# FILTERING AND WAVELET TRANSFORM ALGORITHM FOR OLD DOCUMENT IMAGE RESTORATION

**Ridha Sefina Samosir**

Information System Study Program, Creative Industry, Institute Technology and Business Kalbis
Jln. Pulomas Selatan Kav.22 Jakarta Timur 13210, Indonesia
ridha.samosir@kalbis.ac.id

**Received:** 11th September 2017/ **Revised:** 18th September 2017/ **Accepted:** 27th September 2017

***Abstract*** **-** The aim of this research was to develop image restoration system using filtering and wavelet transform algorithm. Data collection was through observation and system was developed using prototyping model. Result of this research is a computer based on system to restore image containing noise. Based on the research process, filtering and wavelet transform algorithm can used to restore old document image from interferences (noise).

***Keywords:*** old document, image restoration, filtering, wavelet transform

## I. INTRODUCTION

Old document is one of the historical heritage in a nation or state. There is much valuable information that can be extracted from an old document. Some document image comes from different source or place but documents may have a relation or link with each other. It means that many information and the relation among information can be mined to historical heritage. As a relic of history, the main problem of the old document is the appearance of some interference (noise) strokes. It is difficult to read and understand all the content. Interference can be caused by many things such as the storage time, storage media, storage methods, materials paper and ink used, and image apturing technique. One type of interferences that often occurs in the old document image is ink bleed through removal. Ink bleed-through removal is the appearance of various marks or signs that affect the quality of the document such as printed paper from the back side of the document appearing on the front side of the document. Moreover, ink widening is a lack of the ability of the paper to absorb the ink. A sign appears as a result of the digitalization process. In particular, the issue of the appearance of printed paper on the back side of the document is more common in italics documents type. If the gradient is approximately 45, usually printed paper from the other side has a slope of approximately 135. Because of these interferences, much information in the document cannot be recognized.

Indonesia is a country with many cultural heritages from 17.000 islands. Each island consists of many tribes and various languages. One of the cultural heritages is history. Historical stories are poured on various media such as temples or paper. Moreover, writing is a technique that can be used to tell about history. A collection of Indonesian historical heritage is stored in Arsip Nasional Republik Indonesia (ANRI - National Archives of Indonesia). This research uses a collection of document obtained from ANRI and all of documents represent italic hand writing type. From the document obtained by researcher from ANRI, it shows some significant damage. The existing damage includes the appearance of writing from the back side on the front side, ink stains on the document, and unclean document background. This is very unfortunate because many historical stories can be extracted from the document. In addition, there is possible relevance among the documents. Thus, many information and historical events are lost. Figure 1 is an example of the input image.

According to Huiyu, Jiahua, and Jianguo (2010), the technique to minimize or eliminate interference in the old document image is one of the processes in image processing, namely restoration. Restoration is one of the operations in the image processing system aiming to improve the image quality degraded or degradation. Old document image restoration aims to minimize the interference. The beginning stage of the restoration process is digitalization of the old document to the digital image (Rafael & Richard, 2008).

Figure 1 Example of Input Image

There are many methods to perform image restoration. There are classification, serialization, filtering, and wavelet transform. Konidaris, Kesidis, and Gatos (2016) said that digital image classification techniques divided the old documents into three parts. Those were background, original text, and interfering stroke. Through the classification techniques, the front side of the document was extracted to produce original text. Moreover, Hinami *et al.* (2016) used

serialization techniques. This algorithm applied the principle of sequentiality like sliding windows for the entire image. Another previous research uses a thresholding technique with Otsu threshold methods. The principle of this technique is the extraction of text characters that contain background noise based on their gray level distribution (2013). This algorithm is suitable for the degraded document image. If the image size is large, it will cause the gray level to overlap between foreground and background document. Meanwhile, Another algorithm that has been done is the K-Means Singular Value Decomposition (K-SVD) algorithm or by Ren, Lu, and Zeng (2015). The principle of this algorithm is to train a dictionary that represents the semantic structure of the image based on the library of the original image. The main idea of K-SVD algorithm optimization is by updating and adjusting elements in the dictionary continuously until it matches with the image signal people want.

Different from the K-SVD algorithm, ant colony and genetic algorithm are an algorithm which is a combination of two bionic evolutionary algorithms (Gülcü *et al.*, 2016). Ant Colony Optimization (ACO) is adopted from the behavior of ant colonies or the ant systems. Ants can find the shortest route from the nest to the location of food based on footprints on the trajectory that has been passed. The path passed by large ant will be followed by the other ants. It will increase the density of the ants that pass through it, or all the ants will pass that path. This is because of the nature of ants that produce pheromone substances. Such substances can only be identified by similar living things. If many ants cross a path, the substance will be more and more. However, if a track is rarely passed, the substance will be lost. In image restoration, the ant colony algorithm can easily generate the behavior of the image signal. However, ant colony algorithm requires a long search time that also increases the rate of convergence. On the other hand, genetic algorithms conduct searches with random existing techniques to speed up multidimensional nonlinear data computation (Feng, Lu, and Zeng, 2015).

The methods in previous research suggest that the result will be less optimal if the writing on the front side of the document contains more than one color and the image of an old document has many noises (degraded document image). Aside from these two problems, the main problem of using classification techniques and K-SVD is that the system requires supervision or involvement from the users to determine the amount of area (cluster) to be formed and a color sample from each class (cluster) that cannot be done automatically. Meanwhile, the combination of ant colony algorithm and genetic algorithm is more suitable for the restoration process that clarifies the edges and textures of the image.

With the various weaknesses and drawbacks from the application of the algorithm for the document images restoration, this research proposes to combine both of wavelet transform and filtering approach. Therefore, the combination of these two algorithms is expected to provide a solution to improve the quality of damaged old document image. Then, people can read the information contained in it.

## II. METHODS

This research starts from the problems that it is difficult to recognize the contents or text written on many old documents. From 39 documents image obtained from ANRI, the researchers classify the document based on the type of noise subjectively. There are four types of disorders. Those are damage caused by noise in the background document; noise in the form of ink widening and ink splashes from unwanted scratches; damage in the form of printed paper from the back side appearing on the front side of the document; and Damage caused by fail digitalization process.

The instrument research used is two algorithms which are wavelet transformation and filtering. This research proposes to use multi directional wavelet transform algorithm and mean shift filtering algorithm. Then, the software for the system development is Matlab R2008. The principle of the mean shift filtering algorithm is to apply mean shift algorithm to filter the data of the image. The mean shift filter algorithm works iteratively and generates a set of neighborhood pixels (M) based on the spatial radius (hs/spatial kernel bandwidth) and color distance (hr/range or color kernel bandwidth) values. In each set of neighborhood pixels, mean of its spatial and color distance is calculated.

The obtained mean value is the new starting position for the next iteration. This iteration procedure will stop when the mean values of spatial and color distance do not change from the previous iteration. Next, wavelet transformation proposed in this research is multi-directional wavelet transform. With multi-directional wavelet transform algorithm, all posts from various directions can be well identified. Thus, it is easier to process the restoration.

The steps of the research are shown in Figure 2. First, data are collected. Data obtained from ANRI are in the form of an old document image. The researcher also uses literature study from various sources to find out the most appropriate solutions to the problems. The problem statement of this research is how to implement both filtering algorithm and wavelet transformation to perform old document image restoration. Second, it is the design of the application to be built. The design phase is designing navigational structure and graphic interface of the application. Third, there is the development of the applications. Based on the results of the first step, it indicates the algorithm that can be used is a combination of filtering algorithm and wavelet transformation. Filtering is a process of taking the partial signal of a certain frequency and discards the signal on the other frequencies. Filtering the image also uses the same principle which takes the function of the image at a certain frequency and discards the image function at certain frequencies as well. The frequency of the image is affected by the existing color gradation on the image. Image with gradation (threshold level) tends to have a lower frequency and vice versa (Trieu & Maruyama, 2015).

The working principle of a filtering algorithm is divided. First, if people want to maintain the gradation (the number of color levels in an image), the pixels have to be maintained at a low frequency and eliminate pixels at high frequencies (Low Pass Filtering). Second, people will get the image of a certain threshold value or the binary image pixels at high frequency and low frequency maintained is discarded (High Pass Filtering). Third, if people want to maintain gradation, they should reduce the frequency field (bandwidth) and discard unnecessary signal the low frequency, and maintain the high frequency. Then, the mid-frequency is discarded (Band Stop Filtering).

Furthermore, wavelet transformation algorithm works through signal analysis using wavelet function to produce wavelet coefficients. Transformation means a change action that is usually done to help simplify the

problem. Meanwhile, the image is a two-dimensional image on a field that contains much information. Therefore, the image transformation means the process of changing the form of the image to explore or get the information contained in the image and the information is used to solve the occurring problems with the image.

Wavelet is a mathematical function fulfilling the requirements that can be used to represent a signal. Wavelet comes from the word 'wave' and 'let' which means little. Briefly, wavelet can be interpreted as a small wave. Small waves are translated as scale. Therefore, a wavelet is used to analyze the data or functions based on the scale. To ease the process of decomposition of directional wavelet transform algorithm, input images are arranged in a dyadic matrix with a pixel size 2n. If the result of the image is not symmetrical, it can use the zero extension or expansion of the matrix by adding a value of 0 in the row or column. Wavelet function will be divided into the signal components of different frequencies. Then, the frequency components are analyzed using a scale of resolution (scale function) by Hou *et al.* (2013).

From the explanation, a wavelet transform is a tool that can be used to analyze non-stationary signals (frequency content of the signal which varies with time). Wavelet analysis can show the temporal behavior of the signal, filter (filtering) data, and signals, and eliminate unwanted behaviors of signals for image compression. The most important properties of wavelet are the localization of time and frequency so that the analysis of the signals is done locally and detail according to the scales. In other words, wavelet-based analysis splits the signal into several different frequencies, the approximation (A/lowpass) and a detailed section (D/highpass). The approximation is the components of the low-frequency signal while the detail is high-frequency signal components. The detail part consists of horizontal, vertical, and diagonal detail. Low-frequency signal components (approximation) indicates the identity of a high-frequency signal and the nuances/details of the signal (Wang, 2010). Figure 3 is an illustration of decomposition process signals with wavelet transform.

Fourth, it tests the application. Once the application is built, the application is tested using 39 ANRI input images of the institution. To analyze test results easier, 39 input images are divided into four categories based on the type of interference. Last, the researcher can conclude the test from the result.

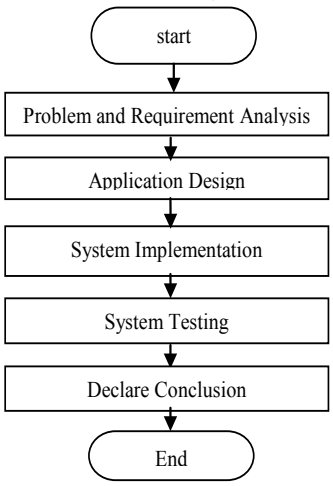


Figure 2 System Development Flowchart

| | | |
|---|---|---|
| LL2 | HL2 | HL1 |
| LH2 | HH2 | |
| LH1 | HH1 | HH1 |

Figure 3 Image Decomposition Using Wavelet

## III. RESULTS AND DISCUSSIONS

Several stages are done in the system development. First, the researcher uploads the input image. The input image is converted from RGB color space to L.U.V. Then, L.U.V color space is represented in the form of data points. This is the pseudocode to generate an initial image and convert an image from RGB color space to LUV color space. Second, pixel input image is subjected without downsampling the advanced wavelet transform to separate the image based on a different frequency in the wavelet domain. Advanced wavelet transformation generates four coefficients (Sub-band). The four coefficients are the approximation coefficient, the detailed horizontal coefficient, vertical detail coefficients, and diagonal detail coefficient. Third, it is followed by a convolution process on the component horizontally, diagonally and vertically. Convolution process filters the signal between the input signals with the impulse response of the filter. The coefficient that acts as the input signal is a horizontal coefficient, vertical coefficient, and diagonal coefficient. Meanwhile, the coefficient of the response impulse is a matrix that represents the direction (direction) specific.

$$\phi_{j,m,n}(x,y) = \varnothing_{j,m}(x_1)\,\varnothing_{j,n}(y_1) \quad (1)$$

$$\Psi^{1}_{j,m,n}(x,y) = \Psi_{j,m}(x_1)\,\varnothing_{j,n}(y_1) \quad (2)$$

$$\Psi^{2}_{j,m,n}(x,y) = \varnothing_{j,m}(x_1)\,\Psi_{j,n}(y_1) \quad (3)$$

$$\Psi^{3}_{j.m,n}(x,y) = \Psi_{j,m}(x_1)\,\Psi_{j,n}(y_1) \quad (4)$$

$$\phi_{j,m}(x) = x/\sqrt{(2^j)}\ \ \phi\,(x/2^j - m) \quad (5)$$

$$\Psi_{j,m}(x) = 1/\sqrt{(2^j)}\ \Psi\,(x/2^j - m) \quad (6)$$

Multi-directional wavelet transforms, horizontal, vertical, and diagonal coefficients are convolved with a matrix representing a certain direction (orientation). The matrix equation is as follows.

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} \quad (7)$$

$$A = c\begin{pmatrix} \cos\varnothing & \sin\varnothing \\ -\sin\varnothing & \cos\varnothing \end{pmatrix} \quad (8)$$

Value of C= $\sqrt{2}$ and Ø represent the direction (orientation/directional) of the posts on the document image. This research proposes the value of Ø as a combination from 0o to 3600. In the process of this convolution, a filtering algorithm is done by applying standard procedures

to obtain the mean shift convergent circumstances. The filtering process begins with the initialization of each pixel of the input image. Then, it proceeds to mean shift standard process until the convergent state is obtained. The calculation of the mean shift vector kernel involves spatial bandwidth and range/color kernel bandwidth.

$$m_{h,g}(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} \quad (9)$$

The convergent state is achieved if the mean value of the current shift vector is not equal to the previous mean shift vector. The next step is the storing pixels ($Yi^s$, $Yi^r$) when the convergent state had reached a certain point. Pixel $Zi = (Xi^s, Yi^r)$ is stored as a filter output value. Filter output value is filtered pixels from the surrounding pixels. Filter output value represents pixels output.

$$K(x) = (\|x^s\ x^r\|) = \frac{c}{h_s h_r} k\left(\frac{x^s}{h^r}\right) k\left(\left\|\frac{x^s}{h^r}\right\|^2\right) (10) \quad (10)$$

Next process is thresholding the output value filter. Thresholding an additional operation is performed to improve the image of the output result. This process means determining thresholding parameter values that suit the image quality output the best. After the reconstruction image is obtained through the filtering process, the data points are reconstructed back into the RGB color space.

The result of this research is an application for image restoration using both mean shift filtering and multi directional wavelet transform algorithm. Figure 4 is a graphical user interface of the developed system. The system is developed with Matlab version 2008. The system consists of one menu. The user can upload an input image, set the parameter value of each algorithm, and display the output image. Moreover, Figure 5 is input and output of old document images after filtering and wavelet transformation implementation.

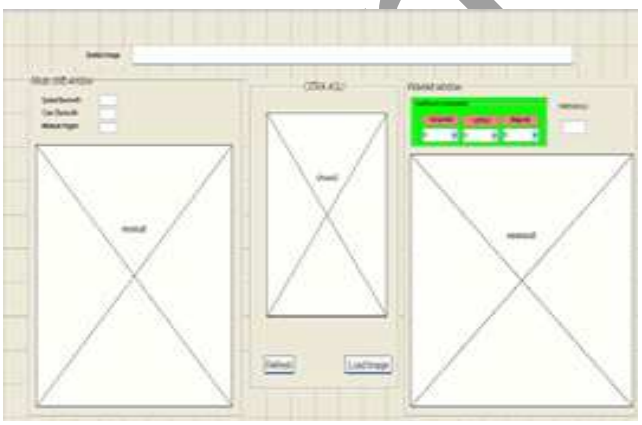

Figure 4 Graphical User Interface of Restoration Image System For Old Document

From the output image, it shows that the result of image restoration is very good. Some noise can be minimized, so it gives some impact such as the cleaner background of the paper and printed text from the back side of the paper. According to the wavelet transformation algorithm, it conducts analysis technique with the image signal. It aims to explore the properties of the image signal. Properties of the image are used to filter and eliminate unnecessary signals. By adding multi-directional approach in wavelet transformation, it can analyze image signal from any direction. Filtering algorithms can eliminate unwanted image frequency and maintain the desired image frequency. In addition, the results show that the parameter values in both algorithms determine the quality of the image restoration. This becomes the advantage of the built system that the users can set the parameter values that match the conditions of the input image the best.

Figure 5 Filtering Output Image

Figure 6 Wavelet Transform Output Image

Then, the input parameters of wavelet transformation algorithm are a combination of the value of θ and the threshold value of the third detailed coefficient factors. The results of tests performed by the filtering algorithm and wavelet transformation algorithm are as follows. First, it is for filtering algorithm. There are some characteristics of the output image based on the parameter value. It has the larger value of hs, and the occurrences level of noise in the background is reduced. It also has a larger value of hr, so the writing becomes unclear. The writing is difficult to be identified. The smaller value of hs causes the appearance level of noise increases. Then, the smaller value of hr causes noise cluster to increase, but the large cluster is smaller. M value influences the quality of image restoration results less significantly. Based on experiment result, the value of the input parameter (hs, hr, M) that provides the best results in the output image is 10, 7, and 20 respectively. After

experiment for all of input image, this parameter value give better result for document image with noise like interference strokes in the document background, paper printed in the front side that appear from backside and noice that appear because digitation process. But its showed not sigificant result for document image with widening ink.

Next, there is some characteristic of the output image based on the wavelet transformation parameter value. If the value of θ at detail coefficients is 15°, 60°, 90°, 135°, 180°, and 360°, the quality of output image is not good. This is because all six grades of θ represent the direction from the back side of a document which is interfering strokes. However, if the value of θ at detail coefficient is 0°, 30°, 45°, 120°, 225°, 270°, and 315°, the quality of output image will be better. The threshold factor is inversely proportional to the quality of the output image. The greater the threshold factor is, the worse the resulting output is. It is because the distribution of wavelet coefficients is centralized at wavelet coefficient value 0 and vice versa. Value of wavelet input parameter is tried for all of input image. From the experiment show that multi directional wavelet transform give good result for document image because interference strokes in document background, paper printed from back side that appear in front side and fail digitalization process.

## IV. CONCLUSIONS

From experiment, it shows that quality of the output image is influenced by the accuracy of input parameter values in the algorithm. Kernel bandwidth (hs), range/ color kernel bandwidth (hr) and M (Minimum Region) are for filtering algorithm. Mean while, θ and threshold are for wavelet transformation. Both filtering and wavelet transformation show optimal performance for interference strokes like printed paper from document backside that appears in the document front side and noise because of digitalization process faulty. Then, the less optimal result is for document image which is ink widening or splash or blobs.

## REFERENCES


Feng, Y., Lu, H., & Zeng, X. (2015). Image restoration based on hybrid ant colony algorithm. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 13*(4), 1298-1304.

Gülcü, Ş., Mahi, M., Baykan, Ö. K., & Kodaz, H. (2016). A parallel cooperative hybrid method based on ant colony optimization and 3-Opt algorithm for solving traveling salesman problem. *Soft Computing, 20*(11), 1-17.

Hetal, J.V., & Astha, B. (2013). A Review on Otsu image segmentation algorithm. *IJARCET (International Journal of Advanced Research in Computer Engineering & Technology), 2*(2), 387-389

Hinami, R., Liu, X., Chiba, N., & Satoh, S. I. (2016). Bidirectional extraction and recognition of scene text with layout consistency. *International Journal on Document Analysis and Recognition (IJDAR), 19*(2), 83-98.

Hou, X., Yang, J., Jiang, G., & Qian, X. (2013). Complex SAR image compression based on directional lifting wavelet transform with high clustering capability. *IEEE Transactions on Geoscience and Remote Sensing, 51*(1), 527-538.

Huiyu, Z., Jiahua, W., & Jianguo. Z (2010). *Digital image processing: Part 1*. Retrieved from www.bookboon.com

Konidaris, T., Kesidis, A. L., & Gatos, B. (2016). A segmentation-free word spotting method for historical printed documents. *Pattern Analysis and Applications, 19*(4), 963-976.

Rafael, C. G., & Richard, E. W. (2008). *Digital image processing*. USA: Addison-Wesley.

Ren, J., Lu, H., & Zeng, X. (2015). Image Denoising Based on K-means Singular Value Decomposition. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 13*(4), 1312-1318.

Trieu, D. B. K., & Maruyama, T. (2015). Real-time color image segmentation based on mean shift algorithm using an FPGA. *Journal of Real-Time Image Processing, 10*(2), 345-356.

Wang, X. (2010). Recovery of blurring scanned manuscript image based on wavelets transform algorithm. In *3rd International Congress on Image and Signal Processing (CISP)* (pp. 844-847). IEEE.