# Simulation System for Document Image Restoration Using Mean Shift Filtering and Multi Directional Wavelet Transform

**Ridha Sefina Samosir**
Department of Information System, Faculty of Computer Science and Communication, Institut Teknologi dan Bisnis Kalbe

**Keywords:**

Documents
Restoration
Mean Shift Filtering
Multi Directional Wavelet Transform

**ABSTRACT**

Indonesia is known as a country that has a lot of heritage, some of them are old historical documents. Each of the old document has meaning and connection towards another documents. Unfortunately, due to the age of the documents, an error in the storage system and quality of the paper or ink cause the documents broken or damaged. The way to solve the problem is using image restoration method like mean shift filtering and multi directional wavelet transform. On one hand, the procedure of mean shift filtering is finding modes from image data set distribution, on the other hand, directional wavelet transform procedure able to exploit the directional property of the strokes which separated the foreground strokes and the interference mainly in different wafelet frequency. Based on the perfomance of each method to restore the damaged document image, this research propose to make a simulation for image restoration using both of these methods. Through this, user can choose which method will give the better output to restore based on the damaged of old document image. The study for this research is an experimental because this research do investigation for the damaged document image. Development and testing of the research output.



***Corresponding Author:***

Ridha Sefina Samosir,
Department of Information System, Faculty of Computer Science and Communication,
Institut Teknologi dan Bisnis Kalbe,
Jalan Pulomas Selatan, Kav.22, Kayu Putih, Jakarta Timur, Indonesia.
Email: ridha.samosir@kalbis.ac.id dan defa.dmk@gmail.com

## 1. INTRODUCTION

Indonesia is a country with abundant numbers of historical and cultural heritage. The most common historical and cultural heritage of Indonesia can be seen through temples, statues, old inscription including some old documents. The documents are prieceless and consist of various meaning and yet they ussually interconected with other old documents. According Viscount St. Davids (1993), Old documents are precious heritage which tell event or moment of earlier time that has probably not been revealed yet. The old documents are one of the important sources of national historical evidence that a nation can realize how the situation of its nation in the past. old document which tells the story of the past is an important legacy.

Old documents stored too long in storage media without any good treatment causing documents to be damaged. In addition to the storage and retrieval, quality factor of materials such as paper and ink of the documents also be another factor of the damage or disruption to the document called *noise*. One type of *noise* that often occurs in old documents is ink bleed removal. Ink bleed removal is a matter of the emergence of a variety of signs and graffiti that affect print quality documents such as the writing that appears on the back side of the document, the ink is widening, it is a sign of the error due to the digitization process and so on. The emergence of a variety of marks on the document causes the old content of the document to be difficult to read and recognize. There are so many ways to improve the quality of the image that has degradation, one of them is using image processing techniques which leads to the methods of image restoration. Imagery is one of the restoration process in image processing which aims to improve the image quality decreased quality or interference (*noise*) [1]. In order to get the restoration of the image from the old documents, the object of the documents should be digitized to produce the image / picture. The benefit of this technique image restoration is that we can avoid the physical contact against old recordings because it works with old

recordings which been digitized (imagery) so that the quality the old document does not increasingly declining. There are 39 image of old documents that will be restored in this study and those documents derived from the National Archives of the Republic of Indonesia (ANRI). Characteristic of the old document is a flowing handwriting font style with a certain incline degree. These type of noise that appeared in the documents grouped into 4 categories which some of the documents included to more than one kind of noise category, namely:

1. Amount of *noise* in the background documents. There are 12 documents that take account in this category.
2. *Noise* in the form of a widening ink, ink set is a result of ink splashes. There are 12 documents that take account in this category.
3. *Noise* arising because of digitizing errors. There are 12 documents that take account in this category.
4. *Noise* in the form of printed paper from the back side of the document that appears on the front side of the document. There are 12 documents that take account in this category.

Figure 1. Ink Bleed Removal Document Image (National Archives of the Republic of Indonesia)

There are so many methods of image restoration has been done in the previous studies, such as K-means algorithm by classification techniques, K-means algorithm with serialization techniques, and K-means algorithm by recursive application techniques (PCA). The working of the K-means clustering algorithm through a classification technique is to divide the document image into 3 classes, namely class background, original text, and interfering strokes. Distribution of this document aims to extract the text on the front side of the document in order to obtain the original text and how to work with the K-means algorithm is serializing it by using the principle of sliding window on the entire image. Whereas the K-means technique PCA is generated through a binary tree where only part logaritmic historgram satisfying condition leaves are processed [2]. Nevertheless, these three algorithm are less effective if the last posts on the front side of the document containing more than 1 color and imagery of the old document restored contains too much *noise*. Besides that the main problem of using K-means classification technique is that the system requires supervision or involvement of the user to determine the number of areas (clusters) that will be formed and the color sample from each class (cluster) is not done automatically. In addition to K-means clustering algorithm, there are other algorithms that can be used for image restoration using the Negishi Tresholding binary techniques. This algorithm is used to extract characters from text image containing background *noise* (noisy background). This algorithm is suitable for image and document that has complex *noise* (too degrade document image) and large in size, but it is very possible that can lead to a state of gray level characters overlap between the foreground and the background of the document section [3]. The failure of this algorithm could be undertaken using the method of morphological techniques to extract text with the condition overlap with the back side of the document [4], however, both of these algorithms were not able to restore the image of the old document effectively.

Based on a literature review, number of methods that have been mentioned are still showing limitations in restoring the image of old documents, especially for the 4 types of *noise* like the previous explanation. Therefore, to overcome the limitations of some of these methods, we need an algorithm that

powerful (robust), and this algorithm can be used to restore the image of an old document with specifications typeface and type of *noise* like the previous explanation. For those purpose, the researcher proposed algorithm; mean shift filtering. This algorithm is an algorithm that applies techniques of finding modes (local maxima), classifying a set of data points from the document image into several groups with great accuracy [5]. This algorithm is suitable for the case of the data points that are not in order (number and irregular distribution of sample data), therefore, it works according to the type of *noise* or damage to the image of an old document.

In the very biginning, this mean-shift algorithm used as a tool to find a group of modes from a set of sampling data (points data) based on the probabability density function, however, this algorithm had been done successfully to bring foward segmentation image processing. Mean shift algorithm is more efficient to work on l.u.v color space, because of some previous studies showed that l.u.v color space is more efficiently used when the search process and the determination of the points nearby have in common (similarity) specific (based on the features of the image). The level of similarity adjacency points are calculated using the Euclidean distance equation. Of some applications. The essence mean shift algorithm shows some important points as follow:

1. The input of this algorithm is a set of data points,
2. Areas of data points which is the densest region of an image identify the modes (local maxima) are formed.
3. Initial procedure of the mean-shift algorithm defines some window search randomly
4. Mean-shift algorithm works iteractievely to achieve convergence.

Each iteration of the algorithm shows window search would shift toward the densest data sets (densest region). Window search will always be headed to the densest region because the mean shift algorithm works based on Parzen windows estimation.

In addition to the mean shift filtering algorithm, there is also method that more widely used lately and this method is proposed as the side by side method of the first one. The second method proposed is directional wavelet transform. The superiority of the method to analyze the image signals on certain components including the degree of slope on component number 45 and component number 135; the horizontal component and vertical component. Directional wavelet transform is used to exploit the directional character in a document and be able to separate the writing on the front side of the document (foregroundstroke) and things that interference in different wavelet frequency domain [6]. On directional wavelet transform, the coefficient of horizontal, vertical and diagonal convoluted with a matrix that represents a particular direction. Matrix equation is as follows:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} \text{ whereas } A = C \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

Where the value of c=$\sqrt{2}$, this means that every point in X and Y convoluted with matrix A then result x1and y1. To recover the original signal (signal reconstruction) then using inverse wavelet transform with the opposite direction of the forward wavelet transform (decomposition). There are 4 stages of directional wavelet transform methods are:

1. Forward wavelet transform without down sampling to separate the image based on different wavelet frequency domain.
2. Convolution process on the horizontal component, diagonal and vertical
3. Thresholding (binarization) on each coefficient of each sub-band
4. Inverse wavelet transform

With the capabilities of the two algorithms; the mean shift filtering and directional wavelet transform, the researcher proposed to create a simulation of the process of image restoration by making use of both these algorithms at once. Through both simulation of this algorithm, the user can select direct or experiment with them to restore the image, or in other words, the user can select which algorithm can generate the best output (the output image quality) according to the level of the damage to the input image. This simulation is implemented with the programming language Matlab based on GUI (Graphical User Interface) making it easier for the user when the application will be used for this simulation.

## 2. RESEARCH METHOD

In accordance to the previous explanation, this research aims to implement a simulation of two algorithms for image restoration of old documents; mean shift filtering algorithms and multi directional wavelet transform. At the beginning of the process, the researcher is collecting data to support both primary data and secondary data. Following the first process, the researcher started to do some literature review to see the development of image restoration algorithms that fit to the data. From the results of the literature review,

the researcher formulated the problem which would be the benchmark of this study beside determine two methode will used.

Design research is an experimental design for this study produces a product in the form of technology to make restoration (improvement) on the image of old documents that were damaged or degraded. This technology is simulation system of the two algorithms, namely the mean shift filtering and multi directional wavelet transform. In carrying out this study, researchers conducted several stages. These stages adapted to systems development methods and each algorithm involved. Generally speaking, there are 9 steps:

1. Finding the problem
2. Literature review of the study that has been done previously to see what methods/algorithms that fit the image restoration problem
3. Formulation of the problem (set the hypothesis as the results of a literature review)
4. Design research methodology
5. Supporting research data collection. Including the image of digitizing old documents that will constitute the input image for the system to be built
6. Analysis of research data. Specify the disturbance or damage to the image of the image based on inputs gathered in the previous stage
7. Development of the simulation system for image restoration involving two algorithms
8. The trial of the system simulation results of step 7
9. The results of the research

## 3. RESULTS AND ANALYSIS

Output of tihs research is a simulation system for document image restoration using both of mean shift filtering and multi directional wavelet transform method. After the simulation system for image restoration involving two algorithms had been developed, next step, researcher tested against the system. Assessment of test results remains subjective through the comments from 10 partisipans about what the most appropriate algorithm used with the input image given. In order to further facilitate the implementation of test systems, researcher divided 39 input image into 4 groups according to the type of damage the image as an explanation on the previous sections. All the participants were asked to analyze the results of the restoration on the entire document in every category of noise. In this case each participant was asked to choose which algorithms provide the best results for each restoration documents (of 12 documents) on each type of noise. The results of testing of the system give the following outcome:

1. In the first group of damaged image which expose much *noise* in the background document images show that directional wavelet transform method is better than the mean shift filtering method
2. In the second group which damaged image is because of the dilated ink or ink splashes showed a significant difference in quality. The test result showed that the two methods are not so good to solve this type of *noise*
3. In the third group, the damaged image due to the digitization errors indicates that the mean shift filtering method is better than the method of directional wavelet transform
4. In the fourth group which damaged image is because of the printed paper from the back side of the document that appears on the front side shows that directional wavelet transform method slightly superior to the method of mean shift filtering.

The following row is a graph showing the results of the comparison of both algorithms for the first and second types of noise:

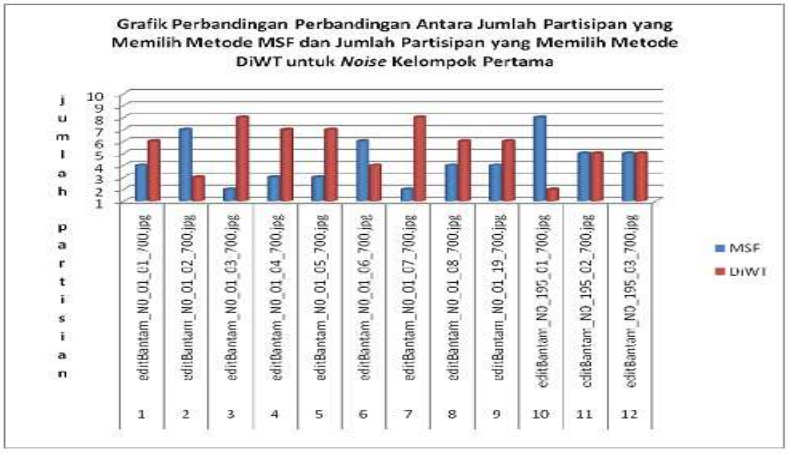


Figure 2. Comparison of two algorithms for type noise I

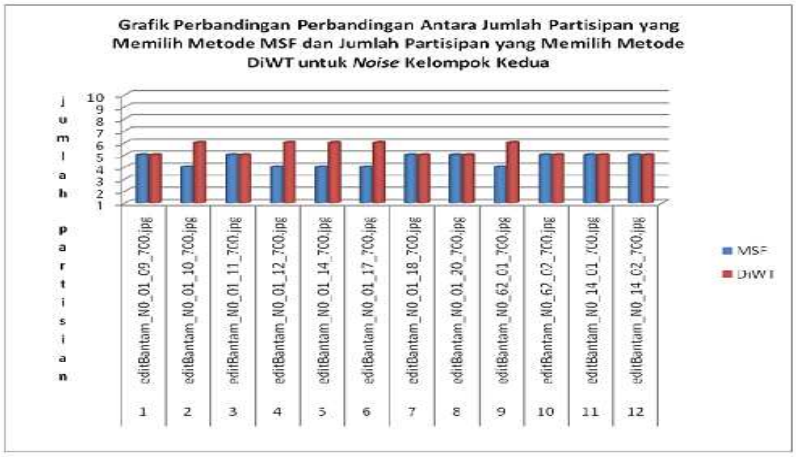


Figure 3. Comparison of the two algorithms for Type noise II

The results from the simulation system can be seen in the following examples: Example image as input, both of the output image (after restoration) by each algorithm**.**

Figure 4. Input Image, Output Image using Mean Shift Filtering and Output Image Using Multi Directional Wavelet Transform

## 4. CONCLUSION

Based on the description of the results of the research on in the form of system simulation old document image restoration application is that users can decide for themselves what is the most appropriate method and provide the best image output between the two existing algorithms based on the type of damage to the image of the old document. Currently research involving two algorithms for 39 input images and concatenated with type writing certain degree slope. Researcher expects future research to develop a system in which the simulation algorithm thus involved more than two types of methode as well as much more disorders that can be restored. In addition to further development researcher expects the system to accommodate a standard format or the characteristics of the input image (*. Bmp, *. Jpg, *. Gif, *. Png, ... etc)

**ACKNOWLEDGEMENTS**

I am willing to present my thanks for everyone who has helped me, especially in this research; my advisor, Prof. Dr. Aniati Murni for her valuable guidance, encouragement, patient, cooperation, advice, and suggestion
which are very helpful in finishing this research.

**REFERENCES**

[1] R.C. Gonzalez dan R.E. Woods, " Digital Image Processing, " Addison-Wesley, USA, 1992
[2] D. Fadoua, F. L. Bourgeois, and H. Emptoz, " A modified Mean Shift Algorithm For Efficient Document Image Restoration, " SITIS, pp. 686-695, 2006
[3] H. Negishi, J. Kato, H.Hase and T. Watanabe, " Character extraction from noisy background for an automatic reference system," 5th International Conference on Document Analysis&Recognition, India, pp.143-146, 1999
[4] S.Liang, M. Ahmadi, A morphological approach to text string extraction from regular periodic overlapping text/background images, Graphical Models and Image Processing. CVGIP, Vol. 56, No. 5, pp.402-413, Sep.1994
[5] D. Comaniciu and P. Meer, " Mean shift: A robust approach toward feature space analysis , " IEEE Transactions on PAMI, Vol. 24, no. 5, pp. 603-619, 2002
[6] C.L. Tan, R. Cao dan P. Shen, "Restoration of achival documents using a wavelet technique," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.24, No.10, pp.1399-1404, 2002

**BIBLIOGRAPHY OF AUTHORS**

Samosir, R.S. (2012). Trigonometry Learning System Based on Multimedia. Proceedings of the National Seminar on Information and Communication Technology Applied. Universitas Dian Nuswantoro. Semarang. Indonesia

Samosir, R.S. (2012). Documnet Image Restoration Using Multi Directional Wavelet Transform. Proceedings of the National Seminar on Informatics. STIMIK Potensi Utama. Medan. Indonesia

Research Report of Institute of Technology and Business Kalbe. (2012). Designing a Data Warehouse System Admissions at Institute of Technology and Business Kalbe (Report). East Jakarta. Indonesia

Samosir, R.S. (2013). Designing a Data Warehouse System Admissions at Institute of Technology and Business Kalbe.Teknologi Informatika Journal. Institute of Technology and Business Kalbe. East Jakarta. Indonesia