Jurnal Teknologi

Full Paper

IDENTIFICATION OF MOST SUITABLE BINARISATION METHODS FOR ACEHNESE ANCIENT MANUSCRIPTS RESTORATION SOFTWARE USER GUIDE

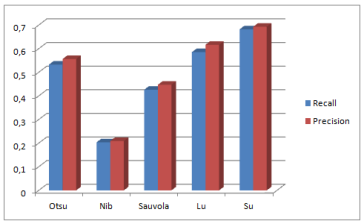
Fardian*, Fitri Arnia, Sayed Muchallil, Khairul Munadi

Electrical Engineering Department, Syiah Kuala University, Banda Aceh, Indonesia

*Corresponding author
fardian@unsyiah.ac.id

Article history
Received
27 April 2015
Received in revised form
15 June 2015
Accepted
25 November 2015

Graphical abstract



Abstract

The Aceh Museum stores many digitized ancient manuscripts from hundreds of years ago. The condition of those manuscripts has degraded into several degradation types such as uneven contrast, show through effects, background spots, and text fading, which cause decreasing readability. A binarisation method is used to decrease the degradation effect on ancient manuscripts. Our research team is currently working on developing application software that consists of five binarisation methods, namely Otsu, Niblack, Sauvola, Lu, and Su for ancient manuscript restoration for the Aceh Museum staff to improve documents' readability. In practice, a user still finds it difficult to choose the best method because there is no method that works best on every ancient manuscript for different types of degradation. This paper intends to determine a binarisation method that suits most manuscript conditions. The method used in this research includes the identification and classification of degradation types from 200 ancient Aceh digital manuscripts, followed by cropping the manuscripts to the size of 256 x 256 pixels. As many as five cropped areas from each degradation type are selected as research samples. These samples are binarisated using the methods. The last step is finding the most suitable binarisation method for each degradation type and classifying which methods are considered to have good readability, and that achieves at least 80% recall and precision values. From our experiments, we found that the Su binarisation methods demonstrate the best performance overall for every degradation type. Otsu, Lu, and Su are suited for uneven background; Sauvola, Lu, and Su are suited for showthrough effects; Otsu, Sauvola, and Su are suited for background spots; and Otsu and Su are suited for both text and background blurring and 'fox'.

Keywords: Restoration Software; Binarisation, Document Degradation

1.0 INTRODUCTION

Aceh province, Indonesia, has many ancient manuscripts. Some of them are stored in the Aceh Museum, and many of those collections have been digitized. The conditions of those manuscripts, which are from hundreds of years ago, have degraded so that their readability has decreased. Our research team is currently working on developing an application software that consists of five binariation methods, namely, Otsu, Niblack, Sauvola, Lu, and Su for ancient manuscript restoration for the Aceh Museum staff. This application provides guidance for the user to choose which methods are suitable for any degradation type of the manuscript available at the museum. Since the degradation types of a manuscript may vary from one to another document, such as uneven background, show through effects, background spots, text fading, text and background blurring, and 'fox' [1], the staff finds it difficult to choose the best method that is considered best for every document. This paper intends to decide which

binarisation method is suitable for most manuscript conditions. The information on the most suitable binarisation method will then be used for user guidance of our application software in order to help the Aceh Museum staff by suggesting the most suitable method every time they use this software.

The method used for this research was started by identifying and classifying degradation types from 200 sheets of ancient Aceh manuscripts. The next step was cropping the manuscripts into 256 x 256 pixels on the degradated location—five samples for six degradation types as research samples. The third step was binarising samples using the methods. The last step was finding the most suitable binarisation method for each degradation type and classifiying which methods are considered to have good readability that achieves at least 80% recall and precision values.

From our expreriments, we found that the Su binarisation method demonstrates the best performance overall for every degradation type. Otsu, Lu, and Su are suited for uneven background; Sauvola, Lu, and Su are suited for showthrough effects; Otsu, Sauvola, and Su are suited for background spots; and Otsu and Su are suited for both text and background blurring and fox.

2.0 LITERATURE REVIEW

There are many ancient, historical documents that have been transformed into digital form to preserve the quality of the original documents, and also to provide scholars with access to the information [2]. Such documents might be in a language that is currently not commonly used; they might contain handwritten or machine printed text that is often hard to read, have a lot of noise, and be corrupted by various artifacts. There is no method that is considered the best to work on every document condition that contains noise. It is quite common that noise found in ancient documents suffer from degradation problems, such as uneven background, showthrough effects, background spots, text fading, text and background blurring, and fox [3]. Suitable binarisation methods are needed to remove noise and improve the readability of historical document images.

Document image binarisation is a process to remove any existing degradation by segmenting pixels of the document image into text and background. It is the pre-processing step on the document image processing that has an important role for the subsequent processes, and contributes on the success rate of OCR (optical character recognition) performance [4]. Global thresholding that calculates statistical properties of the document is employed in some binarisation methods such as Otsu. This thresholding is suitable for documents in nearly ideal condition, where texts can be completely segmented from the background. However, global thresholding is not ideal for a document that contains noise, such as an ancient manuscript [5].

Much research has developed local thresholding methods to improve binarisation results of a document with noise, and that can adapt to the variation of readability degradation in the documents. The Niblack method [6] results in a binary document that suffers from a large background noise, especially in no-text areas [7]. The Sauvola method is not able to segment characters when the pixel values between characters and background are close. However, it has been reported that the Niblack and Sauvola methods improve the binarisation results for document images with extremely low intensity variation and a whiter image [7]. The Su and Lu methods are reported to be more stable in document images that have different types of degradation, by combining the local image contrast and local image gradient to reduce the background with more variation, avoiding the over-normalisation of document images with less variation [8]. But the performance of the Su and Lu methods still needs to be improved on documents that suffer from different types of degradation, such as water stains, ink bleed-through, and significant foreground text text intensity [8].

Several projects have worked on preserving ancient manuscripts in order to help librarians, historians, researchers, and other users to access information contained in the manuscripts. Such projects include the Timbuktu manuscripts project, which works on digitisation and electronic document management (by developing databases, image capture, storage and backup, and image retrieval) [9]. The Madonne research project funded by the French government is to establish a large database of digitised manuscripts by providing efficient navigation, an indexing system, and data organisation to provide general services to the users [10]. Journet et al. have built assistance tools for humanists and historians to retrieve information from old books. This project characterises old books by indexing their layout extraction to help the user classify the books by their contents [11]. DEBORA (Digital Access to Books of the Renaissance), in addition to digital manuscripts management, also implements binarisation methods for their application software in order to ease user access of the information contained in ancient documents by applying the indirect Sauvola binarisation method on their software to reduce existing degradation of the Renaissance documents. They use multistage segmentation to improve the weaknesses found with the direct Sauvola method implementation when processing their Renaissance dataset [12]. The Binarisation Shop project is similar to our research presented in this paper. It is concerned with assisting the user to operate their binarisation software. However, this software is only suitable for the user who has good knowledge of image processing, especially binarisation methods, because it allows

the user to tune parameters available in binarisation methods. In practice, especially for a user from the Aceh Museum, this parameter-tunable flexibility may not be fully applicable since there are not many users familiar with how binarisation methods work [13].

3.0 METHOD AND EXPERIMENT

The experimental flow used for this research is shown in Fig. 1.

It begins by placing the degradation type into six categories, namely, uneven background, show-through effect, background spots, text fading, text and background blurring, and fox. The next step is classifying the document images according to their degradation types. Then each document is sampled by cropping it into the size of 256 x 256 pixels, but only on its degraded area. The purpose of this step is to ensure that each cropped image represents the degradation type optimally, as shown in Fig. 2. Next, all the cropped images are binarised using five binarisation methods. After binarisation, the next step is to calculate recall and precision values of the binarised image tiles. Finally, the data is analysed to determine which method is the most suitable for each degradation type. The analysis in this research is also to determine which method produces good readability by having recall and precision values above 80% on average.

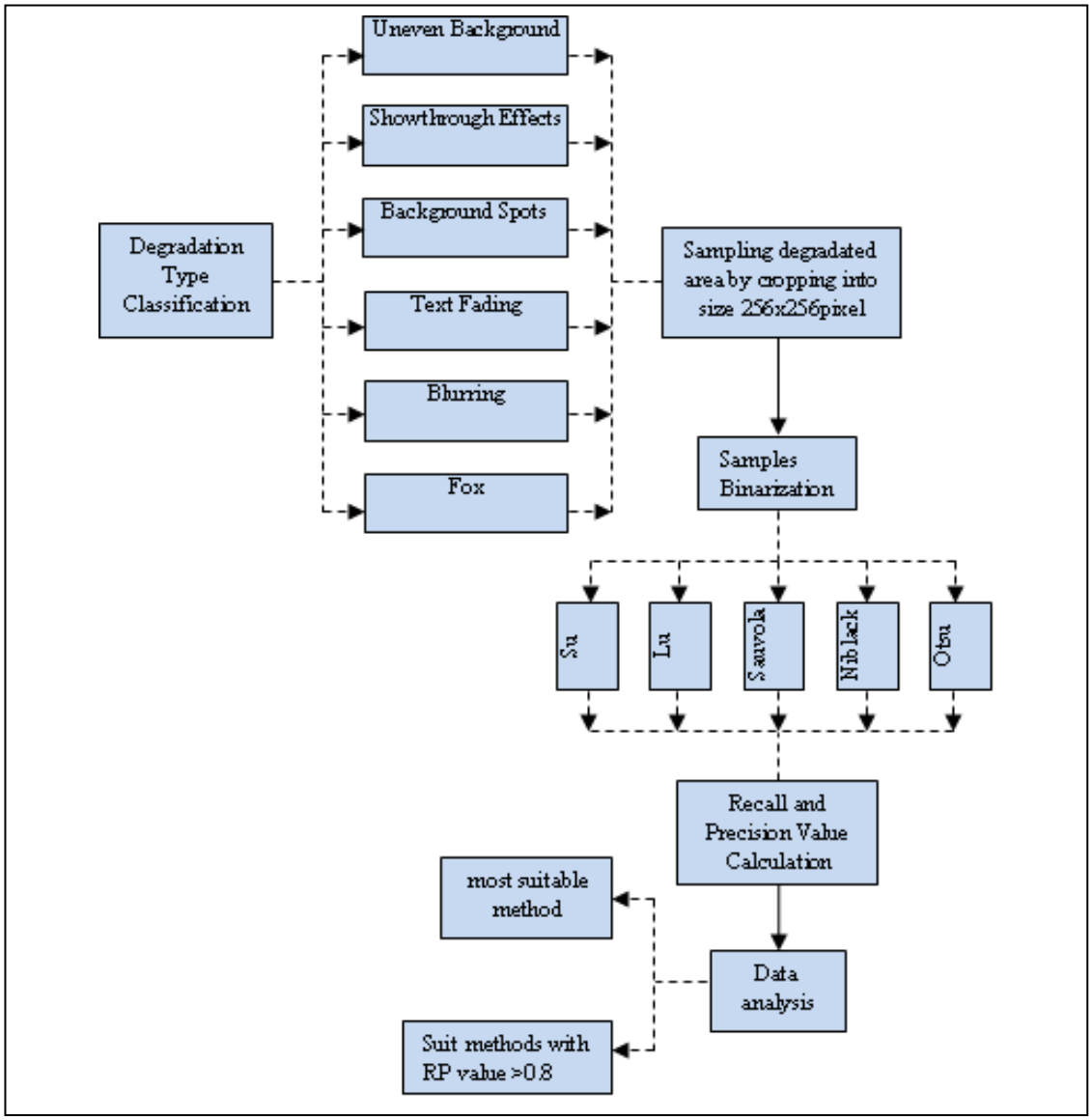


Figure 1 The Experimental Flow

Figure 2 Cropped Original Degraded Document Image Samples (from top left to bottom right: uneven background, show-through effects, background spots, text fading, text and background blurring, and fox)

3.1 Experiment Conditions

In this experiment, 200 sheets of ancient Acehnese manuscripts with Jawi characters are used. Six degradation types are identified and used to classify the documents, namely, uneven background, see through effects, background spots, text fading, blurring, and fox. Each degraded document is sampled by cropping the degraded area with the size of 256 x 256 pixels, and five tile samples for each degradation type, so that the total sampling taken for six degradation types is 30 samples. There are five binarisation methods used, namely, Otsu, Niblack, Sauvola, Lu, and Su. All samples are processed using these methods, and 150 binarised samples are generated in total. Fig. 3 is an example of a noiseless document; Figs. 4 to 9 are examples of documents classified by the kind of degradations.

Figure 3 Example of noiseless document image

Figure 4 Example of document image with uneven background

Figure 5 Example of document image with showthrough effects

Figure 6 Example of document image with background spots

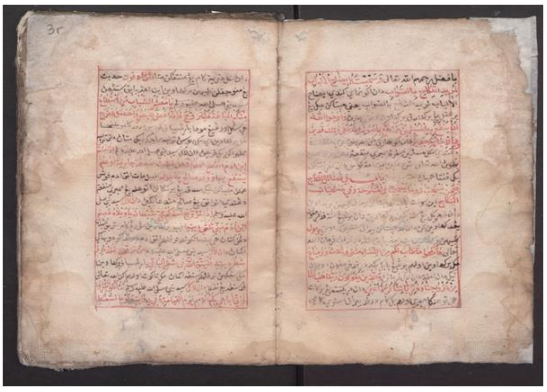


Figure 7 Example of document image with text fading

Figure 8 Example of document image with text and background blurring

Figure 9 Example of fox document image

Figure 10 Cropped original and binarised document images samples (from top left to right: original document, Otsu image, Niblack image, Sauvola image, Lu image, and Su image)

Figure 10 shows the result of image binarisation, after the cropping process to the size of 256 x 256 pixels. From the figure, we can see that there are different binarisation results for each method.

3.2 Evaluation of Binarisation Results

Recall and precision values are used to measure the binarisation performance where the calculation approach is similar to the method in [11]. NCD refers to the number of correctly detected characters in a binarised document, GT (ground truth) refers to the total number of characters in the original document images, and TR refers to the total number of characters detected in binarised documents, including correctly detected and broken characters. Recall and precision are defined by Eq. 1 and Eq. 2, respectively.

$$recall = \frac{NCD}{GT} \quad (1)$$

$$precision = \frac{NCD}{TR} \quad (2)$$

A ground truth image is generated manually by calculating the number of readable and broken characters in the original document images. The NCD and TR are detected with the guidance of its ground truth.

In this experiment, the suitable method is defined as the method that is able to demonstrate a recall and precision value higher that 0.8 (80%) [5].

4.0 RESULTS AND DISCUSSION

4.1 Results

The results of the experiment are provided in Figures 11 to 16. Figure 11 shows the averaged recall and precision values after applying five binarisation methods for uneven background; Figures 12 document images for showthrough effects; Figure 13 for background spots; Figure 14 for text fading; Figures 15 for text and background blurring; and Figures 16 for fox.

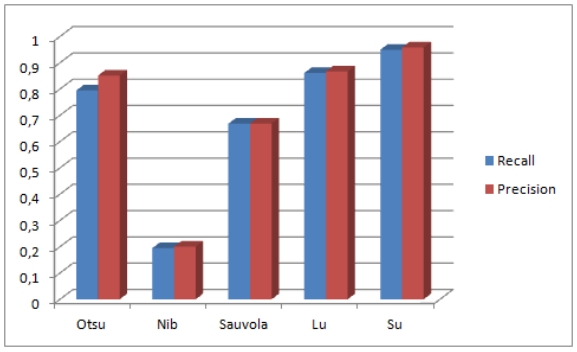


Figure 11 Recall and precision values of uneven background

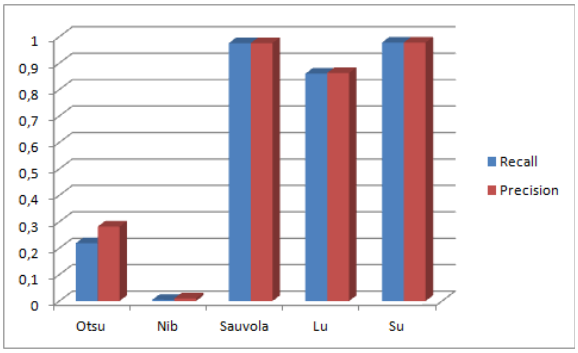


Figure 12 Recall and precision values of showthrough effects

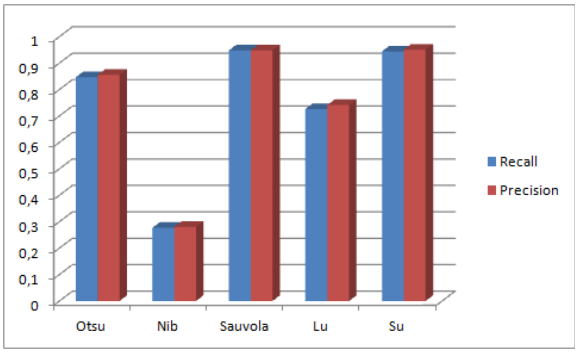


Figure 13 Recall and precision values of background spots

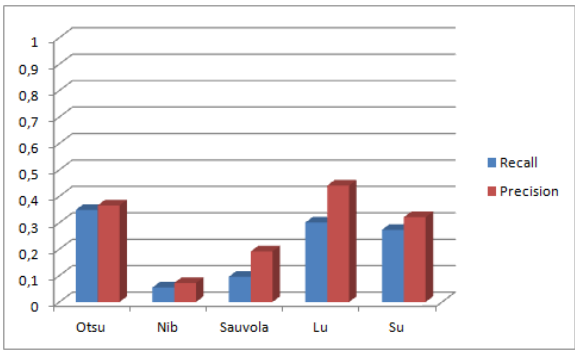


Figure 14 Recall and precision values of text fading

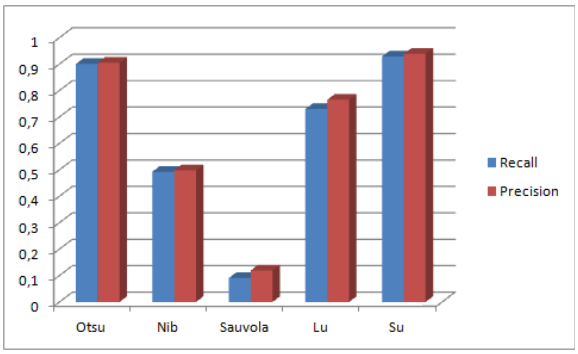


Figure 15 Recall and precision values of text and background blurring

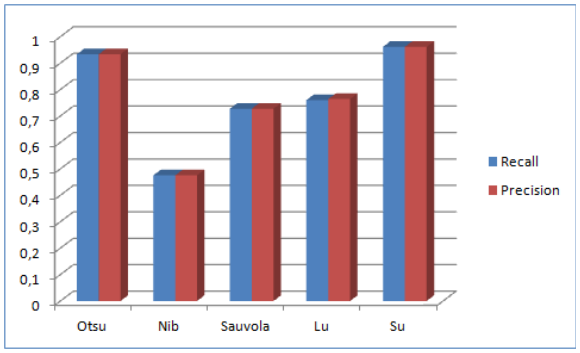


Figure 16 Recall and precision values of fox

4.2 Discussion

The experiment results show that overall, the Su binarisation method demonstrates the best performance among other methods for all degradation types, while the Niblack binarisation method shows the opposite performance. Su is suitable for nearly all degradation types except for text fading, as can be seen on Figure 14

As mentioned in the evaluation of binarisation methods in the Method and Experiment section, this experiment uses recall and precision values above 0.8 to define the suitability of binarisation methods. Using this indicator, Otsu, Lu, and Su are categorised as suitable methods for uneven background. For showthrough effects, Sauvola, Lu, and Su are defined as suitable methods. For background spots, recall and precision values of Otsu, Sauvola, and Su are above 0.8. Interesting results are shown for text fading, as there are no recall and precision values in any method that are above 0.8. This means that no method is categorised as suitable for this degradation type, so we will explore it in the future. Figure 17 shows the binarisation results of all methods for fox degradation type. For text and background blurring type, the suitable methods are Otsu and Su, and this result is also similar to the fox degradation type.

Of all methods used in this research, Niblack is the only method that is not suitable for any kind of degradation.

Figure 17 Cropped original and binarised document images samples for fox degradation type (from top left to right: original document, Otsu image, Niblack image, Sauvola image, Lu image, and Su image)

5.0 CONCLUSION

This paper proposes an identification process for choosing the suitable binarisation method for ancient Acehnese manuscripts that are written in Jawi characters. The identification process resulting from this research is used as user guidance for application software for ancient Acehnese manuscripts' restoration at the Aceh Museum. The experiment conducted for this research involved five binarisation methods—namely, Otsu, Niblack, Sauvola, Lu, and Su—in our software application. Six degradation types, such as uneven background, showthrough effects, background spots, text fading, text and background blurring, and fox, are used, since these types mostly appear on ancient Acehnese documents stored in the Aceh Museum. The research results show that Su is the most suitable method for nearly all degradation types. By defining the recall and precision values higher than 0.8 to indicate the appropriate level of document readability, we found that Su, Lu, Otsu, Sauvola are suitable for several degradation types.

Acknowledgement

The work reported in this paper is the partial result of research projects funded by the Directorate General of Higher Education (DGHE) of the Republic of Indonesia, under the National Strategic Research Grant, with contract no. 013/UN11.2/LT/SP3/2013.

References

[1] F. Stanco, L. Tenze and G. Ramponi, 2007.Technique to correct yellowing and foxing in antique books, *IET Image Process.* 1(2):123133,.

[2] E.Kavallieratou, E.Stamatatos, 2006. Improving the quality of degraded document images, IEEE proceedings of dial, 340-349, *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*.

[3] Ntirogiannis, K. et al. 2012. A Combined approach for the binarization of handwritten document images. *Pattern Recognition Letters*. 35: 3-15

[4] Fitri, A., M. Fardian, M. Sayed, and K. Munadi. 2014. Improvement of Binarization Perfornance By Applying DCT as Pre-Processing Procedure. Communications, Control and Signal Processing (ISCCSP), 6th International Symposium. 128-132. http://dx.doi.org/10.1109/ISCCSP.2014.6877832

[5] Niblack, W. 1986. Introduction to digital image processing. Prentice Hall, New Jersey, pp 115-116.

[6] Khurshid, K., I. Shiddiq, C. Faure, and N. Vincent. 2009. Comparison of Niblack inspired binarization methods for ancient documents. *SPIE Proceedings, 16th Document Recognition and Retrieval Conference, DRR-09, col 7247. 1-10.*

[7] Bolan, S., L. Shijian, T. Chew Lim. 2013. Robust Document Image Binarization Technique for Degraded Document Images. *IEEE Transactions on Image Processing.* 22: 408 - 1417.

[8] Albrecht, H. 2014. Timbuktu Manuscripts Project for the Preservation and Promotion of African Literary Heritage. Department of Culture Studies and Oriental Languages University of Oslo. Accesed: December 28, 2014 from http://www.hf.uio.no/ikos/english/research/projects/timbuktu/

[9] Ogier, J.-M., K. Tombre. 2006. Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. *International Conference on Digital Cultural Heritage, Aug. 2006,* Vienna, Austria.

[10] Journet, Nicholas, et al. *Dedicated texture based tools for characterisation of old books.* Document Image Analysis for Libraries, 2006. DIAL'06. *Second International Conference on. IEEE, 2006.*.

[11] Le Bourgeois, Frank, and Hubert Emptoz.2007. "Debora: Digital access to books of the renaissance." *International Journal of Document Analysis and Recognition* (IJDAR) 9. 2-4: 193-221.

[12] Deng, Fanbo, et al. BinarizationShop: a user-assisted software suite for converting Old Documents To Black-And-White. *Proceedings Of The 10th Annual Joint Conference On Digital Libraries*. ACM, 2010.

[13] Wang, Q., C. L. Tan. 2001. Matching of Double-Sided Document Images to Remove Interference. *Proceedings from the IEEE Computer Vision and Patter Recogintion.* 1: 1084-1089.