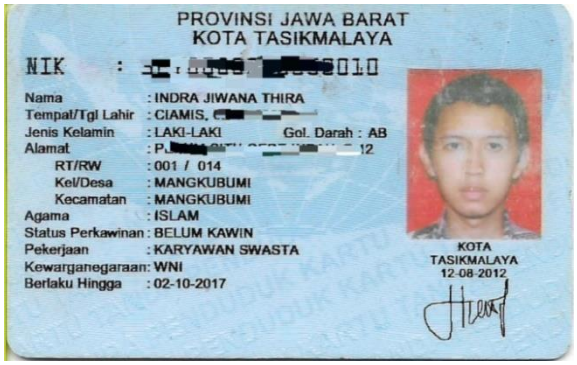
Sources : (Kholifah et al., 2020)

On character recognition, OCR onlineocr.net is more accurate than convertio.co and smallseotools.com. In word recognition, smallseotools.com OCR is superior to onlineocr.net and convertio.co. In the introduction area, Smallseotools is also superior to the two OCRs. And in terms of precision, onlineocr.net is superior to smallseotools.com and convertio.co.

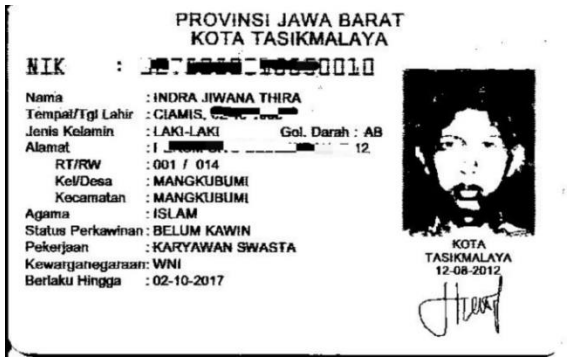
The remove background image method used in this study is the thresholding method, where the RGB image is converted into a binary image and then returned to an RGB image with a background value of 255 (white).

RESULTS AND DISCUSSION

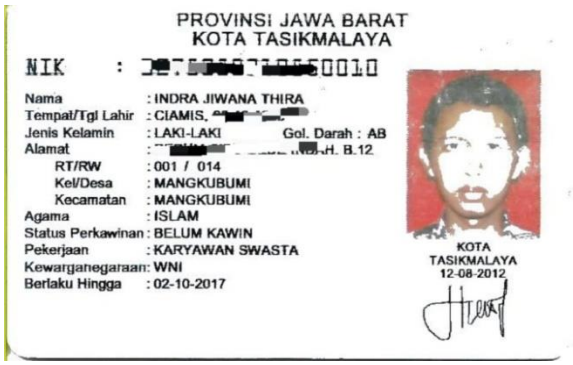
Processing the original image in Matlab with the thresholding method to eliminate the background image with the following results :



(a)



(b)



(c)

Sources : (Kholifah et al., 2020)

Figure 6. Original KTP Image (a), Image Segmentation (b), Result Image (c)

(a)

(b)

(c)

Sources : (Kholifah et al., 2020)

Figure 7. Original SIM Image (a), Image Segmentation (b), Image Result (c)

(a)

(b)

(c)

Sources : (Kholifah et al., 2020)

Figure 8. Original Image Certificate (a), Image Segmentation (b), Result Image (c)

The results of the image that has been thresholded are uploaded to the OCR application, then the OCR results are calculated by evaluating the performance of the OCR with the new image that has been processed with the following results :

Table 2. OCR Performance Values Using New Image

Metric	Optical Character Recognition (OCR)					
	onlineocr.net		convertio.co		smallseotools.com	
	Before	After	Before	After	Before	After
Character Recognition	91,10%	95,39%	79,30%	92,50%	89,74%	91,60%
Word Recognition	76,91%	90,67%	78,33%	93,75%	86,18%	88,83%
Area Introduction	53,39%	67,46%	62,19%	75,37%	64,89%	68,60%
Precision	91,99%	95,60%	83,47%	93,04%	90,83%	92,34%

Sources : (Kholifah et al., 2020)

From table 2 above it can be seen that there is a difference in the accuracy of the original data before it is processed by removing the background image. In terms of character recognition, onlineocr.net remains the most superior with a value of 95.39%, its percentage increased by

4.29%, the percentage increase in convertio.co by 13.20% and smallseotools.com by 1.86%.

In the word recognition metric, convertio.co is superior to the other two OCRs, namely 93.75% from 78.33% with a percentage increase of 15.42%, an increase in onlienocr.net by 13.76% and smallseotools by 2, 65%. The area recognition metric, convertio.co is superior with a percentage increase of 13.8%, percentage increase from onlineocr.net by 14.07% and smallseotools.com the percentage increase is 3.71%. On the metric level the precision of convertio.co was higher with the percentage increase of 9.57%, onlineocr.net the percentage increase was 3.61%, and the percentage and smallseotools.com increased by 1.51%.

From the above data, it can be concluded that the removal of background image using thresholding can improve the performance of three OCRs which are used as research objects with an average increase in all performance and OCR of 8.07%.

CONCLUSION

Conclusion of this research is the removal of background image using the thresholding method can improve the performance of three OCRs that are used as research objects tested by OCR tools onlineocr.net, convertio.co, and smallseotools.com, and the result is an increased accuracy in the image under study with average increase in accuracy of 8.07%. There is still much that needs to be developed from this research, namely the use of more and more varied image datasets, the determination of OCR objects based on performance using training data first, and the use of other remove image background methodologies.

REFERENCES

Ahmad, N., & Hadinegoro, A. (2012). Metode Histogram Equalization untuk Perbaikan Citra Digital. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan (SEMANTIK)*, *2012*(Semantik), 439–445. Retrieved from http://publikasi.dinus.ac.id/index.php/semantik/article/view/185

Alginahi, Y. M., & Munawarah, M. (2008). *Digital Image Computing : Techniques and Applications THESHOLDING AND CHARACTER RECOGNITION IN SECURITY DOCUMENTS WITH WATREMARKED BACKGROUND*. https://doi.org/10.1109/DICTA.2008.90

Bahtiar, A. (2016). Sistem Deteksi Nomor Polisi Mobil dengan Menggunakan Metode Haar Classifier dan OCR guna Mempermudah Administrasi Pembayaran Parkir. *Journal of Information and Technology*, *Volume 04*(9), 40–46. https://doi.org/10.1017/CBO9781107415324.004

Budianita, E., Jasril, & Handayani, L. (2015). Implementasi Pengolahan Citra dan Klasifikasi K-Nearest Neighbour Untuk Membangun Aplikasi Pembeda Daging Sapi dan Babi. *Jurnal Sains, Teknologi Dan Industri*, *12*(2), 242–247.

Gusa, F. R. (2013). Pengolahan Citra Digital Untuk Menghitung Luas Daerah Bekas Penambangan Timah. *Jurnal Nasional Teknik Elektro*, *2*(2), 27–34. https://doi.org/10.20449/jnte.v2i2.71

Hasugian, A. H., & Zufira, I. (2018). Perancangan Sistem Restorasi Citra Dengan Metode Image Inpainting. *Jurnal Ilmu Komputer Dan Informatika*, *03*(November), 31–45.

Kholifah, D. N., Nawawi, H. M., & Thira, I. J. (2020). PENGOLAHAN LATAR BELAKANG CITRA UNTUK MEMBANDINGKAN NILAI AKURASI TERHADAP KINERJA OCR. *Jurnal PILAR Nusa Mandiri*, *Vol. 16*, *N*.

Kumaseh, M. R., Latumakulita, L., & Nainggolan, N. (2013). Segmentasi Citra Digital Ikan Menggunakan Metode Thresholding. *Jurnal Ilmiah Sains*, *13*(1), 74. https://doi.org/10.35799/jis.13.1.2013.2057

Priyawati, D. (2013). Teknik Pengolahan Citra Digital Berdomain Spasial Untuk Peningkatan Citra Sinar-X. *Jurnal KomuniTi*, *II*(2), 44–50.

Santi, C. N. (2011). Mengubah Citra Berwarna Menjadi Gray-Scale dan Citra biner. *Teknologi Informasi DINAMIK*, *16*(1), 14–19.

Shen, M., & Lei, H. (2015). Improving OCR performance with background image elimination. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2015*, 1566–1570. https://doi.org/10.1109/FSKD.2015.7382178

Susanto, A. (2019). Penerapan Operasi Morfologi Matematika Citra Digital Untuk Ekstraksi Area Plat Nomor Kendaraan Bermotor. *Pseudocode*, *6*(1), 49–57. https://doi.org/10.33369/pseudocode.6.1.49-57

Yogi, M. (2016). Aplikasi Deteksi Kematangan Buah Semangka Berbasis Nilai RGB Menggunakan Metode Thresholding. *Jurnal Riset Komputer (JURIKOM)*, *3*(6), 84–89.