Enhancement of Old Document Image after Restoration with Morphologycal Approach

Ridha Sefina Samosir

***Abstract*—An old document contains important information about culture, heritage and past story. Quality decreasing of old document images is caused by method and time storing failure, ink or paper quality, and digitizing process failure. Quality decreasing means the appearance some noise such as printed writing, widened ink, fading paper colour, some missing text and some noise caused of bad digitizing process. This study aims to do additional operation after restoration using morphologycal approach. Morphologycal approach is used to improve image quality after restoration process. Restoration use filtering algorithm. Evaluation technique for this study use statistic approach through calculation average percentage. Morphologycal give best result for old document images with noise such as widened ink and dirty background. But not good enough results for old document images with printed graffity from back side which appear in front side because input image contains italic character.**

***Index Terms*—Morphologycal, old document image, noise restoration.**

I. Introduction

An old document image contains many important information even sometimes it has relationship among the documents. The old documents can presents historical story in the past in an area. Even an old document can be guide for archeologist to find missing historical place and heritage. An old documents image has role as historical noted, identity source both as personal and as community. Viscount St. Davids said that a country can know the condition of their nation in the past based on the story in the old document. Usually, old document contains many particularity such as a large variability noise and degradation: page skew, random alignment, specific fonts, the presence of embellishments, spacing variation, words, line, paragraph, margins, and object boundaries [1].

Identify the character which is contained in old documents image will determine the apropriate methods or techinque for the image. This identification result will avoid any assumption about the structure or content of the documents through texture feature calculation [2]. Documents that are stored too long with poor storage methods cause documents to be damaged. The quality of paper and ink is also another factor that causes the damage. The damage that appears in the documents called as noise. Noise is one of phenomena on image processing.

Noise often arises as result of processing failure, transmision failure, and when taking images with sensor devices [3]. One kind of noise that often arises in the documents is ink bleed trough removal. Ink bleed trough removal is a condition that some various sign and graffity appear in the document which interfere the document quality. Various sign such as printed writing from document back side, widened ink, sign because of lack paper ability to absord the ink, sign appear while digitalization process and other factor. Usually, a case with printed writing from back side document appears in document with italic hand writting. If the slope degree of the text is 45° then printed text has slope of appoximately 135°.

One method that can be used to overcome noise in the old document is the restoration method. Restoration can be done on digital images. It means that old document images used as input must be go through digitalization process first. In additon, digitization can prevent further damage because frequency of physical contact with the document is fewer. There are many algorithm for restoration in image processing, one of them is filtering. Filtering algorithm works at pixel level. Pixel is the smallest unit of a digital image. Filtering means selecting pixel values from digital images that can describes the appearance original image clearly. Filtering method used for this study is mean shift filtering. Mean shift filtering aims to find modes from a set of data based on probabilily density finctions. Mean shift filtering effectively works in L.U.V colour spaces when searchs nearby points that have certain similarities. Other research previous use RGB colour space for speed sign recognition. Color thresholding in RGB is being used to segment road sign images. Color-based methods use sign's color information to remove non road sign objects from the scene [4].

The restoration algorithm choosed is powerful enough to improve image quality so the contents in the document can be recognized again. The impact of image restoration is the paper document more clear and some unexpected noise or annoying writing is lost. But sometimes in some case, new problem arises such as the loss of some scratches even though that scratches are still part of documents contents. The other words, that scratches are still needed to complete the document contents. In addition, after restoration some strokes that are not too large but can affect the process of contents recognition is appear in the document. We need tecnique or method to reassert the missing scratches. The appearance of new scratches that are no needed due to the restoration process needs to be removed or minimized.

Some previous study relevant to this topic such as research by Teo Asplund and Cris L. Luengo Hendriks in 2016. Both of them use opening to solve line detection problems. Exactly, they proposes opening technique and called as path opening. They concern not only about line measurement but also the

Manuscript received May 28, 2019; revised August 20, 2019. This work was supported in part by the Kalbis Institute.

Ridha Sefina Samosir is with the Kalbis Institute and Information System Department, Pulomas, East of Jakarta, 13210 Indonesia (email: ridha.samosir@kalbis.ac.id).

DOI: 10.7763/IJCTE.2020.V12.1256

structure point during selection step. They proposed upper skeleton path opening algorithm to avoid the problem occlusion while also inhibiting zig-zagging. Upper skeleton path opening approximates the path opening has been presented. The method also fast and able to be reconstructed as an approximation of the traditional path opening [5]. Sameena Pathan, Siddalingaswamy, and Gopalakrishna use opening method to connect unconnected small circular object with four neighborhood for retinal vessel images. Consequences arises after algorithm implemented for retinal vessel image is filtering out some unexpected pixels. In addition, some labeled pixels will be connected automatically [6]. Suman Rani also use opening method for pre processing medical image before edge detection. In this study, opening used as de-noised technique for images with salt and pepper noise in the document backgorund. [7]. Sebastián Salazar *et al.* use opening methods with filtering, exactly with Gaussian Filtering. This study compare between original dark Channel Prior algorithm and DCP with morphologycal approach. Experiment result shows that algorithm effectively reduce the processing time. Input image used for this study is fog images. Denoising process for fog imageges needs to be done because fog images is obtained from external environment so it has opportunity contains low contrast and modified colurs. Sebastián Salazar *et al.* use MSE (Mean Square Error) and SSIM (Structural Similarity Index) as performance measurement indicator for DCP algorithm with Morphologycal approach [8]. Sourabh, Satish and Dhara with research title: Color sensing and image processing-based automatic soybean plant foliar disease severity detection and estimation, this study use opening technique for segmenting soybean plant foliar disease images. Opening method in segmentation process can identify infected area by foliar disease. After segmentation few background pixels contribute falsely to the count of segmented leaf area pixels. This pixels are treated as noise and removed with opening technique [9]. Sarabpreet and Sahambi use opening techique to detect cell existence. Usually, problems exist in cell images is low contrast with poor edge information so that opening method can be used to extract cell regions from the low contrast cell images. Opening technique using erotion and dilation. Erotion removes all small objects and dilation restore the shape object. Sarabpreet dan Sahambi use accuration, precision and sensitivity as measurement performance to the algorithm [10].

Form some reseach has been done before show that morphologycal approach is done in pre-processing step before restoration process, segmentation process, and edge detecton process. Previous research has shown that all algorithm provide better result according to the requirement. But, its can not resolve new problems exist after restoration such as the loss of some scratches and the appearance of several strokes that are not too large but can affect the process of contents recognition in the document. Due to that proplems, this study proposes an experiment using morphologycal approach which used not in pre-processing but after main process such as restoration, segmentation and edge detection. This additonal operation aims to press as little as possible the appearance of new annoying strokes and reinforces scratches that are almost lost but still needed. Additonal operation used in this study is morphologycal approach exactly opening process which consist of dilation and erotion methods. Through restoration and additional operation, the image quality of old document not only repaired but also improved so that it is easier to interpreted.

II. Proposed Algorithm

A. Mean Shift Filtering

In this study, Mean shift filtering is used to restoration process. Mean shift fltering work with filtering approach. The way of filtering methods is highlight technique for the appearance of the image so it is easier to distinguish from other feature. Filtering can also disguise the appearance of images that dont want to displayed. Filtering means taking part of a signal from a certain frequency and discarding a signal at another frequency. Images frequency is influenced by colour images gradient existing. Images with high level gradients tend to be low frequencies, and vice versa [11]. The principle of mean shift filtering is iterating for n times until a convergent data points state is obtained. Iteration is done to update the position of center from each window based on mean shift vector value. Mean shift vector value is obtained from the calculation of several variables including spatial kernel bandwidth (hs) and color kernel bandwith (hr). Both of these values are needed while formation modes from nearest neighborhood pixels set. The following are the steps of mean shift filtering algorithm:

1) Image initialization: In this section, all pixels from the images is initialized as spatial and color information pairs.
2) Mean Shift: Mean shift is occur to distributes data points from input images into a set of search window then determine the center point and mean shift vector for each search window. After that, search window shifts towards the mean shift vector. Then the mean shift vector is calculated again. This stages continuously iterate until the convergence of data points is reached.
3) Convergence: After convergence reached, then filtered pixel pairs from nearest neigborhood is stored as an output pixels.

B. Morphologycal Approach

Mathematically, morphologycal is a concept which concern in processing and analysis of images or signals using filters and other operators that modify them. Fundamental theory in morphologycal are integral geometry and lattice algebra. Morphologycal filter is obtained by means of erotion and dilation [12]. In this study, morphologycal approach as additonal operation after restoration process. The use of morphologycal approach as additonal operation is to improve the quality of restored images. Sometimes restoration process produces new objects that are less than 2 pixels and eliminate objects that are less than 3 pixels even though the objects is needed. Morphologycal approach used is opening algorithm precisely erotion and dilation. Opening opertion produces images with less noise and the documents background more in accordance with original image. Opening serves to remove small and thin objects, smoothing the boundary points of large objects, and reinforcing slightly faded objects.

Erotion operation works by removing the boundary points of an objects smaller than 1 pixels of the overall objects. The effect is that erotion process can eliminates interference oject less than 2 pixels. Refer to (1) below show that binary image (E) is obtained from erotion process of B by S.

$$E = B \otimes S = \{x, y | S\ xy \subseteq B\ \} \tag{1}$$

Dilation process is process of combining points so that they become an objects. If a separate objects less than 3 pixels the objects will blend into an obejct.

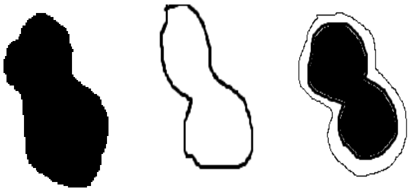

Fig. 1. (a) Binary image, (b) Dilation result, and (c) Erotion result.

Usually, dilation used to fill the hole objects into object after it has been segemented. Refer to (2) below show that binary image (D) is obtained from dilation process of B by S. Fig. 1 show illustration of opening process precisely erotion and dilation. From that figure, clearly seen the differences between erotion and dilation.

$$D = B \oplus = \{x, y | S\ xy\ \cap B\ \neq 0\} \tag{2}$$

III. Research Method

This section contains all research methods used in this study include theoretical background for all methods.

A. Research Cronological

This study is a continuation of previous research about old document image restoration using mean shift filtering in 2013. Experiment result from previous research show that there are some type of input images need to be enhancement although it has been restored. Some case presents new problems after image is restored. This new problems appear for old document images which contains italic character. Based on that new problems, this study then continued with additional operational precisely with morphologycal approach.

B. Research Design

This study is belongs to the type of applied research because the main objective is to solve the real problems. The design research used in this study is a case research study. A case research study means a research that focus on certain problems or collect deeper data related to the object under study. A study case research will do a empirical investigation about a phenomenon to be solved by the researcher. Case study research is descriptive and eksploratory. Usually, data used in case study research is primary data [13].

C. Data Acquisition

Data used for this study is taken from National Archives Institution in the form of old documents image. All data collected has been digitized by the National Archives Institution so researcher not direct physical contact with the old documents. There are some type of interferring graffity in the document. Almost all documents contain italics type. From data collected show almost all documents contains hand writing with italic type. There are 48 old document image used for experiment. Restoration process will be easier if all documents image is classified with a certain label. Classification for all data set determine classifier baseline for each input images. This Baseline classifier will determine some appropriate parameter value from the algorithm [14]. First step for image input is image conversion. Old document image is converted in dyadic form. Dyadic form formulated with 2^n. All image size is made to be 215x215.

D. Research Procedure

This flowchart Fig. 2 describes all procedure done for this study, its start with initialization until evaluation performance measurement:

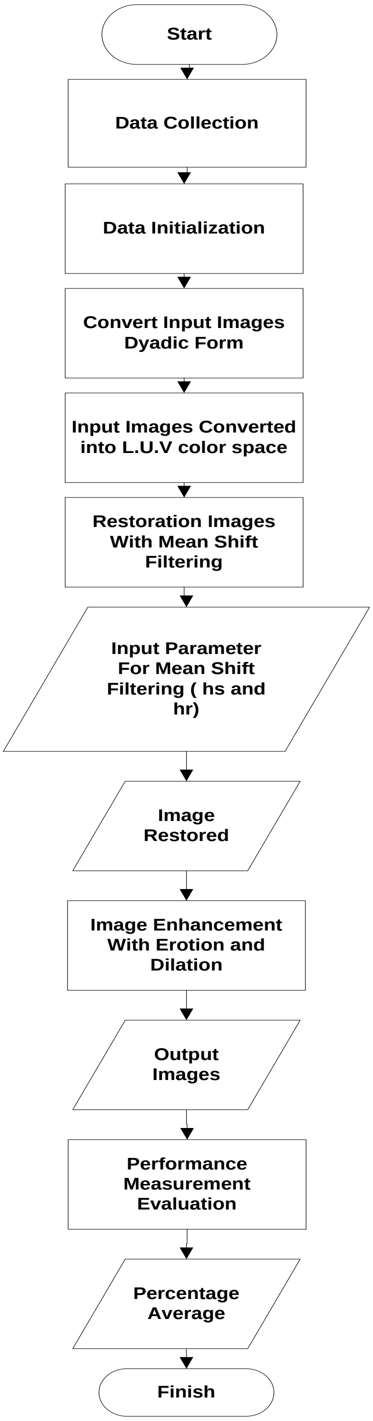
Start
Data Collection
Data Initialization
Convert Input Images Dyadic Form
Input Images Converted into L.U.V color space
Restoration Images With Mean Shift Filtering
Input Parameter For Mean Shift Filtering (hs and hr)
Image Restored
Image Enhancement With Erotion and Dilation
Output Images
Performance Measurement Evaluation
Percentage Average
Finish

Fig. 2. Flowchart for research procedure.

E. Performance Research Evaluation

Performance measurement evaluation carried out for images after opening process as additional operation. This study proposes subjective approach to evaluate the quality of output image after enhancement process. All of old douments image is divided into 4 group according to noise type. Evaluation process involves respondent to evaluate the quality of output image. Each image group is evaluated by 10 respondent. Respondents provide an assessment of whether the image quality is good or not. The evaluation results data were collected and then processed with statistically approach by calculating the percentage of output images that were succesfully repaired and improved in quality. Evaluation is done for restored image after additional porcess with erotion and dilation.

IV. Results and Discussion

In this section will presents and explain some results from the experiment such as image initialization, grouping input image results, restored image results with mean shift filtering, and final output images after morphologycal approach implemented precisely with opening methods.

A. Data Preparation Result

This section explain result about detail data input used and data pre processing. First step is group input data based on noise categories. This classification will help to restorate the old documents image easier. In this study, data clustered into 4 cluster based on noise type:

1. Old document where noise occur in document background, maybe because of paper age.
2. Old document with widened ink or ink splash
3. Old document with interference graffity printed from back side of document. Graffity printed from back side occur some because of character type in the document. Almost all documents contain italic type for the character. It the character slope is 45° then printed text form back side has slope approximately 135°.
4. Old document with noise because of digitation process error.

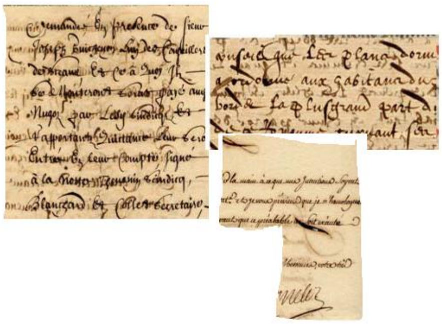

Fig. 3. Old document image.

According to Chansong Liu *et al*. that sometimes camera captured distorted document image. The case that often appears while document digitization process is that the character in the document becomes curved. Changsong Liu *et al*. proposed thin plates splines algorithm to estimateds warping shapes of each text line and rectifies it [2]. There are 48 images which used in this study are devided into 4 groups based on noise type. Each group involves 12 images. After data classification then data input used is convert in dyadic form and uniform size 512x512 pixels. It's mean all image will be convert into dyadic form and uniformly size. Dyadic form means that each pixel in the matrix is multiple of two. Matrix is representation for all pixel from the image. After uniformity size and format is achieved then image convert from RGB (Reg Green Blue) into L.U.V colours space. Fig. 3 is one example of data input with some kind of noise.

B. Restoration Result

In the explanation above, restoration process is done for data input representing 4 types of noise. The following is the result of image restoration using mean shift filtering or the first group inut images, namely old document image where noise is occured in document background, maybe because of paper age.

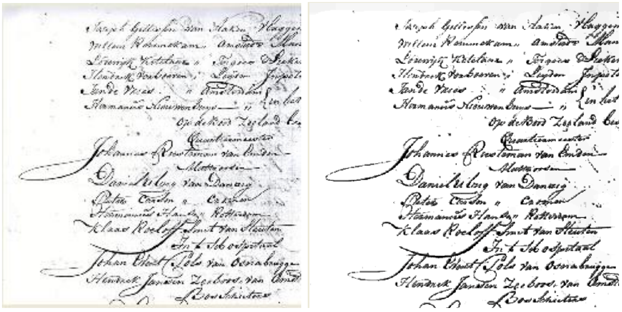

Fig. 4. (a) Input old document image and (b) Image after restoration.

From Fig. 4, it can been seen that there have been significant changes to the old document images. Cleaner document background from original images. Another visible effect is that writing is easier to recognize because disturbing objects have diminshed. But it can also be seen that after the restoration process, a number of small objects withs a size of less 3 pixels appear 3 pixels appear or vice versa, the loss of several small new objects but the obejcts is part of the character in the document.

C. Morphology Result

Previous explanation show that after the restoration process several small object is appear and some needed small object is loss. After restoration process then morpohological approach applied to all output image. The purposes of morphologycal approach is to eliminate some unnecesarry graffity but reaffirming writing lost due to the restoration process. Fig. 5 show output image after morphology approach exactly after erotion and dilation.

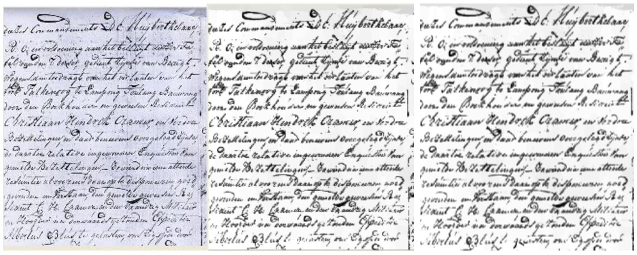

Fig. 5. (a) Input image (b) Image after restoration (c) After erotion dilation.

First image is original digital image from National Archive Agency.That image show some noise in the background, some widened ink and some printed graffity from back side occures in front side. Then second image is restoration result image. Second image show the image background more clear than the first one. Other than that, the appearance of a widened ink decreases. Last image is output image. This output image as erotion and dilation process result. Form third image show that some unnecesarry graffity decreases and some lost graffity needed appear.

From result above show that opening process successfully eliminating interfering graffity. Eliminate interfering object that less than 3 pixels and merge separetted object that less than 3 pixels into an object. Based on explanation above described that accuration result found trough statistical approach. Here are performance result for all images which classified in fourth type of noise after last process with morphologycal approach.

D. Performance Measurement Result

TABLE I: SUMMARY PERFORMANCE MEASUREMENT RESULT

Noise Type	Details	% Agree	% Disagree
Type 1	Artifact in Document Background	58%	42%
Type 2	Widened Ink or Ink Splash	50%	50%
Type 3	Interference graffity printed from back side of document.	25%	75%
Type 4	Noise because of digitalization process error	60%	40%

Based on explanation above described that performance measurement result found trough statistical approach. This study used subjective evaluation by expert to evaluate the quality of image after algorithm implemented. There are 10 expert that give opinion about the quality. Every expert evaluate 12 image for each noise category. From that 10 expert, they give argument whether the result good or not good. After all image evaluated for each category then find the average percentage. Table I shows summary for performance measurement of each noise type. Then some chart below presents details performance result and average percentage for fourth type of noise after last process with morphologycal approach.

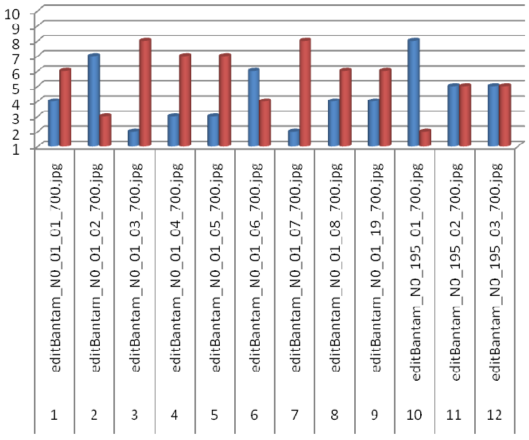

Fig. 6. Statistical report for image with first type noise.

1. This section explain about performance result for noise type 1. Noise type 1 shows some artifacts that occur in documents background, maybe because of paper age. Fig. 6 show average percentage for noise in image background. Fig. 6 show that there is ten participant for twelve input image. Input image is output image after restoration process. Red bar indicate number participants who declare good quality and the blue bar indicate the opposite. Summary for first type noise show that for twelve image, 58% agree that erotion and dilation give better result.
2. This section explain about performance result for noise type 2. Noise type 2 shows widened ink or ink splash on the paper. Fig. 7 show average percentage for noise type 2.

Fig. 7 show there is 10 expert who evaluate the output result after opening process. Red bar indicate good quality and blue bar indicate the opposite. Evaluate result for first image is 50% respondent agree opening process give good quality and 50% disagree, second image show that 60% respondent agree that opening process give good quality and 40% disagree. Evaluation is done for all image to obtained the average result between red and blue bar. Average result for all image is 50% agree that erotion and dilation give better result for old documents image with widened ink or ink splash.

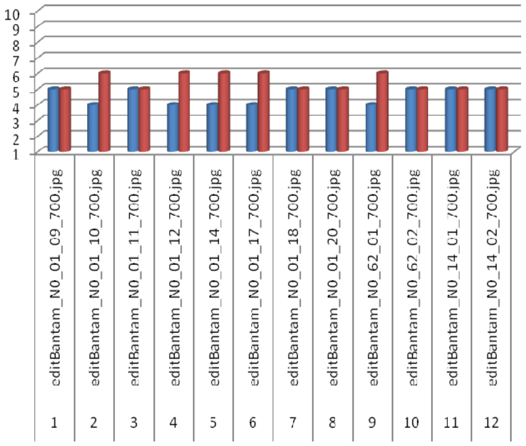

Fig. 7. Statistical report for image with second type noise.

3. This section explain about performance result for noise type 3. Noise type 3 show interference graffity printed from back side of document. Fig. 8 show average percentage for noise type 3.

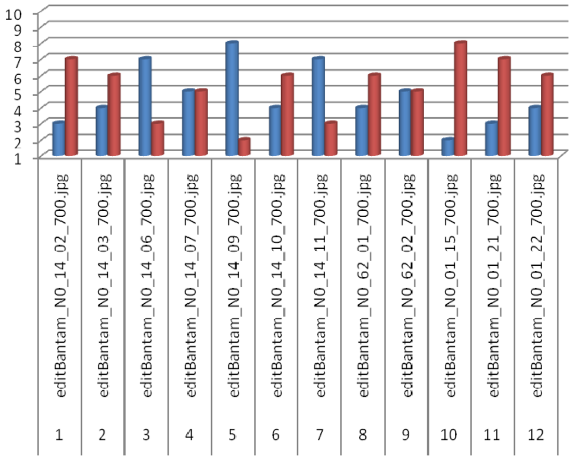

Fig. 8. Statistical report for image with third type noise.

Fig. 8 show 10 respondent who evaluate the output result after opening process. Red bar indicate good quality and blue bar indicate the opposite. Evaluate result for first image is 30% respondent agree opening process give good quality and 70% disagree, second image show that 40% respondent agree that opening process give good quality and 60% disagree. Evaluation is done for all image to obtained the average result between red and blue bar. Average result for all image is 75% disagree that erotion and dilation give better result for image with printed graffity from back side which appear in front side.

4. This section explain about performance result for noise type 4. Noise type 4 shows some artifact which occur because of digitation process error. Fig. 9 show average percentage for noise type 4.

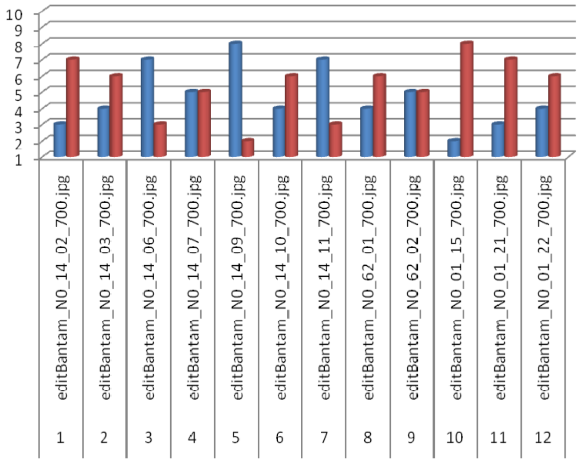

Fig. 9. Statistical report for image with fourth type noise.

Fig. 9 show 10 respondent who evaluate the output result after opening process. Red bar indicate good quality and blue bar indicate the opposite. Evaluate result for first image is 70% respondent agree opening process give good quality and 30% disagree, second image show that 60% respondent agree that opening process give good quality and 40% disagree. Evaluation is done for all image to obtained the average result between red and blue bar. Average result for all image is show 58% agree that erotion and dilation give better result for image with noise caused by failure of digitalization process.

V. Conclussion

From all study has done, the conclusion that can be taken are:

- Morphologycal approach did not show good enough result for image with third type noise. Image with printed graffity from back side which appear in front side. Its because almost image contains italic type character. Italic type character will produce printed text in the opposite direction on the back side of the documents. Dilation and erotion will be effective while we used directional approach for restoration algorithm. Directional approach for restoration will help to minimize noise in the back side of document then erotion and dilation task is easier.
- Morphologycal approach exactly dilation and erotion give better result for image after restoration because of widened ink and background noise. Its because the ability of morphologycal approach to eliminate unnecessary graffity and reinforce the disconnected character needed.

Conflict of Interest

All of input, process, and output for this study does not contains conflict of interest. As author, I declare no conflict of interest in this study.

Author Contributions

All step in this study is done by RSS. But this study based on final task that given by lecturer of image processing.

Acknowledgment

We would like to thank to National Archive Institution for the data given. Thanks to all friends I have for the support given for me to finish this study.

References

[1] M. Mehri, P. Gomez-Kramer *et al.*, “A texture-based pixel labeling approach for historical books,” *Journal Pattern Analysis & Applications*, vol. 20, no. 2, pp. 325-364, May 2017.

[2] C. Liu, Y. Zhang, B. Wang *et al*. “Restoring camera-captured distorted document images,” *IJDAR*, vol. 18, no. 2, pp. 111-124, June 2015

[3] M. H. Lee and S. Y. Shin, “Modified pixels based fast median filter in impulse noise environments,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 3, pp. 731-741, March 2018.

[4] A. M. Mansour, J. Abukhait, and I. Zyout, “Speed sign recognition using independent component analysis,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, pp. 1-7, November 2013.

[5] T. Asplund, C. L. Luengo, and A. Faster, “Unbiased path opening by upper skeletonization and weighted adjacency graphs,” *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5589-5600, September 2016.

[6] S. Pathan, P. C. Siddalingaswamy, and K. G. Prabhu, “A pixel processing approach for retinal vessel extraction using modified Gabor functions,” *Progress in Artificial Intelligence*, vol. 7, no. 1, March 2018.

[7] S. Rani, “A novel mathematical morphology based edge detection method for medical image,” *CSI Transactions on ICT*, vol. 4, no. 2, pp. 217-225, December 2016.

[8] S. Salazar-Colores, J. M. Ramos-Arregu ń, C. J. Ortiz-Echeverri *et al.*, “Image dehazing using morphologycal opening, dilation and Gaussian filtering,” *Signal, Image and Video Processing*, vol. 12, no. 7, pp. 1329-1335, October 2018.

[9] S. Shrivastava, S. K. Singh, and D. S. Hooda, “Color sensing and image processing-based automatic soybean plant foliar disease severity detection and estimation,” *Multimedia Tools and Applications*, vol. 74, no. 24, pp. 11467-11484, December 2015.

[10] S. Kaur and J. S. Sahambi, “Cell detection in very low contrast images using discrete curvelet transform and radon transform with morphologycal operations,” presented at 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences, Chandigarh, India, December 21-22, 2015.

[11] R. S. Samosir, “Filtering and wavelet transform algorithm for old document image restoration,” *COMTECH*, vol. 8, no. 3, pp. 177-181, September 2017.

[12] C. Alcalde, A. Burusco, and R. Fuentes-Gonz ález, “Application of the *L*-fuzzy concept analysis in the morphologycal image and signal processing,” *Annals of Mathematics and Artificial Intelligence*, vol. 72, no. 1, pp. 115-128, October 2014.

[13] L. Cohen, L. Manion, and K. Morrison, *Research Methods in Education: An Introduction*, 6th ed. Routledge, New York: Taylor and Francis Group, 2007, ch. 1, pp. 41-46.

[14] Y. Shao, C. Wang, and B. Xiao, “A character image restoration method for unconstrained handwritten Chinese character recognition,” *IJDAR*, vol. 18, no. 1, pp. 73-86, March 2015.

Ridha Sefina Samosir was born in September 9, 1982 in small village named Aek Nabara. This village is located in North Sumatera Island in Indonesia. She graduated from Computer Science Department of Sanata Dharma University in Jogjakarta of Indonesia. She completed her bachelor program in 2004. In 2011, she completed her postgraduated program for major of computer science from Indonesia University.

She began her carrier as a teacher in National Plus School Named Djuwita School in Batam. She teach for playgroup until senior high school for IT and mathematics subject. She enjoy her career as teacher from 2004 until 2005. Next, she works as civil servant in finance department. Her responsibility is calculate the salary of civil servant of health department in Central Tapanuli. Her activity there for 1 years too in 2005 until 2006. Next career as a teacher in child education institution named KUMON. She works for 1 years from 2006 until 2007. She teach there for early chilhood and focused for mathematics subject. In 2008, she joined potgraduated program for computer science in Indonesia University. She finished the potsgraduated program in 2011. Her publication from her thesis is comparation between mean shift filtering and multi directional wavelet transform for old document image restoration. This article is safed as research report in the library of Indonesia University. From this thesis, she write three publication, one of them is published in Journal COMTECH in 2017 which Bina Nusantara University as the publisher. The title of this publication is Filtering And Wavelet Transform Algorithm For Old Document Image Restoration. After finished her study then she joined in one higher education in Jakarta Known as Kalbis Institute until now. Mrs. Ridha Sefina Samosir is an active member for Association of Information System (AIS). Mrs. Ridha Sefina Samosir always join for some activity from this organization to increase her knowledge, relation, and skill. Mrs. Ridha Sefina Samosir also acts as editorial board in internal journal which managed by Kalbis Institute.