Implementation of Average Filter and Median Filter for OCR Pre Processing of Incoming Letters Image

1 I Made Avendias Mahawan, 2 I Putu Agus Eka Darma Udayana
1,2 Department of Informatic Engineering, STMIK STIKOM Indonesia, Indonesia
1 avendias@stiki-indonesia.ac.id, 2 agus.ekadarma@stiki-indonesia.ac.id

Abstract. Letters are one of the tools to communicate through writing and are useful for conveying information. STMIK STIKOM Indonesia campus (called STIKI Indonesia) receives letters from other institutions reaching an average of 50 letters a month, an automatic system is needed that can help manage the incoming letters and can determine the purpose of the letter. OCR (Optical Character Recognition) is one of the ways that can be used to find a text contained in an image, but to get good results in the OCR process requires a pre-process in order to reduce noise in letter images, combining average filters and median filters be an option because it succeeded in reducing noise and smoothing the image of the organization's stamp and signature that piled on the name of the sender of the letter. The results obtained from OCR after going through pre-processing are an array of words from letters. The accuracy achieved based on the date of the letter, letter number, subject of the letter and the sender of the letters from 50 test images reached 90%.

1. Introduction
STMIK STIKOM Indonesia is a campus that currently has 3,953 active students and 1,400 new students entering the academic year 2019/2020, with the large number of students making this campus popular in Eastern Indonesia especially in Bali. As an educational institution under Kementerian Riset Teknologi dan Pendidikan Tinggi this campus has good cooperation between educational institutions or other non-educational institutions in Indonesia and especially in Bali, such as research collaboration with the government of Badung Regency in 2018-2019. One example of good collaboration is invitations to attend or participate in activities organized by other institutions. At present, incoming letters or invitations received by STMIK STIKOM Indonesia reach an average of 50 letters a month, to assist front office staff (FO) in managing incoming letters, STMIK STIKOM Indonesia is starting to make improvements by developing a letter filing system so that it can make it easy to record, search and report letters every year. The system developed for mail management has not been able to run automatically, such as tracing letter numbers, letter dates, senders and letter destinations. In this case it need to add features that can automatically search letter details based on Optical Character Recognition (OCR) technology. To get good results in the OCR process, we need a way that can be done to reduce noise in the letter image, in this study the author tries to combine two methods, namely averaging filter and median filter in the OCR preprocess.

Several studies have been carried out to conduct character recognition in print media on digital media, OCR is an effective solution for the process of converting print documents to digital documents. In research [1]a word input system was designed in a text to speech application by utilizing digital image processing. OCR technology is also applied to research conducted by [2], in this study proposed a...
feature extraction method using Centroid to Boundary and classification using Backpropagation Neural Network that can improve system performance better including time and accuracy of computational processing. In research [3] evaluating OCR performance on Arabic letters data, the results obtained are low accuracy, this is due to the absence of performance evaluation metrics and tools for proper evaluation to evaluate Arabic letters. Based on the three previous studies, OCR is an appropriate technology for managing incoming letters automatically, the researcher considers that an appropriate preprocess is needed so that it can produce good output on OCR, in this study the researcher tries to combine average filter and median filter to reduce noise created during the data acquisition process.

2. Methodology

2.1. Digital Image Processing
According to [4] digital image processing can be interpreted as processing an image on the two dimensions using a digital computer, image processing is also interpreted as a technique used to process, manipulate and modify images. Digital image is a function of two-dimensional light intensity $f(x, y)$, where x and y show spatial coordinates. The value of “$f$” at each point $(x, y)$ shows the level (value) of the color of the image at that point [5].

2.2. Pattern Recognition
Pattern recognition is a study of science that can describe something based on the characteristics of an object and then classify it based on the characteristics of the object. The pattern itself is an entity that is defined and can be identified and given a name. Pattern is a collection of measurement results and expressed in vector or matrix notation [4]. Pattern recognition can also be one way to preserve culture in Indonesia, especially in Bali, such as the classification of Balinese letters or also as a classification of types of Balinese carvings [6].

2.3. System Overview
The process flow in this study is explained in Figure 1.
The main process in this study starts from data acquisition, followed by preprocessing, OCR process and character output obtained from the image of the letter. Pre-processing consists of grayscaling, filtering consists of a combination of averaging filter and median filter. The OCR process consists of a feature extraction process followed by a pattern recognition process.

2.4. Data Acquisition

The method used to collect data is observation and interviews, the results obtained are 50 incoming letters used as test data. The obtained letter is then scanned in the .jpg file format.

2.5. Process Flow

Before the OCR process, the previous process was the grayscaling process which was calculated using the following equation [6]

\[ I(x, y) = \alpha R + \beta G + \gamma B \]  

Information:

- \( I(x, y) \) = Grayscale Image
- \( R \) = Red value in the color image
- \( G \) = Green value in the color image
- \( B \) = Blue value in the color image

With a constant \( \alpha = 0.2989; \beta = 0.5870; \gamma = 0.1140 \)

The results of the grayscaling process are shown in Figure 2

\[ \text{Figure 2. Grayscaling Image} \]

The next process is filtering, the filter used is the average filter and the median filter. Averaging filters can be calculated by the example applied to the following pixel image.

\[
\begin{array}{ccc}
65 & 66 & 56 \\
68 & 70 & 68 \\
65 & 64 & 64 \\
\end{array}
\]

\[ \text{Figure 3. Example Pixel f}(x,y) \]

The average filter results of the replacement pixel for \( g(y, x) \) are calculated as follows:

\[ g(y, x) = \frac{1}{9} \times (65 + 66 + 56 + 68 + 70 + 68 + 65 + 64 + 64) = 65.111 \cong 65, \text{Based on the calculation of } g(y, x), \text{the pixel in image 3 is replaced by a value of } 65\] [7]

Median filter works using filters and is calculated by sorting the intensity of a group of pixels, then the pixel value is replaced by a certain value, the filter on the median filter contains odd pixels, the filter is
shifted point by point reaching the entire image area, the pixel values are sorted ascending then the median value is calculated. The value obtained is used to replace the value at the center of the filter [7]. An example of a filter median filter is shown in Figure 4.

![Median Filter Example](image)

**Figure 4.** Median Filter

The feature extraction process uses the template matching method, where the pixels of each image are matched with the template provided, the character that has the highest match is considered as the character.

### 3. Result

This study experienced problems when doing OCR on the image of letters that have noise. Through literature studies and trials, researchers find solutions to overcome these problems. The researcher combines two filters, the average filter and the median filter in the OCR preprocess. The results obtained by the system can recognize more words, compared to before using additional filters. Figure 5 shows the results before and after using the filter.

![Result Before and After Filter](image)

**Figure 5a.** Result Before Using Filter
Figure 5b. Result After Using Filter

Figure 5a shows the OCR results have not been able to find the sender of the letter, but after going through the filtering process the results obtained in Figure 5b can recognize some characters that are not readable in Figure 5a. The characters that are read in Figure 5b show the sender of the letter, based on the testing process carried out on 50 images, the results obtained 45 images of the letter can indicate the existence of a number, date, subject, destination and sender of the letter. There are 5 letters that cannot indicate the sender of the letter, it is caused by the stamp of endorsement piling up in the name of the sender of the letter.

4. Conclusion

Based on the analysis of 50 test data, it can be concluded that the level of accuracy resulting from the merging of the averaging filter and the median filter in the OCR preprocess is 90%. 45 OCR image images can indicate the number, date, subject, destination and sender of the letter, 5 images cannot show the sender of the letter because the endorsement stamp merges with the sender of the letter. The researcher suggests that for further research a segmentation method is added to separate the endorsement stamp and the sender of the letter.

References

[1] K. Apriyanti and T. Wahyu Widodo, “Implementasi Optical Character Recognition Berbasis Backpropagation untuk Text to Speech Perangkat Android,” IJEIS (Indonesian J. Electron. Instrum. Syst.), vol. 6, no. 1, p. 13, 2016.

[2] B. A. Pratama, W. T. A. Budi, R. K. Nur, and F. Sthevanie, “Mobile OCR using centroid to boundary and Backpropagation Neural Network,” 2015 3rd Int. Conf. Inf. Commun. Technol. ICoICT 2015, pp. 364–368, 2015.

[3] M. A. Alghamdi, I. S. Alkhazi, and W. J. Teahan, “Arabic OCR evaluation tool,” Proc. - CSIT 2016 2016 7th Int. Conf. Comput. Sci. Inf. Technol., 2016.

[4] D. Putra, Pengolahan Citra Digital. Yogyakarta: Andi, 2010.

[5] R. C. Gonzalez and R. E. Woods, Digital Image Processing, USA: Pearson/Prentice Hall, 2008.

[6] I. M. A. Mahawan and A. Harjoko, Pattern recognition of balinese carving motif using learning vector quantization (LVQ), vol. 788. 2017.

[7] A. Kadir and A. Susanto, Teori dan Aplikasi Pengolahan Citra Digital. Yogyakarta: Andi, 2013.