# HISTOGRAM E QUALIZATION SMOOTHING FOR DETERMINING THRESHOLD ACCURACY ON ANCIENT DOCUMENT IMAGE BINARIZATION

Mahendar Dwipayana[1,a)], Fitri Arnia[2], Zuhar Musliyana [1,b)]

*[1]Department of Information System, Faculty of Computer Science,*
*Ubudiyah Indonesia University,*
*Jalan Alue Naga, Desa Tibang, Banda Aceh*
*[2]Magister Electrical Engineering, Syiah Kuala University,*
*Darussalam Banda Aceh*
[a)] mahendar@uui.ac.id, [b)] zuhar@uui.ac.id

## ABSTRACT

Ancient documents are inheritance that must be preserved. The documents contain historical, scientific, social, religious information, etc. Converting ancient documents into digital image formats is one of ways to preserve the inheritance and can be stored into a computer. However, images of ancientdocuments have many blemishes caused by age, moisture, flood, etc. Therefore, special techniques are needed for those images to be restored and can improve the legibility of the ancient documents' images. In this study, the image restoration process uses separation of background and foreground/text on histogram equalization such as research conducted by Fitri Arnia in 2008. Through histogram equalizationimages can be seen the distribution of pixels from the intensity of black color "0" to white "1". The distribution of pixels on histogram equalization describes the curves of foreground/text and curves of background. Among the histogram curves, the determination of thresholdvalues can be done so as to clarify the foreground/text and background areas on images of ancient documents. The lowest point between the two curves is the lowest pixel (local minima) which is used as the threshold value. However, the selection of such threshold values in some cases is very difficult to determine because there are still many fluctuations in the curve at the lowest curve. Therefore, this study proposesa histogram smoothing method in the ancient documents' images to minimize curvature fluctuations and to determine more accurate threshold values. In this research, average filtering method is used for smoothing the histogram image. This filter successfully refines the histogram and makes the image of the restoration or binary image display the value of the ancient document image readability increases.

**Keywords**: HistogramEqualization, Smoothing Histogram, Average Filtering, Thresholding

## INTRODUCTION

Many of ancient documents found so far are in very bad condition due to their age, humid storage and so on. In those documents there are many disturbances that make the document difficult to read. Therefore, it is necessary to restore the information contained in ancient documents by converting it first into digital format / digitalization so that reconstruction can be done.

In this study the process of restoration of ancient documents using background and foregraound separationtechniques such as in existing researches and by using histogram equalization [1]. Histogram equalization is useful in fulfilling pixel gradation level and adding color contrast between background and foreground/text. In histogram, the lowest curve or threshold value is obtained which is the reference point for separation between background and foreground.

Histogramequalization method on that study [1] has successfully eliminated fox and noise. This method is not like method used in the research of Otsu [2] that automatically divides the gray level image. The method Otsu used was a large threshold value so that the pixels obtained accumulate on the black color causing some text to be affected. In contrast to *wafa Bousella*, et al., they used the maximum likelihoodand k-means clustering -based estimation methods [3] and iterations with recursive algorithms in separating the background and foreground/text [4].

In this study, the process of restoration of ancient documents' images using four smoothing histogram methods. Those are mean filtering, median filtering, wiener filtering, and cubic spline which are smoothing methods in histogram equalization to facilitate the determination of threshold values. Of the four smoothing methods will be obtained a different binary image on each method. The difference in the results of this binary image will be measured using the recall and precision parameters. These parameters are useful to determine the ability of each smoothing method in restoring ancient document images.

## METHODOLOGY

Subjectin this research is the field of science of digital image processingby using processing method for image quality enhancement. The objects of this study are images of ancient documents derived from Acehs inscribedin Arabic Malay. The steps of test performed in this study can be seen in Figure 1 as follows.

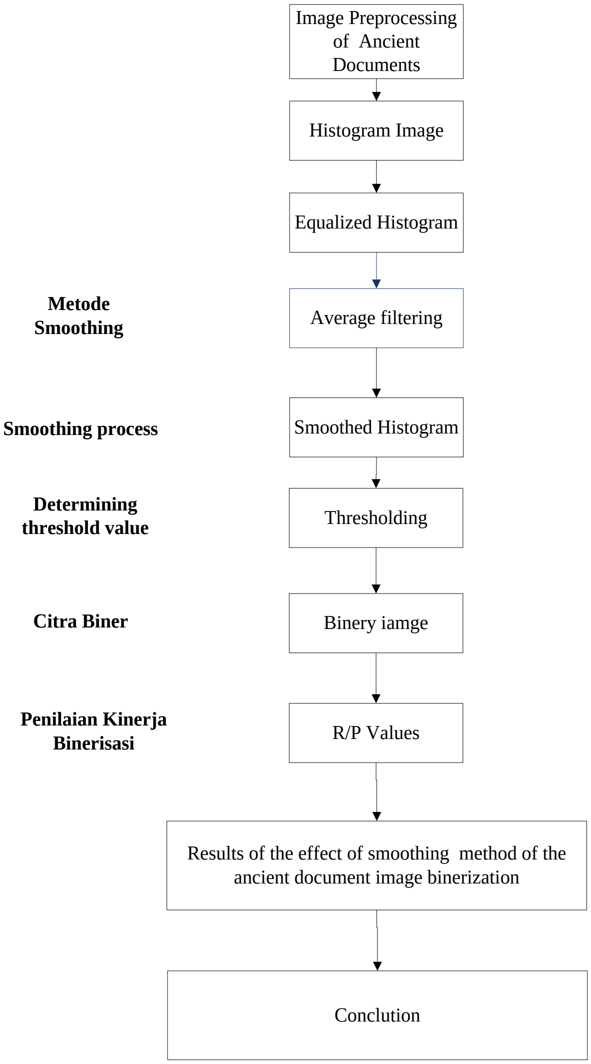


Image Preprocessing of Ancient Documents
Histogram Image
Equalized Histogram
Metode Smoothing
Average filtering
Smoothing process
Smoothed Histogram
Determining threshold value
Thresholding
Citra Biner
Binery iamge
Penilaian Kinerja Binerisasi
R/P Values
Results of the effect of smoothing method of the ancient document image binerization
Conclution

Figure 1. Flow of Implementation Method

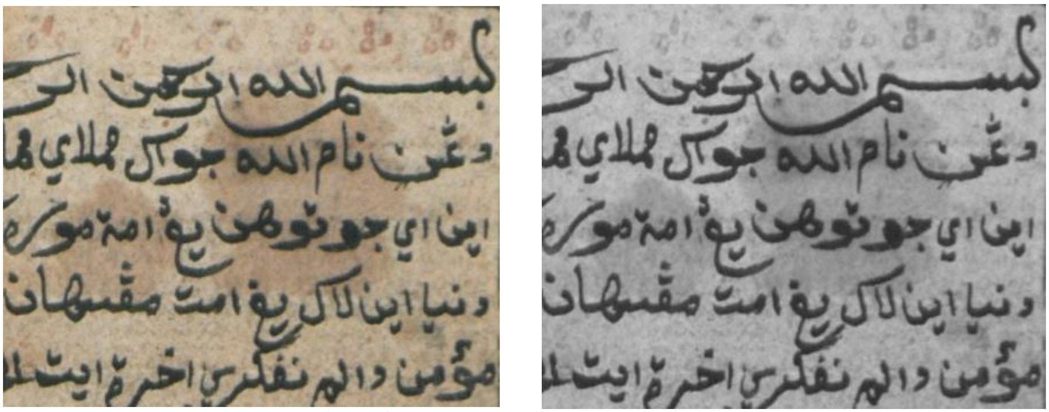
## Image Preprocessing

Pre-precessing is the processing of image data for further analysis. Pre-processing includes mostly is by changing the colored image (RGB) into a grayscale image. Grayscaling is served to simplify an image model to make the image easier to be processed

In general, to generate grayscale image the following formula is written:

$$S = \frac{r + g + b}{3} \qquad (1)$$

Where S is the grayscale image by searching for the mean of each layer of r (red), g (green), and b (blue). Below is achange image from RGB image to grayscale image.



a. Colour Image(RGB)

b. Grayscale Image

Figure 2. Changes of RGB image to grayscale

## Histogram Equalization and Normalization

A histogram in a digital image is a graph that represents the color distribution of a digital image showing the intensity of pixel values of an image. The mapping of pixel values in the histogram is as follows [1]:

$$h(n_k) = n_k \qquad (2)$$

Where $n_k$is the axis denoting the pixel value (k = 1-255) and $h(n_k)$ is the ordinate representing the number of pixels for each pixel.

In this study the histogram in an image should be normalized before the histogram equalization process. The benefits of histogram normalization is to see statistics of an image divided by the total number of pixels in the image. Normalization of histogram can be defined as below [1]:

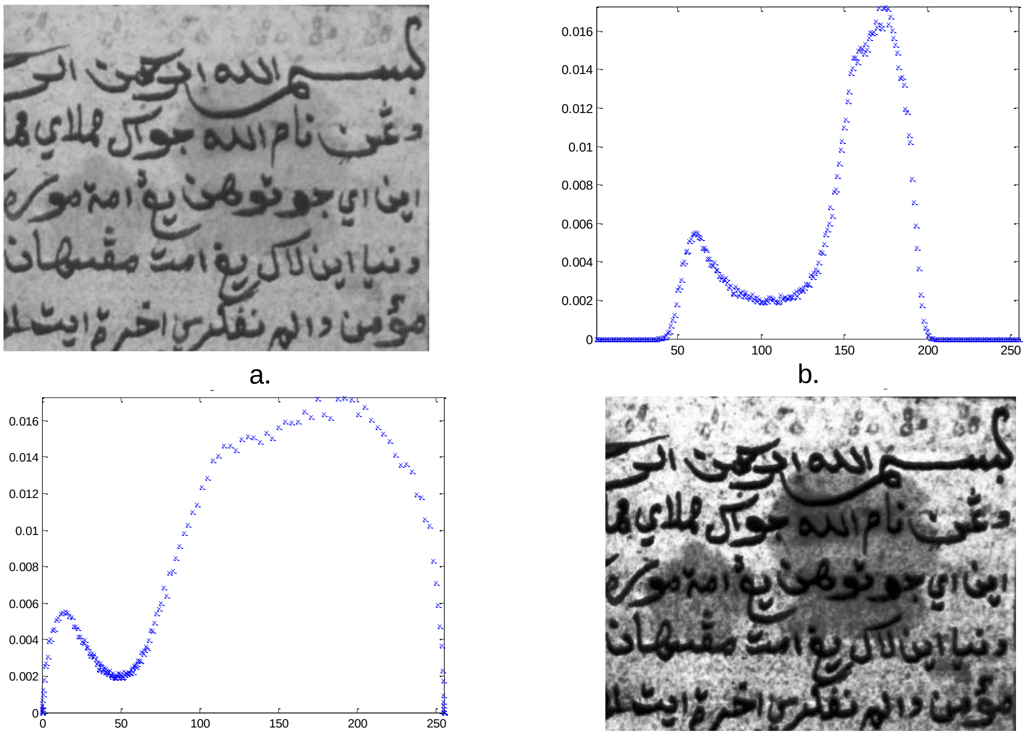
$$p(n_k) = \frac{h(n_k)}{n} = \frac{n_k}{n} \qquad (3)$$

Histogram equalization is an image enhancement technique by manipulating each image pixel in which the spread of the original image histogram is not evenly distributed because the pixel distributiondoes not keep the entire level of gradation available on histogram [1]. This process results in pixel values evenly distributed in the interval (0-255).

Histogram is equalized mathematically and can be performed with the following equation [1]:

$$T(n_k) = \sum_{j=1}^{k} p(n_k) \qquad (4)$$

Where $p(n_k)$ is the ordinate that states the number of pixels for each pixel. While $T(n_k)$ is the location where the $n_k$intensity value will be mapped.



a.

b.

c.

d.

Figure 3. (a) Document image, (b) Image histogram, (c) Equalized histogram, (d) Imagery of histogram equalization

## Average filtering

Average filtering is filter which issearching foraverage values of data set [5]. The formula of calculating Average filtering is as follows:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (5)$$

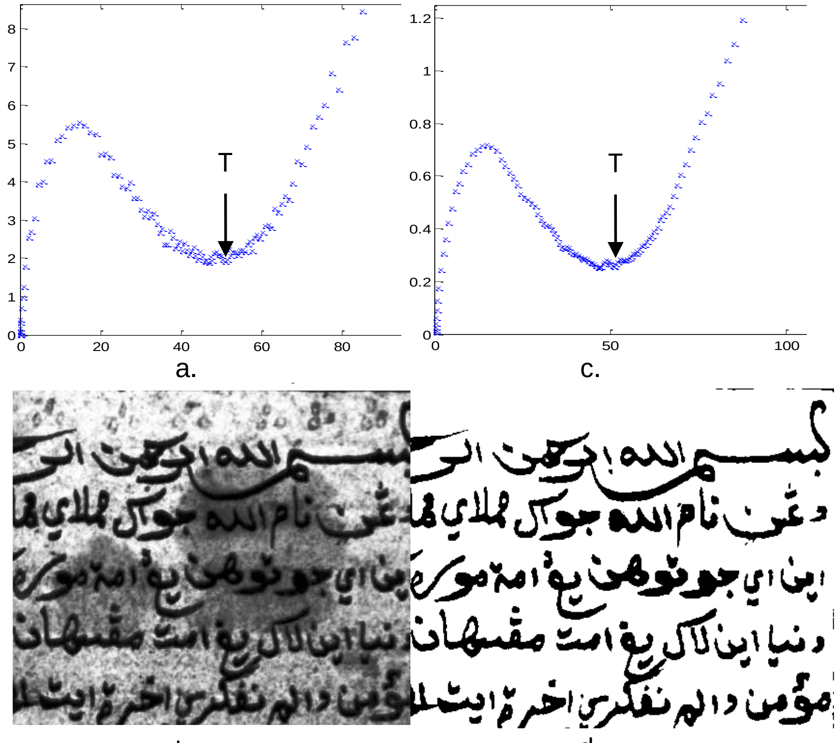
Where $\overline{X}$ is the average, $n$ is the number of data, $x_i$is $i$ value and $i$ is the initial value.

## Thresholding dan Binerisasi

Thresholding or determining the threshold value is the process of separating pixels according to the degree of gray they have. [7] The threshold value of the histogram represents the object and the background. The provisions in determining the threshold values are as follows.

$$g(x,y) = \begin{cases} 1 & \text{if } f(x,y) > T \\ 0 & \text{if } f(x,y) \leq T \end{cases}$$

Where *g (x, y)* is the image segment backround, T is threshold value, and *f(x, y)* is the image dimension. If *f*(*x*,*y*) >*T* then it is called background, if *f*(*x*,*y*) ≤ *T* then it is called object or foreground.



a. c.

b. d.

Figure 6. (a) equalization histogram (b) the image of equalization histogram, (c) the smoothed histogram, (d) binarization result

*Materials*

The imagesof ancient documents used were ancient documentsinscribed with Arabic that had been digitized and had been pre-processed i.e changed it in grayscale to facilitate the binarization process. Images used were in the format of ".tif" which ahve dimensions of 1320x2000 pixels per image. The images of ancient documents used in this study were 10 images with characteristics of 5 low-noise images and 5 high-noise images. Software testing used in this study was MATLAB application.

*Methods*

The method used to test readability value using Recall and Precision.Recall parameters is the size of the number of relevant documents retrieved from document set at the time the query is applied. While precision is a measure of the accuracy or relevance of query results. A application of method research is by counting all characters of readable texts and unreadable texts. With these parameters,the precentage of readability value before and after the proposed method appliedwill be obtained.

Recall and precision in this study were used to evaluate retrival of text characters against applied methods by measuring the number of relevant and irrelevant characters. Recall and precision are typically rated in percentages of 1 to 100%. High recall value means little false negative and high precision value means little false positive. Recall and precision can be calculated by the following equation.

$$Recall = \frac{NCD}{GT} \quad (7)$$

$$Precision = \frac{NCD}{TR} \quad (8)$$

Where NCD is the correct number of characters detected in the binarization result document. GT is the total number of characters contained in the original document. And TR is the number of characters detected in the binarization result document including the correct and damaged characters.

The GT (Ground-truth) of the document image is searched manually by counting the number of characters read and the damaged characters in the original document image. The detection of NCD and TR by following GT (Ground-truth) [13].

## RESULTS

Ancient document images produce a histogram which is a representation of the appeared color intensity. An equalized histogram has unfavorable image curve that causes the determination of the threshold value to be very difficult. Histogram smoothing method is required so thresholding can be done. In this research method used in smoothing curve at local minima curve is Average filtering. Image histogram before and after average filtering performed can be seen in figure 7.

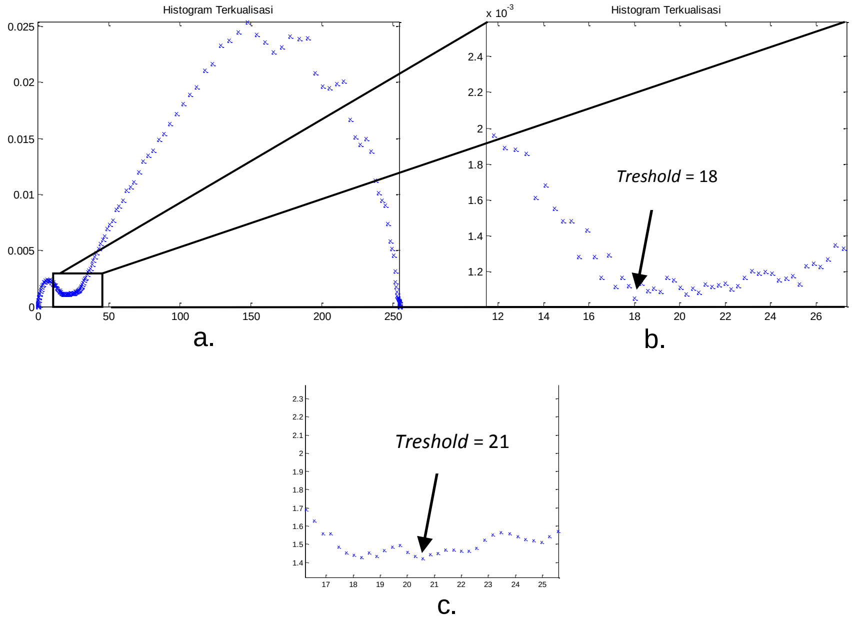


Figure 7. (a) Equalized histogram, (b) Enlarged Histogram, (c) H. Average filtering

From the above smoothing histogram, we get different threshold values before and after filtering process. To find out the results of threshold values generated by some ancient document images, it can be seen in table 1 and table 2. Table 1 shows some images that have low noise with difference before and after smoothing process performed. While table 2 shows images that has high noise qualification and show the difference before and after the smoothing process performed.

Table 1. The Results of Threshold ValuesOf Some Low-Noise Images

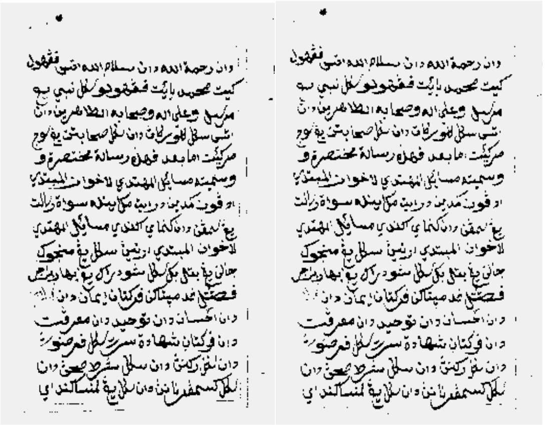
| File | Histeq | Average |
|---|---|---|
| 2l.tif | 18 | 21 |
| 2r.tif | 14 | 14 |
| 3l.tif | 28 | 29 |
| 3r.tif | 26 | 25 |
| 4l.tif | 26 | 26 |

Table 2. Results of Threshold Values From Some High-Noise Images

| File | Histeq | Average |
|---|---|---|
| 1l.tif | 26 | 24 |
| 20l.tif | 30 | 29 |
| 23l.tif | 14 | 15 |

| | | |
|---|---|---|
| 23r.tif | 18 | 18 |
| 24l.tif | 13 | 13 |

From Table 1 and Table 2 above there are several variations of threshold values of each method. If we look carefully at the result of determination of threshold values in each method is not much different from the result of the threshold values in the histogram. But from the side of the binarization results it becomes important because one number on the different threshold value will affect the number of successfully restored characters.From the binarization results there are differences that on**figure 8.a** noises contained in the image are less than in image of **figure 8.b**



a. b.

Figure 8. Binarization image difference (a) "21" threshold value and (b) "18" threshold value

To find out the binarization performance or readability values of documents in the ancient documents' images above, this research uses Recall and Precision parameters to find out howmuch the successful percentage is in the restoration of characters contained in those images. Here are some Recall and Precision results.

Table 3. table of recall and precision of low-noise images

| Image Name | Recall and Precision |
|---|---|
| | Average Filtering |
| 2l.tif | 98.81% |
| 2r.tif | 97.78% |
| 3l.tif | 99.08% |
| 3r.tif | 99.52% |
| 4l.tif | 99.28% |

Table 4. table of recall and precision of high-noise images

| Image | Average Filtering |
|---|---|

| name | Recall | Precision |
|---|---|---|
| 1l.tif | 94.50% | 96.30% |
| 20l.tif | 97.51% | 98.73% |
| 23l.tif | 90.93% | 97.34% |
| 23r.tif | 97.75% | 99.54% |
| 24l.tif | 68.73% | 98.31% |

## DISCUSSION

The determination of threshold values on histogram has constraints on unsmoothed histogram curves. Using the smoothing histogram method in this study has proven very helpful. Average Filtering successfully refines the histogram and clarifies curves so that the lowest points (local minima) look more clearly. The result of binarization on each method is known its difference after Recall and Precision calculations. Recall and Precision count how many characters are successfully restored and the damaged characters are compared to the characters in original images.

## CONCLUSION

From the table of determination of threshold values, the different valuescan be seen. The determination values of threshold values before and after the smoothing histogram process look very much different. However, in binarization results the ancient documents'imagesresult in slightly different readability values. This depends on the level of noise or condition of the ancient document images.

## REFERENCES

[1] F. Arnia dan K. Munadi, "Metode Restorasi Citra Manuskrip Kuno Berbasis Histogram Terekualisasi", Seminar Nasional Teknologi Informasi, hal. A12, 59-63, 2008.

[2] R. C. Gonzalez dan R. E. Woods, *Digital Image Processing,* 2nd Ed. Practice Hall, 2002.

[3] R. Munir. "Pengolahan Citra Digital dengan Pendekatan Algoritma" Bandung, Penerbit Informatika, 2004.

[4] J. Utama, "Akuisisi Citra Digital Menggunakan Pemrograman Matlab", Jurnal Majalan Ilmiah UNICOM, Vol. 9, No.1, 2011.

[5] B. Yuwono, "Image Smoothing Menggunakan Mean Filtering, Media Filtering, Modus Filtering dan Gaussian Filtering", Jurnal UPN Veteran Yokyakarta, Vol 7, No.1, 2010.

[6] R. S. Lasijo, "Fitting Kurva dengan Menggunakan Spline Kubik", INTEGRAL, vol. 6, no. 2, 2010.

[7] E. Kavallieratou dan H. Antonopoulou, "Cleaning and Enhancing Historical Image", LNCS3708, pp. 681-688, 2005