Performance Comparison Online and Offline for Printed Jawi Character Recognition

Ahmadiar and Sayed Muchallil*

Department of Electrical and Computer Engineering, Faculty of Engineering, Syiah Kuala University, Indonesia; ahmadiar@unsyiah.ac.id, sayed.muchallil@unsyiah.ac.id

Abstract

Objectives: OCR for Jawi character both using offline or online application have been built. In this paper, we present performance comparison between online and offline feature extraction for Jawi character recognition. Methods/Analysis: This comparison is carried out to choose which version of the application that run faster to be applied in an OCR system. Both applications uses moment invariant feature extration, the online system uses PHP. On the other hand, the offline method uses Matlab. This research uses Database Printed Jawi (DPJ) as character samples. A 10 percent of total sample is taken as data-set for experiment. Findings: The result of this research shows that online application executes faster than offline process. Online OCR application is suggested to develop OCR systems. Novelty/Improvement: This comparison proposed for comparing and suggesting what kind of OCR system based would be considered.

Keywords: Moment Feature, OCR Performance Comparison, Online OCR, Printed Jawi Characters

1. Introduction

Aceh had a lot of old and historical manuscripts. Some of these documents have been swept away by Tsunami at December 24th, 2004. Others have been processed for physical restoration, in Rehabilitation and reconstruction phase after the Tsunami. Restoration process cannot guarantee that the document can be read by many people because this process objective is to protect the hardcopy of the documents from physical damage.

However, the information that is contained in the documents need to be disseminated. Digital restoration will play an important part for the distribution process. This process started from digitalized all the documents and will be ended after the documents can be read as computer text. Denoising process is the second level practice after digitalized process. This process for historical Aceh manuscripts has been investigated.

In addition, Optical Character Recognition (OCR) is the last phase of digital restoration. OCR is a system that is developed for recognizing the character from a digital image to be identified as digital characters. Character recognition is the most vital phase in pattern recognition. Beside character recognition, OCR technology also consists of pattern recognition step, artificial intelligent step and computer vision step (machine vision). OCR system has at least four phases. The phases are 1. Pre-processing which is including image denoising and binarization, 2. Segmentation, 3. Feature extraction and 4. Pattern classification.

Performance comparison is significant in developing applications. A user will consider to use an application which is executed faster and has a better performance. Previous research has been done and suggested that online printed Jawi character had 96.94% successful rate. Execution time comparison between online and offline Jawi character recognition applications has to be done to grasp which version works faster. Previously, many applications in character recognition were built and used offline, Matlab for Example. One of the offline application weakness is that the application need to be reinstalled when it is removed and reused in other computer or by other user. On the other hands, the online application can be used via browser only that has been installed by
default. In addition, online version removes installing process which consumes more times.

Arabic language has a lot of variances like Farsi or Urdu. Indonesia and other Melayu countries such as Malaysia and Brunei Darussalam also adopted its writing style named Jawi. Jawi is an Arabic variants alphabet beside Farsi and Urdu that using in Malay World. Jawi alphabet was using to write Malay language. Malay language is a lingua franca that used in South East Asia that including Indonesia, Malaysia, Singapura and Brunei Darussalam. The research in Jawi character recognition was start in 2000 and 2002. After that many research in Jawi character was published.

As one of Arabic alphabet variant, Jawi alphabets also have same characteristics with Arabic alphabet. However, there are six additional characters in Jawi alphabet, they are Nya, Ca, Nga, Pa, Ga, and Va.

Binarization is a method to get an image with two level representation from a grayscale image. 0 and 1 will be used to represent an image in binary image. Black color is represented as 0. This color considered as background. On the contrary, white is represented as 1 and known as foreground.

This technique should use a threshold to determine which part of an image will be used as background and the latter will be determined as foreground. There are some binarization algorithms like Sauvola algorithm, Otsu algorithm, Su and Lu algorithm and modification algorithm such as. In this paper the threshold was manually set to get the best result.

Feature extraction is a process to get features as attributes from an image. To identify an object in an image, the features must be extracted from the image. Feature extraction aims to get significant features on an image depending on their intrinsic characteristics.

Hu moment invariant is one of extraction feature methods that is commonly used to get image features. Hu extracted the feature from an image using moment method. Hu moment invariant also known as geometrical moment invariant. Hu moment is defined as:

\[
\begin{align*}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{20} - 3\eta_{11})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{20} + \eta_{11})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (3\eta_{21} - \eta_{03})(\eta_{20} + \eta_{11})^2 - (\eta_{21} + \eta_{03})^2
\end{align*}
\]

2. Materials and Methods

This research used the DPJ dataset for experiment. The dataset which is using is printed Jawi character image. Total number of character in the dataset that were used in this experiment is 1524 characters. This study only used 30% for every font style as sample data in experiment. The samples which were taken are considered to represent every different shape from each character. This is to ensure that every different shape of each character drawn as samples in this experiment. Figure 1 is the sample of experiment characters.

![Figure 1. Example of experiment characters.](image)

The offline application for extracted Hu moments was built using MATLAB programming language. The MATLAB script which is developed as the applications was taken from RC Gonzales’s book. The online applications that are tested in this experiment is an application that built using PHP: Hypertext Preprocessor (PHP) programming language. The PHP script was ported from the Matlab source code. The offline application contains two functions. On the other hand, the online application only consists of one function. In online version, binarization and feature extraction process was combined in a function, while in offline application binarization and feature extraction process was packed in a different package.

Comparison of time processing in the application is performed on the same Personal Computer (PC) with specificaion. Time calculation process start by providing the timer function at the beginning and at the end of
the moment invariant calculation script. The similar was applying to binarization process of the grayscale images.

Finally, the result is compared between online and offline to figure out which application is run faster in binarization process and extracted Hu moment feature from an image. The comparing data would be used as references of applying OCR systems for Jawi characters.

Figure 2. Processing time of execution for type font Courier New a) Offline application process b) Online application process c) Comparison figure of offline and online application process.

3. Result and Discussion

In this section, we provided the experimental result of time comparison between offline and online feature extraction in Jawi character recognition. The results of the experiment are showed in Figure 2 to 5. Figure 3 shows the application that tested using courier new font type. Figure 4 shows the application that tested using Sakkal Majalla font type. Figure 5 shows the application that tested using courier new font type. Figure 6 shows the application that tested using times new roman font type.

Figure 2 shows execution time process for extracted moment invariant feature of Courier New font style. Figure 2a shows time process using offline application which is addition from binarization and feature extracted process. Offline application process needs time for execution program between 18 milisecond (ms) until 20 ms but online application only needs 0.00015 ms till 0.00028 ms. Figure 2b shows execution time process for online application. Figure 2c shows comparison time execution process between offline application and online application. This figure shows that online application has much faster processing than offline application.

Figure 3 shows execution time process for extracted moment invariant feature of Sakkal Majalla style. Figure 3a shows time process using offline application which is addition from binarization and feature extracted process. Offline application process needs time for execution program between 18 ms until 19 ms but online application only needs 0.00015 ms till 0.00017 ms. Figure 3b shows execution time process for online application. Figure 3c shows comparison time execution process between offline application and online application. This figure shows that online application has much faster processing than offline application.

Figure 4 shows execution time process for extracted moment invariant feature of Segoe UI font style. Figure 4a shows time process using offline application which is addition from binarization and feature extracted process. Offline application process needs time for execution program between 18 ms until 19 ms but online application only needs 0.00015 ms till 0.00017 ms. Figure 4b shows execution time process for online application. Figure 4c shows comparison time execution process between offline application and online application. This figure shows that online application has much faster processing than offline application.

Figure 5 shows execution time process for extracted moment invariant feature of Times New Roman font style. Figure 5a shows time process using offline application which is addition from binarization and feature extracted process. Offline application process needs time
Figure 3. Processing time of execution for type font Sakkal Majalla, a) offline application process b) online application process c) Comparison figure of offline and online application process.

Figure 4. Processing time of execution for type font Segoe UI, (a) Offline application process (b) Online application process (c) Comparison figure of offline and online application process.
for execution program between 17 ms until 19 ms but online application only needs 0.00013 ms till 0.00016 ms. Figure 5b shows execution time process for online application. Figure 5c shows comparison time execution process between offline application and online application. This figure shows that online application has much faster processing than offline application.

4. Conclusion

This paper presented a comparison of execution time between offline and online OCR systems to find out their performance. The system consists of binarization and feature extraction process. Hu moment was used to extracted image feature. In offline application, binarization and feature extraction process was combined. Vice versa, in online application, the two steps were divided. So that execution time in offline application is combination from binarization process and feature extraction process.

The process of Binarization and feature extraction in online application take 0.00018 ms but in offline application take 18 ms. According to the result we can conclude that binarization and feature extraction processes which are using online application is better and faster than using offline application.

5. References

1. Muchallil S, Arnia F, Munadi K, Fardian. Performance comparison of denoising methods for historical documents. Jurnal Teknologi. 2015; 77(22):137–43. Crossref
2. Omar K. Jawi handwritten text recognition using multilevel classifier. PhD Thesis, Universiti Putra Malaysia; 2000.
3. Manaf M. Jawi handwritten text recognition using recurrent bama neural networks. PhD Thesis, Universiti Kebangsaan Malaysia; 2002.
4. Nasrudin MF, Petrout M, Kotoulas L. Jawi character recognition using the trace transform. 2010 Seventh International Conference on Computer Graphics, Imaging and Visualization (CGIV); 2010. p. 151–6. Crossref
5. Nasrudin MF, Petrout M. Offline handwritten Jawi recognition using the trace transform. 2011 International Conference on Pattern Analysis and Intelligent Robotics (ICPAIR); 2011. p. 87–91. Crossref
6. Razak Z, Rosli S, Mashkuri Y. Hardware design of on-line jawi character recognition chip using discrete wavelet transform. Proceedings of Eighth International Conference on Document Analysis and Recognition; 2005. p. 91–5.
7. Razak Z, Zulkiflee K, Noor NM, Salleh R, Yaacob M. Off-line handwritten Jawi character segmentation using histogram normalization and sliding window approach for hardware implementation. Malaysian Journal of Computer Science. 2009; 22(1):34–43.
8. Sauvola J, Pietikäinen M. Adaptive document image binarization. Pattern Recognition. 2000; 33(2):225–36. Crossref
9. Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 1979; 9(1):62–6. Crossref
10. Lu S, Su B, Tan CL. Document image binarization using background estimation and stroke edges. International Journal on Document Analysis and Recognition. 2010; 13(4):303–14. Crossref
11. Su B, Lu S, Tan CL. Robust document image binarization technique for degraded document images. IEEE
Performance Comparison Online and Offline for Printed Jawi Character Recognition

12. Hu M-K. Visual pattern recognition by moment invariants. IRE Transactions on Information Theory. 1962; 8(2):179–87. Crossref

13. Arnia F, Munadi K, Fardian F, Muchallil S. Improvement of binarization performance by applying dct as preprocessing procedure. Proceedings of 2014 6th International Symposium on Communications, Control, and Signal Processing, ICCSP 2014. IEEE Press; 2014. p. 128–32.

14. Gonzalez RC, Woods RE, Eddins SL. Digital image processing using Matlab, New Jersey: Pearson Prantice Hall; 2004.

15. Shih FY. Image processing and pattern recognition: Fundamentals and techniques. New Jersey: John Wiley and Sons, Inc; 2010. Saddami K, Munadi K, Arnia F. A database of printed Jawi character image. Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015. IEEE Press; 2015. p. 56–9.

16. Saddami K, Munadi K, Arnia F. A database of printed Jawi character image. In: Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015. IEEE Press; 2015. p. 56–59. Crossref