History Document Image Background Noise and Removal Methods

Ganchimeg Ganbold*

ARTICLE INFO

Article history:
Received 14 February 2015
Revised 11 April 2015
Accepted 19 April 2015

Keywords:
Binarization,
History Document Noise,
Noise Removal Algorithms

ABSTRACT

It is common for archive libraries to provide public access to historical and ancient document image collections. It is common for such document images to require specialized processing in order to remove background noise and become more legible. Document images may be contaminated with noise during transmission, scanning or conversion to digital form. We can categorize noises by identifying their features and can search for similar patterns in a document image to choose appropriate methods for their removal. In this paper, we propose a hybrid binarization approach for improving the quality of old documents using a combination of global and local thresholding. This article also reviews noises that might appear in scanned document images and discusses some noise removal methods.

1. Introduction

Document image binarization refers to the conversion of a gray scale image into a binary image. It is the initial step of most document image analysis and understanding systems. Usually, it distinguishes text areas from background areas, so it is used as a text locating technique (Gatos, Pratikakis, & Perantonis, 2006). Nowadays, with the increase in computer use in everybody's lives, the ability for people to convert documents to digital and readable formats has become a necessity. Scanning documents is a way of changing printed documents into digital format. A common problem encountered when scanning documents is 'noise' which can occur in an image because of paper quality, the typing machine used, or it can be created by scanners during the scanning process (Farahmand, Sarrafzadeh, & Shanbehzadeh, 2013). Noise removal is one of the steps in pre-processing. Among other things, noise reduces the accuracy of subsequent tasks of OCR (Optical character Recognition) systems. It can appear in the foreground or background of an image and can be generated before or after scanning. Historical manuscripts and scanned document images often have degradations like uneven contrast, show through effects, interfering strokes, background spots, humidity absorbed by paper in different areas, and uneven backgrounds (Gatos, Pratikakis, & Perantonis, 2004). These problems cause challenges similar to those in an OCR system. Such degradations can destroy the blank spaces

* Senior Lecturer, School of E-Open Institute of The Mongolian University of Science and Technology, Mongolia (ganaa_mzb@yahoo.com)
International Journal of Knowledge Content Development & Technology, 5(2): 11-24, 2015.
http://dx.doi.org/10.5865/IJKCT.2015.5.2.011

between lines and words. There are many methods in the literature to enhance background degradations in document images.

2. Binarization and Thresholding Based Methods

One of the methods to enhance background quality of gray scale images employs thresholding and binarization techniques (Farahmand et al., 2013). Some resources divide thresholding techniques into two major groups. The methods in the first group use global algorithms that employ global image features to determine appropriate thresholds to divide image pixels into object or background classes (Farahmand et al., 2013). The second group uses local image information to calculate thresholds, similar to the locally adaptive thresholding method that uses neighborhood features such as the mean and standard deviation of pixels (Sauvola & Pietikainen, 2000). However, the methods of the second group are much slower than the first, but their accuracy is higher. Niblack (1986), and Sauvola and Pietakainen (Sauvola & Pietikainen, 2000) use the local variance technique. Sauvola and Pietakainen's method is an improvement on Niblack's method, especially for stained and badly illuminated documents.

2.1. Niblack's Technique

Niblack's (1986) algorithm calculates a pixel-wise threshold in a rectangular window over the gray level image. The computation of threshold is based on the local mean *m* and the standard deviation *s* of all the pixels gray levels in the window and is given in formula 1 below:

$$N_{Niblack} = m + k * s \qquad (1)$$

$$N_{Niblack} = m + k\sqrt{\frac{1}{NP}\sum_{i=1}^{NP}(p_i - m)^2} = m + k\sqrt{\frac{\sum_{i=1}^{NP} p_i^2}{NP} - m^2} = m + k\sqrt{B} \qquad (1^*)$$

NP, the number of pixels in the window *k* is a constant fixed to 0.2 by the authors. The advantage of Niblack's method is that it always identifies the text regions correctly as foreground, but on the other hand, it tends to produce a large amount of binarization noise in non-text regions and text boundaries.

2.2. Sauvola's Technique

Sauvola's algorithm (Sauvola & Pietikainen, 2000) improves Niblack's method by computing the threshold using the following formula:

$$T_{Sauvola} = m * (1 - k * (1 - \frac{S}{R})) \qquad (2)$$

where *k* is set to 0.5 and *R* to 128. This method outperforms Niblack's algorithm (1986) in images where the text pixels have near 0 gray-values and the background pixels have near 255 gray-values; however, in images where the gray values of text and non-text pixels are close to each other, the results degrade significantly.

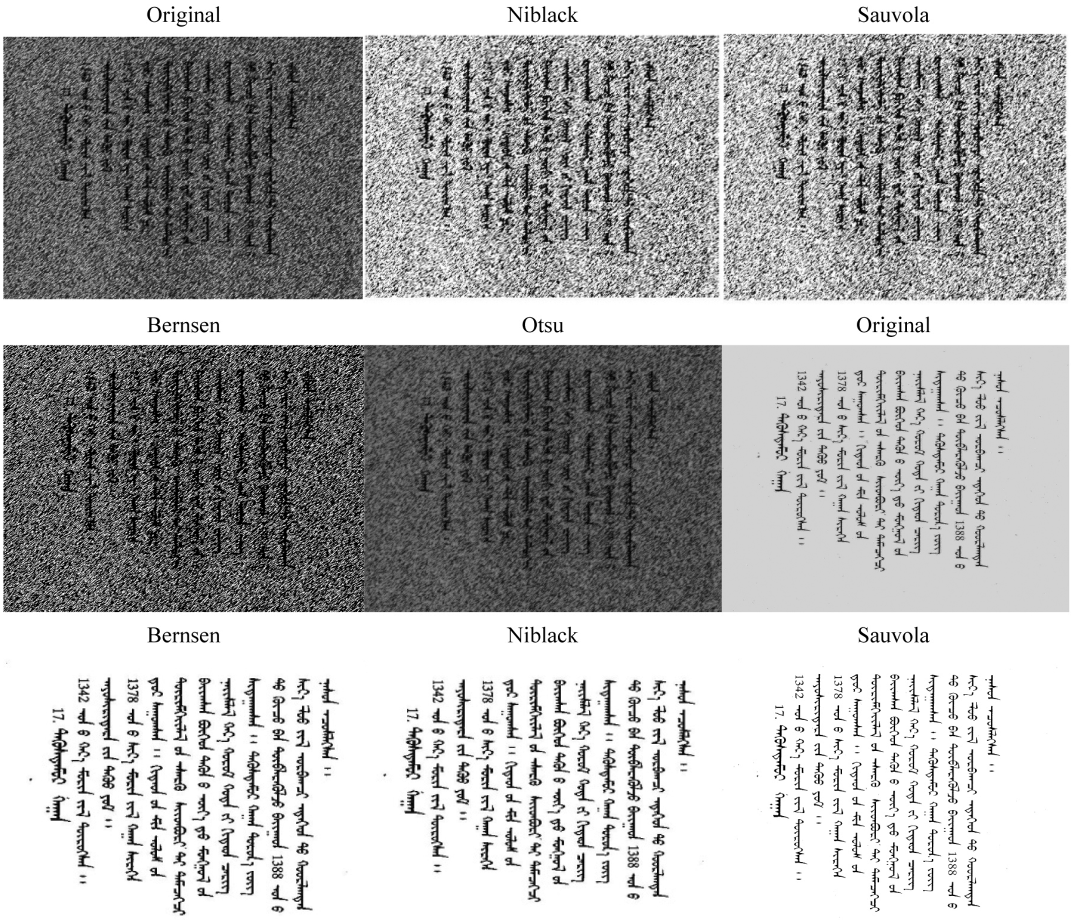


Fig. 1. Comparison of local variance technique

2.3. Bernsen's Technique

Bernsen (1986) uses the local gray range technique. In this technique the range between the maximum and minimum pixel gray range within the local window is used to determine the threshold value. In this method the local threshold value $T(x, y)$ at (x, y) is calculated within a window of size *wxw* as:

$$T(x,y) = 0.5(I_{\max(i,j)} + I_{\min(i,j)}) \qquad (3)$$

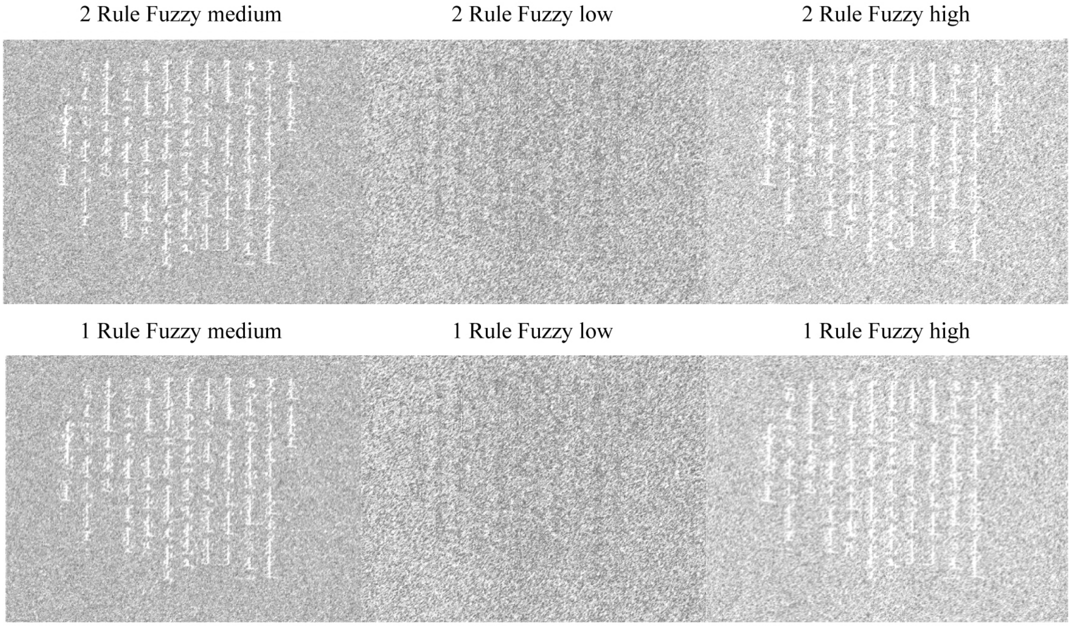
where $I_{max(i,j)}$ and $I_{min(i,j)}$ are maximum and minimum gray values within the local window provided contrast

$$C(i,j) = I_{\max(i,j)} - I_{\min(i,j)} \geq 15 \qquad (4)$$

In this method, the threshold is set at the midrange value, which is the mean of the maximum and minimum gray values in a local window of size *wxw* A value of *w*=31 gives satisfactory results; however, if the contrast *C(i,j)* is below a certain threshold (15), then that neighborhood is said to consist of only one class, foreground or background, depending on the value of *T(x, y)*. There is no bias to control the threshold value.

3. Fuzzy Based Methods

Enhancing image quality using fuzzy logic operators is based on mapping gray levels of an image to fuzzy space, and we know that defining an appropriate membership function requires experience and prior knowledge. Enhancement with fuzzy operators employs weighing features proportional to some image features, like average intensity to increased contrast (Zadeh, 1965). In 1997, H.R. Tizhoosh proposed a fuzzy approach to image enhancement using a contrast intensification operator. This operator increases the difference between gray levels by increasing membership functions higher than 0.5 and decreasing those lower than 0.5 values so the contrast in image is improved (Kuppannan et al., 2006).



2 Rule Fuzzy medium

2 Rule Fuzzy low

2 Rule Fuzzy high

1 Rule Fuzzy medium

1 Rule Fuzzy low

1 Rule Fuzzy high

Fig. 2. Result of Fuzzy logic based methods

4. Histogram Based Methods

Histogram equalization provides a sophisticated method for modifying the dynamic range and contrast of an image by altering that image such that its intensity histogram has a desired shape (Fisher et al., 2003). Unlike contrast stretching, histogram modeling operators may employ non-linear and non-monotonic transfer functions to map between pixel intensity values in the input and output images. Histogram equalization employs a monotonic, non-linear mapping which re-assigns the intensity values of pixels in the input image such that the output image contains a uniform distribution of intensities (Fisher et al., 2003).

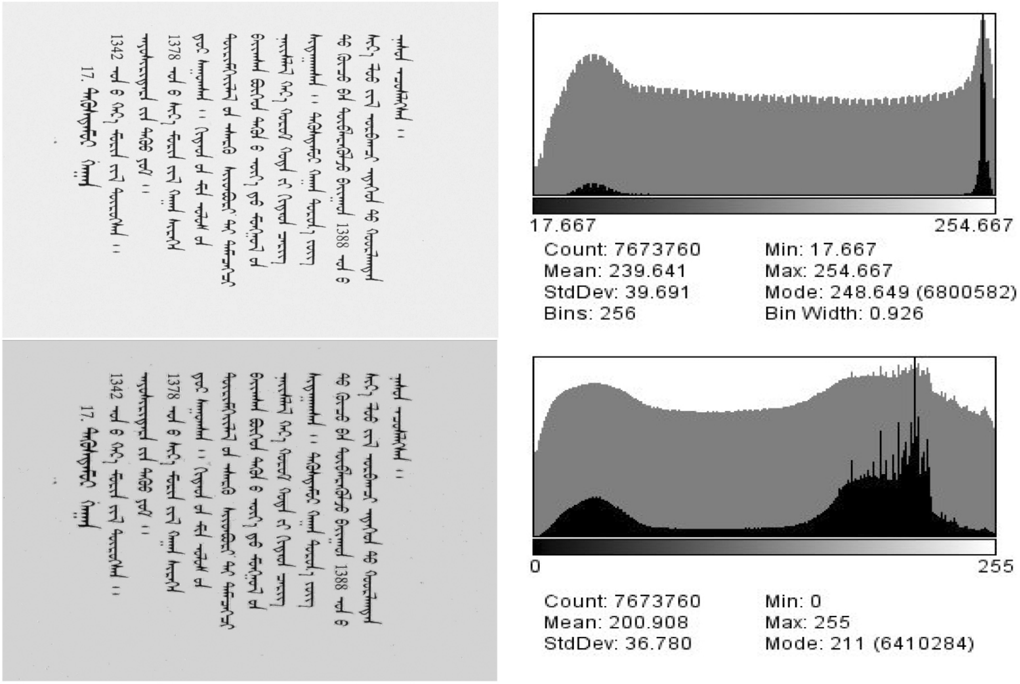


Fig. 3. Example of histogram equalization process

An image histogram acts as a graphical representation of the intensity distribution in an image. It plots the number of pixels for each intensity value. The histogram for a very dark image will have the majority of its data points on the left side and center of the graph. Conversely, the histogram for a very bright image with few dark areas will have most of its data points on the right side and center of the graph, so the contrast in an image will be improved by using histogram equalization (Farahmand et al., 2013). Histogram-based methods solve most of the fuzzy logic-based method's problems. The POSHE (partially overlapped sub-block histogram equalization) algorithm is a typically adaptive local histogram equalization algorithm; its effective pixel number, to be calculated for histogram equalization, extends from one to multiple, thus reducing the local computing times (Kim, Kim, L, & Hwang, 2001). The global histogram equalization algorithm cannot overcome the problem of local degradation caused by a different depth of field in a dust image, thus the PAL fuzzy algorithm can be adopted to enhance the local parts of an image, but this method can result in a partial block effect (Hao, 2008). Band-limited local histogram equalization is that which controls the range of the local contrast by limiting the height of local histogram, for reducing the amplification of noise and over-enhancement of the local contrast. As a result a better observed effect of image can be achieved. R. Cromartie proposed the local histogram equalization method with a limited band (Bao-ping et al., 2005). This method not only considers the histogram within the window, but also contemplates the area outside the window.

5. Morphology Based Methods

Mathematical morphology is a powerful means of enhancing uneven backgrounds. Processes in this group search for noise patterns, which appear as shadows in the background, with defined structuring elements (Farahmand et al., 2013).

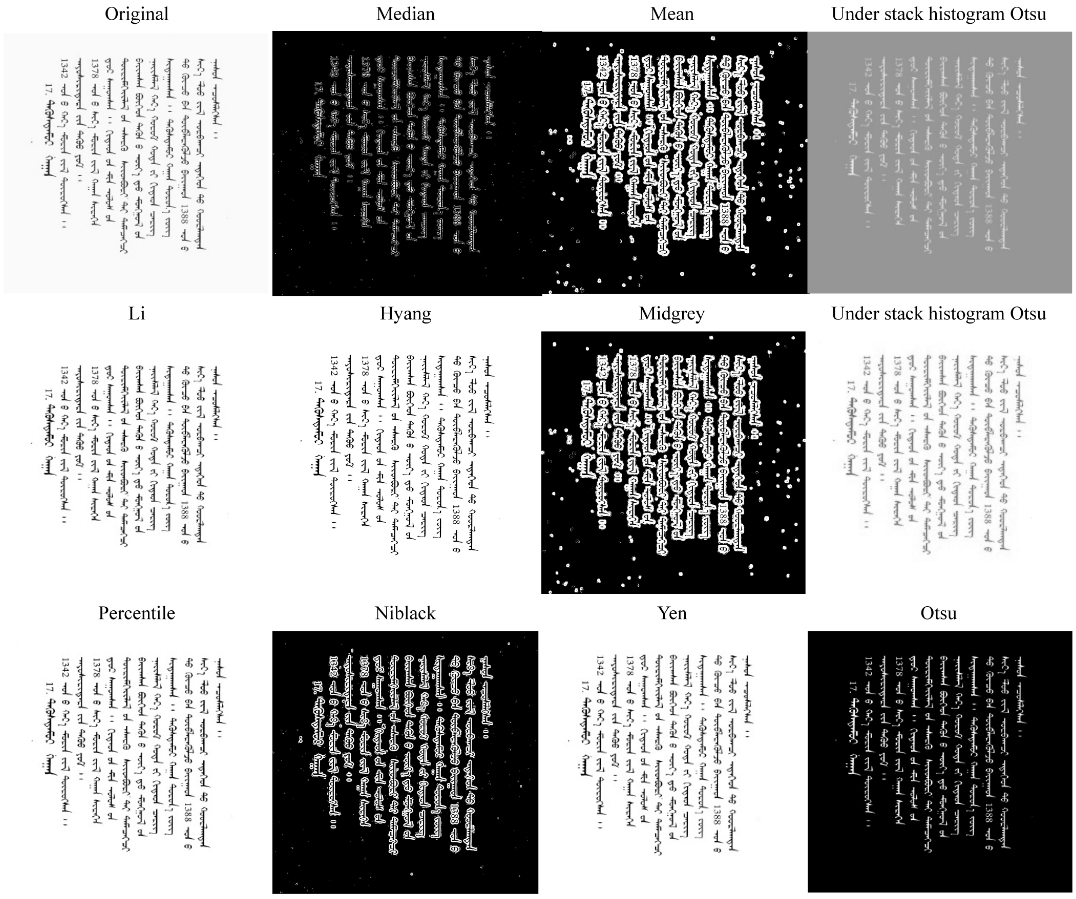
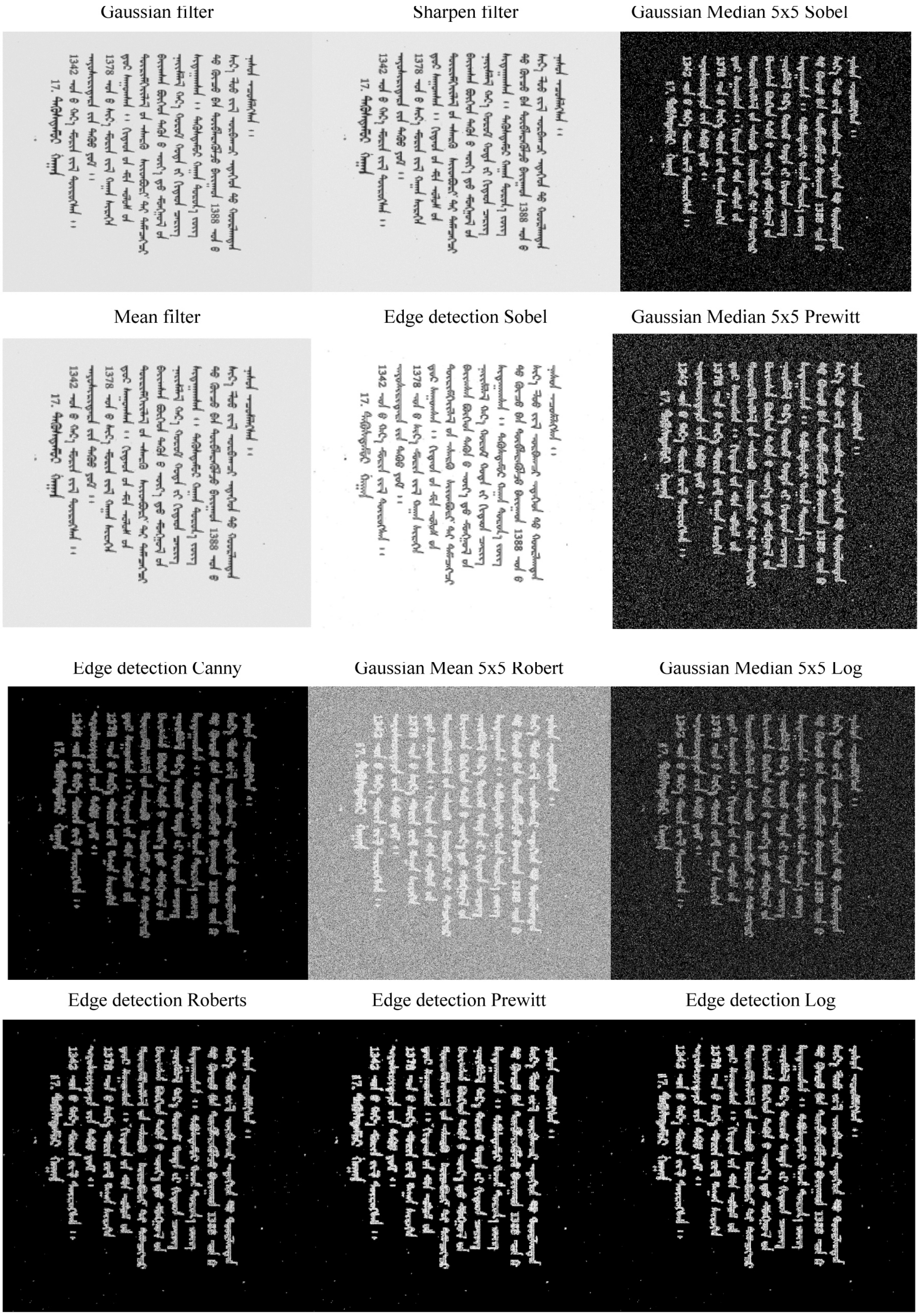


Fig. 4. Result of Mathematical morphology based methods

After discovering these patterns in one or more steps, morphological operators, like thickening and pruning remove the shadows. Some algorithms in this group start with a pre-processing stage. The Shadow Location and Lightening (SL*L) method uses thickening to highlight features that cause shadows in images, then pruning is used to remove the shadows (Nomura et al., 2009). With an even background without noise, binarization can also be done using higher quality or even global methods like Otsu, which produces better results (Farahmand et al., 2013). Otsu's Method of Global thresholding uses only one threshold value, which is estimated based on statistics or heuristics on global image attributes, to classify image pixels into foreground or background. The major drawback of global thresholding techniques is that they cannot differentiate those pixels which share the same gray level but do not belong to the same group (Feng & Tan, 2004). They work well with dearly scanned images, but perform unsatisfactorily for those poor quality images that have low contrast and non-uniform illumination (Otsu, 1979). Otsu's thresholding method involves iterating through all the possible threshold values, and calculating a measure of spread for the pixel levels on each side of the threshold, i.e. the pixels that either fall in the foreground or background. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum (Otsu, 1979).

6. Genetic Algorithm Based Methods

The majority of difficulties arise during the separation of characters from the background. Backgrounds can have complex variations and a variety of degradations (Farahmand et al., 2013). Genetic Algorithm has many advantages in obtaining the optimized solution in image processing. It has been proven to be the most powerful optimization technique in a large space. Various tasks from basic image contrast and level of detail enhancement, to complex filters and deformable models parameters are solved using this paradigm (Paulinas & Usinskas, 2007). The algorithm allows for performing a robust search without trapping in local extremes. The genetic algorithm allows for performing a robust search to find the global optimum. The result of the optimization depends on the chromosome encoding scheme and involvement of genetic operators as well as on the fitness function (Paulinas & Usinskas, 2007). Kohmura and Wakahara (2006) extended previous work and used the algorithm for color images. A filter bank of 17 well-known filters (mean, min, max, Sobel, etc.) was created in this approach to search for an optimal filtering sequence. There are some problems, however, in using a genetic algorithm. The first is that the optimization procedure is rather slow, as every fitness evaluation requires the comparison of two images. The second problem is the algorithm's inability to automatically select appropriate filters for the optimization procedure (Paulinas & Usinskas, 2007). In 2010 (Deborah & Arymurthy, 2010) genetic algorithms were used to estimate the degradation function of an image. A degradation model has a degradation function that, together with an additive noise term, operates on an input image to produce a degraded image (Ganchimeg & Turbat, 2014). In general, the more we know about the degradation function and the additive noise term, the better we are able to restore the image (Gonzales & Woods, 2002).



Gaussian filter

Sharpen filter

Gaussian Median 5x5 Sobel

Mean filter

Edge detection Sobel

Gaussian Median 5x5 Prewitt

Edge detection Canny

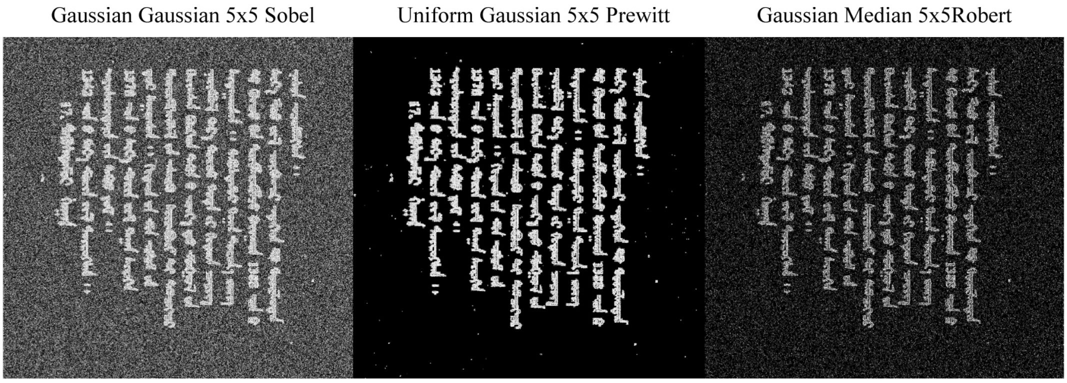
Gaussian Mean 5x5 Robert

Gaussian Median 5x5 Log

Edge detection Roberts

Edge detection Prewitt

Edge detection Log



Gaussian Gaussian 5x5 Sobel Uniform Gaussian 5x5 Prewitt Gaussian Median 5x5Robert

Fig. 5. Result of various filters

7. Marginal Noise

Marginal noises are dark shadows that appear in the vertical or horizontal margins of an image. This type of noise is the result of scanning thick documents or the borders of pages in books; it can be textual or non-textual. Methods to remove marginal noise can be divided into two categories. The first category identifies and removes noisy components; the second focuses on identifying the actual content area or page frame of the document (Zhang & Tan, 2001).

- *Identifying Noise Components*
  The methods in this group search for the noise patterns in an image by extracting their features, then removing areas which contain those patterns (Fan, Wang, & Lay, 2002; Peerawit & Kawtrakul, 2004).

- *Identifying Text Components*
  Another group of methods finds the page frame of the document, which it defines as the smallest rectangle that encloses all the foreground elements of the document image. This group performs better than the previous one because searching for text patterns is easier than searching for the features of noise in a document (Shafait et al., 2008; Shafait & Breuel, 2009).

8. Existing Challenges Problem Solving

In order to evaluate the proposed approach, we collected an large amount of historical document images. In more detail, we formed a collection of 233 document images that included those taken from Mongolian National Archive of old documents. The quality of materials that were kept in the archive for many years was very poor and each item was first restored and then transferred to a virtual document type (Ganchimeg, 2013).

Table 1. Details about the condition of the document collection used in this study (Total number of document images is 13312)

In good condition	6.65%
Degradation problems	
Uneven illumination	65.17%
Holes on the page	9.63%
Seepage ink	41.07%
Stain, smearing	42.7%
Page crumple	32.95%
Other problems	
Mixed text (print/handwrite)	17.3%
Presence of page lines	5.87%

In this collection, we attempted to cover a wide range of document types, which included both printed and handwritten, with variable degrees of degradation as well as diversity in the form of included information (different languages, musical notation, images, etc.). Therefore, a representative amount of degradation problems is included in this collection, similar to any historical document collection. Table 1 shows details about the condition of the document images used in this study. As can be seen, only a few documents are considered to be in good condition (e.g., homogeneous background noise not affecting the readability of the document). These documents assist the evaluation of a noise removal algorithm to a considerable degree, since they are not drastically affected. The rest of the documents have at least one problem, due to either degradation or the content of the document. The most common problems are uneven background illumination and seepage of ink.

9. Discussion

In this paper we presented a hybrid binarization approach aimed at removal of background noise from historical and ancient documents. We attempted to combine the advantages of global and local thresholding, that is, better adaptability of various kinds of noise at different areas of the same image based on low computational and time cost. The evaluation results, using a historical document collection, indicate that the proposed approach is able to deal with hard cases while maintaining precision on a high level. Thus, it can be used in the framework of libraries willing to provide public access to their historical document collections, as well as a preprocessing step in document image analysis systems. After a thorough examination of our experimental results, important observations can be summarized follows:

- Using Niblack’s approach, the resulting binary image generally suffers from a great amount of background noise, especially in areas without text.

- The approach of Sauvola et al. Solves the background noise problem that appears in Niblack's approach but in many cases characters become extremely thin and broken.
- The global thresholding technique of Otsu is not satisfactory for degraded documents that exhibit local variance problems.
- Although the approach of Bernsen et al. performs well in most of the testing cases, we observed occasions that a great amount of noise remains in the resulting image as well as that characters become broken.
- For binarization techniques, by choosing the most suitable threshold and parameter values, results show that Niblack's and Sauvola's algorithm work better than Otsu's global methods and Bernsen methods.
- Surprisingly, even though our aim is to enhance the image quality by applying Histogram Equalization, each equalized image shows worse results than the original gray scale images. From the findings, we can conclude that histogram equalization does not work on every historical document image type due to illumination and degraded paper problems.
- The major problem of recognizing the symbol (Mongolian script) is to identify dots and curves in each symbol. Because the number of dots and curve forms in each Mongolian writing character represents a different type of character, this results in misinterpretations or different meanings.
- Filtering techniques such as Gaussian, Sharper and Mean filters also impact to the quality of images. For comparison, the Gaussian filter is better than other filters in terms of providing smooth and clear images.
- Global threshold binarization has the poorest result, performing worse than any of the local thresholding algorithms.
- Finally, researchers could improve the post processing step by adding edge detection techniques and it can be further enhanced by an innovative image refinement technique and a formulation of a proper class method.

References

Bao-ping, W., Huai-liang, L., Nan-jing, L., & Wei-xin, X. (2005). A novel adaptive image fuzzy enhancement algorithm. *Xi'an*, *32,* 307-313.

Bernsen, J. (1986). Dynamic thresholding of gray-level images, *Proceedings 8th International Conference on Pattern Recognition,* 1251-1255.

Deborah, H., & Arymurthy, A. (2010). Image enhancement and image restoration for old document image using genetic algorithm. *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies,* 108-112. doi:10.1109/ACT.2010.24

Fan, K., Wang, Y., & Lay, T. (2002). Marginal noise removal of document images. *Pattern Recognition, 35*(11), 2593–2611. doi:10.1016/S0031-3203(01)00205-9

Farahmand, A., Sarrafzadeh, A., & Shanbehzadeh, J. (2013). Document image noises and removal methods. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2013, 1*, 436-440.

Feng, M., & Tan, Y. (2004). Adaptive binarization method for document image analysis. *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), 1,* 339-342. doi:10.1109/ICME.2004.1394198

Fisher, R., Perkins, S., Walker, A., & Wolfart, E. (2003). Histogram equalization. Retrieved from http://homepages.inf.ed.ac.uk/rbf/HIPR2/histeq.htm

Ganchimeg, G. (2013). Application exhibits of historical virtual museum. *ICEIC 2013*, 188-190.

Ganchimeg, G., & Turbat, R. (2014). Detection of edges in color images. *Journal of IEEK Transactions on Smart Processing and Computing*, *3*(6), 345-352. doi:10.1109/IFOST.2013.6616886

Gatos, B., Pratikakis, I., & Perantonis, S. (2004). An adaptive binarization technique for low quality historical documents. *Document Analysis Systems VI Lecture Notes in Computer Science, 3163,* 102-113. doi:10.1007/978-3-540-28640-0_10

Gatos, B., Pratikakis, I., & Perantonis, S. (2006). Adaptive degraded document image binarization. *Pattern Recognition*, *39*, 317-327. doi:10.1016/j.patcog.2005.09.010

Gonzales, R. C., & Woods, R. E. (2002). *Digital Image Processing 2nd Edition.* New Jersey: Prentice-Hall.

Hao, N. B. (2008). Fuzzy enhancement algorithm based on rough fuzzy sets theory for the medical volumetric data, Micro-electron. *Com put*, *25*, 137-140.

Kim, J., Kim, L., & Hwang, S. (2001). An advanced contrast enhancement using partially overlapped sub-block histogram equalization. IEEE Trans. Circuits Syst. Video Technol. *IEEE Transactions on Circuits and Systems for Video Technology*, *11*, 475-484. doi:10.1109/76.915354

Kohmura, H., & Wakahara, T. (2006). Determining optimal filters for binarization of degraded Characters in Color Using Genetic Algorithms. *18th International Conference on Pattern Recognition (ICPR'06), 3,* 661−664. doi:10.1109/ICPR.2006.446

Kuppannan, J., Rangasamy, P., Thirupathi, D., & Palaniappan, N. (2006). Intuitionistic fuzzy approach to enhance text Documents. *2006 3rd International IEEE Conference Intelligent Systems,* 733-737. doi:10.1109/IS.2006.348511

Niblack, W. (1986). In an introduction to digital image processing. *Englewood Cliffs* (p. 198), N.J.: Prentice-Hall International.

Nomura, S., Yamanaka, K., Shiose, T., Kawakami, H., & Katai, O. (2009). Morphological preprocessing method to thresholding degraded word images. *Pattern Recognition Letters, 30*(8), 729-744. doi:10.1016/j.patrec.2009.03.008

Otsu, N. (1979). A threshold selection method form gray-level histograms. *Proceedings of the 1986 IEEE Transactions Systems*, *9*(1), 62-66.

Paulinas, M., & Usinskas, A. (2007). A survey of genetic algorithms applications for image enhancement and segmentation. *Information Technology and Control*, *36*(3), 278-284.

Peerawit, W., & Kawtrakul, A. (2004). Marginal noise removal from document images using edge density. Proceedings of Fourth Information and Computer Eng. Postgraduate Workshop.

Sauvola, J., & Pietikainen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, *33*(2), 225-236. doi:10.1016/S0031-3203(99)00055-2

Shafait, F., & Breuel, T. (2009). A simple and effective approach for border noise removal from document images. *2009 IEEE 13th International Multitopic Conference,* 126-137. doi:10.1109/INMIC.2009.5383115

Shafait, F., Beusekom, J., Keysers, D., & Breuel, T. (2008). Document cleanup using page frame detection. *IJDAR International Journal of Document Analysis and Recognition (IJDAR)*, *11*(2), 81-96.

Yoo, S., & Shin, S. (2014). Design of object-based information system prototype. *International Journal of Knowledge Content Development & Technology, 4*(1), 79-91.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, *8*(3), 338-353.

Zhang, Z., & Tan, C. (2001). Recovery of distorted document images from bound volumes. *Proceedings of Sixth International Conference on Document Analysis and Recognition,* 429-433. doi:10.1109/ICDAR.2001.953826

• About the authors:

Ganchimeg Ganbold
E-mail: ganaa_mzb@yahoo.com

Key Achievements:

- *PUBLICATION*

1. Flash programming, 2006, Press of “Shorkhon shar”, UB, Mongolia
2. Data structure handbook, 2008, Press of “Enigma”, UB, Mongolia
3. C++ programming language, 2011, Press of “Soymbo”, UB, Mongolia /copyright certificate

- *DEVELOPED SOFTWARES AND WEB SITES (2011~2013)*

1. The text data parameters processing algorithm and software, № 2013/5591, copyright certificate
2. Organize textual data database optimization algorithm, № 2013/5590, copyright certificate
3. Academic scheduling and testing software, № 2011/223

- *PUBLISHED PAPERS (2013~2015)*

1. G.Ganchimeg, “Обработка монгольской текстовой информаций ”, The 3rd Russian-Mongolian Conference for Young Scientists on Mathematical Modeling, Computing Technologies and Control 2015, (accepted), June 30, 2015, pp
2. G.Ganchimeg,”History Document Image Background Noise and Removal Methods”, International Journal of Knowledge Content Development & Technology, (accepted), Vol.5, №2, 2015, pp
3. G.Ganchimeg, “Image edge detection algorithm for linear and parametric model”, Conference on Mongolian Information Technology, MMT 2015, pp 127-131, Mongolia
4. G.Ganchimeg, R.Turbat, “Image processing development and 3D printing technology”, Journal of scientific transactions, MUST, 2015, pp 189-191, № 07/170
5. G.Ganchimeg, “History Document Image Processing”, Journal of scientific transactions, MUST,

2015, pp 62-67

6. G.Ganchimeg, R.Turbat, “Detection of Edges in Color Images”, Journal of IEEK Transactions on Smart Processing and Computing, Vol.3, №6, Dec30, 2014, pp 345-352
7. G.Ganchimeg, R.Turbat, “ Color Edge Detection with The Linear Fitting Algorithm”, The 5th International Conference on Creative Science and Technology ICICT-2014, MUSTAK, pp 241-245
8. G.Ganchimeg, “Comparison of Image Edge Detection Techniques”, The 5th International Conference on Creative Science and Technology ICICT-2014, MUSTAK 2014, pp 236-240
9. G.Ganchimeg,”Research of multimedia data processing methodology”, Conference on Mongolian Information Technology, MMT 2014, pp 206-2012, Mongolia
10. G.Ganchimeg, “Image processing methods”, Journal of scientific transactions, MUST, 2013, pp 165-171, № 14/147
11. G.Ganchimeg, R.Turbat, “The Processing of Multimedia Data Using Image Processing Techniques”, International Conference, Information and Media Technology-IMT 2013, 20 Dec, Ulaanbaatar, Mongolia, № 23
12. G.Ganchimeg, “Museum websites and museum visitors in Mongolia”, The 4th International Conference on Creative Science and Technology, ICCST2013, pp 232-234
13. G.Ganchimeg, “Representing Narrative in Multimedia Information Systems”, The 8th International Forum on Strategic Technology 2013, Proceedings volume II, pp 196-200
14. G.Ganchimeg, “Application exhibits of historical virtual museum“, International Conference on Electronics Information and Communication ICEIC-2013, pp 188-189, Indonesia
15. G.Ganchimeg, “Multimedia information processing”, E-Governance and Information Technology, E-IT 2013, pp 159-161, Mongolia