# Document Image Noises and Removal Methods

Atena Farahmand, Abdolhossein Sarrafzadeh, and Jamshid Shanbehzadeh

***Abstract*- 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. After a brief introduction, this paper reviews noises that might appear in scanned document images and discusses some noise removal methods.**

***Index Terms*— Pre-processing document noise, OCR, noise removal algorithms**

## I. INTRODUCTION

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. 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. Examples of noise in scanned document images are as follows. The page rule line is a source of noise which interferes with text objects. The marginal noise usually appears in a large dark region around the document image and can be textual or non-textual. Some forms of clutter noise appear in an image because of document skew while scanning or are from holes punched in the document, or background noise, such as uneven contrast, show through effects, interfering strokes, and background spots, etc. Next, we will discuss each type in detail.

## II. RULED LINE NOISE

Handwritten documents are often written on pre-printed, lined paper. The lines can cause the following challenges: (i) the ruled lines interfere with and connecting to the text; (ii) variable thicknesses in the ruled lines cause problems for the noise removal algorithms; (iii) broken ruled lines cause problems for algorithms detecting them; (iv) some letters, for example 'z', which have horizontal lines are removed by the algorithms as they are incapable of detecting differences between them and the ruled lines.

Several methods have been proposed for ruled line removal. The methods can be divided into three major groups. First, there are mathematical morphology-based methods that depend on prior knowledge. The second group contains methods which employ Hough Transform to extract text features and to find lines in every direction. The methods in the last group use Projection Profiles to estimate lines and, hence, reduce the problem's dimensions, which then improves the accuracy of the first step in some methods of noise removal. We will discuss each group in detail.

### *A. Mathematical Morphology Based Methods*

The mathematical morphology-based methods are limited by the design and application of the structuring elements which often require knowledge of the font size or use trial and error. Structuring elements are used to probe an image, and draw conclusions on how they fit or miss the shapes in the image. Following that step, some operations such as dilation are used to highlight the extracted features from the patterns in order to remove them more easily.

Methods in this group are based on tracing line like structures as candidates for rule lines for removal [1]. In these methods, a structuring element is used to find the line patterns to facilitate removal of the ruled lines by dilation and erosion. Because the structuring elements are designed for special purposes, these methods are incapable of handling large variations in the thickness of the ruled lines. On the other hand, with these methods no difference is perceived between the ruled lines and characters with horizontal strokes (such as 'z'), so removal of too many text pixels makes the recognition phase more difficult.

### *B. Hough Transform Based Methods*

The purpose of Hough Transform is to find imperfect instances of objects within a certain class of shapes using a voting procedure. The voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm to compute Hough Transform. It can be used to find straight lines, such as ruled lines, in an image. By extracting the dominant features of an image, Hough is able to find lines in every direction; this group of methods, therefore, is robust against document rotation as earlier group. Methods using Hough Transform are computationally expensive but are more robust against

Manuscript received Jan 26, 2013; revised Jan 30, 2013.

A. Farahmand is an M.Sc. student with the Department of Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, I.R. Iran (e-mail: farahmand.atena@yahoo.com).

A. Sarrafzadeh is an Associate Professor and Head of Department of Computing, Unitec Institute of Technology, New Zealand (email:hsarrafzadeh@unitec.ac.nz).

J. Shanbehzadeh is an Associate Professor with the Department of Computer Engineering, Faculty of Engineering, Kharazmi University (Tarbiat Moallem University of Tehran), Tehran, I.R. Iran (phone: +98 26 34550002; fax: +98 26 34569555; e-mail:jamshid@tmu.ac.ir).

noise; they also cope better with broken lines in comparison to other methods.

A Hough Transform-based method was proposed to remove ruled lines in 1990 [2]. However, the method had problems which were mentioned earlier, so Random Hough Transform was proposed which performed better but, because of the high computational cost, neither one is used.

### C. Projection Profile based Methods

Projection Profile- based methods work by creating a horizontal histogram in which the hills of the histogram are the center locations of the horizontal ruled lines. Projection profiles ignore the line's thickness, therefore, in the removal phase, the characters with horizontal strokes will be broken up. Another problem with this group of methods is sensitivity to rotation. However, in comparison to the methods mentioned before, reducing the problem's dimensions makes this group faster.

The successful methods in this group have two phases [3,4]: First, the projection profile of an image helps to estimate the ruled lines. Second, we make our estimation more accurate using some other methods such as searching vertical run lengths [4]. These groups of methods solve the third problem of ruled lines, as mentioned earlier.

## III. MARGINAL NOISE

Marginal noises are dark shadows that appear in 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. Figure 1 shows two sorts of marginal noise. 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.

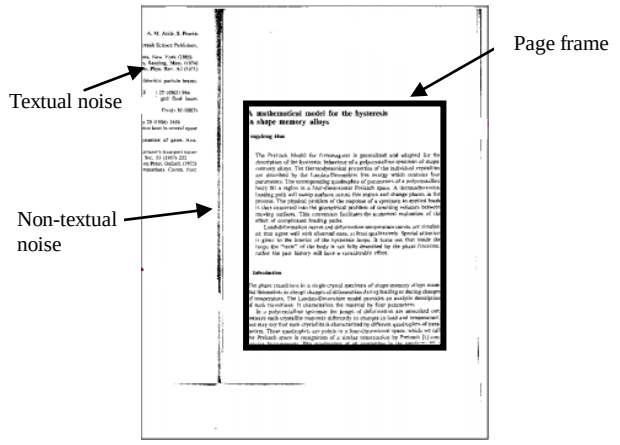



Fig. 1. Example of marginal noise

### A. Identifying Noise Components

The methods in this group search for the noise patterns in an image by extracting its features, then remove areas which contain those patterns. Zheng Zhang et al.'s [5] method employed vertical projection to recover document images that contain marginal noise, and decided whether the marginal noise was on the left or right side of the image based on the location of peaks in the profile. Then, by using extracted features, it detects the boundary between the shadows and cleans the area. However, this method suffers from the following problems:

1. Because of using features like black pixels, in images that have marginal noise areas which are smaller than the text areas, there is no peak in projections to locate marginal noise. Thus, it is not suitable for noises with variable areas.

2. Because of ignoring the extraction of features in horizontal directions, this method is incapable of locating marginal noises in the horizontal margins of a page.

To overcome these problems, another algorithm was proposed in 2002 [6]. This algorithm has three steps:

Step 1: Resolution reduction

Step 2: Block splitting to find possible local boundaries between connected blocks

Step 3: Block identification to determine which blocks contain marginal noise

In 2004, Peerawit [7] employed Sobel edge detection and identified noises to be removed by comparing the edge density of marginal noise and text. This method uses density as the selected feature because edge density is higher in noise than text. If the document has only non-textual marginal noise, this method is unable to find significant difference between edge densities and, hence, is unable to detect marginal noise. Moreover, this method is not suitable for detecting marginal noise in a small area.

### B. 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.

In 2008 Shafait [8] proposed a method that works in two steps. First, a geometric model is built for the page frame. Then a geometric matching method is employed in finding the globally optimal page frame with respect to a defined quality function. Although the method works well in practice, it requires prior extraction of the text line which increases the computational cost and is hard to implement. To overcome the shortcomings of this method, another algorithm was proposed that works in three steps [9]:

Step 1: A black filter is used; if the black regions are bigger than a pre-defined threshold area, it selects them.

Step 2: Connected component removal is used; first, all connected components are extracted from the image after applying a black filter. All components that are close to the border of the image are considered noise and, hence, removed from the image. Selecting an appropriate value for the threshold is dependent on prior knowledge.

Step 3: A white filter is used; it extracts features similar to the black filter and removes everything up to the border if it finds a big white block.

## IV. CLUTTER NOISE

Clutter noise refers to unwanted foreground content which is typically larger than the text in binary images. This results from numerous sources such as punched holes, document

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

image skew, or connecting huge amounts of pepper noise. The significant feature of clutter noise is that it is larger than the text objects in the document image. One of the major challenges facing clutter is its connectivity with text. Clutter often touches or overlaps some parts of the text which reduces segmentation and recognition accuracy in OCR systems.

Wang and Fan [10] proposed a method that can detect and remove clutter noise. The proposed method reduces the resolution of the image, splits it into blocks and detects blocks that contain noise based on the three assumptions of shape, length and position. The technique performs fairly well to remove the marginal noise only without attached text, but assumptions cause some limitations in detecting all types of clutter noise in an image.

Agrawal [11] proposes a method that is independent of clutter's position, size, shape and connectivity with text. A half residual image is achieved as the result of analysis of the distance transform, and by removing parts which are less than half of the maximum distance measured. The clutter element is then identified with an SVM, Support Vector Machine, classifier.

## V. STROKE LIKE PATTERN NOISE

Stroke Like Pattern Noise (SPN) is a kind of noise which is independent of the size or other properties of the text in the document image. SPN is similar to diacritics so its presence near textual components can change the meaning of a word. This noise is formed primarily due to the degradation or unsuccessful removal of underlying ruled lines that interfere with the foreground text, or it is formed by the remaining clutter noise after clutter removal approaches.

The situation is challenging where the ruled lines are broken and degraded, as they cannot be perceived in straight lines even by the human eye. Thus, techniques like Hough Transform and projection profiles are inappropriate in such cases. Furthermore, because of their similarity in shape and size to smaller text components, morphology-based removal approaches are unsuitable because the successive erosion and dilation steps needed tend to degrade the text.

In 2011, Agrawal [12] described the difference between SPN and ruled-lines for the first time and proposed a solution. The method works in two steps. First, independent prominent text component features are extracted with a supervised classifier, then it uses their cohesiveness and stroke-width properties to filter smaller text components with them using an unsupervised classification technique.

## VI. SALT AND PEPPER NOISE

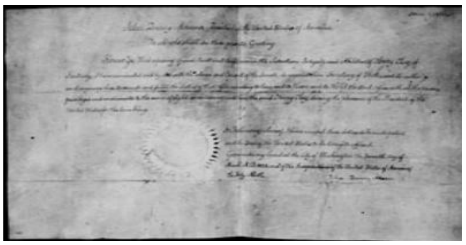
Pepper noise can appear in a document image during the conversion process and is also caused by dirt on the document. This noise can be composed of one or more pixels but, by definition, they are assumed to be much smaller than the size of the text objects. Isolated pepper noise can be removed by simple filters like median [13] but if they are larger than that, algorithms like k-fill [14] or morphological operators [15] will be more effective for noise removal.

Printed documents come in many forms and in infinite varieties of writing ink, and salt noise looks like a lack of ink in the document image. If the fragmentation is very high, it reduces segmentation and recognition accuracy.

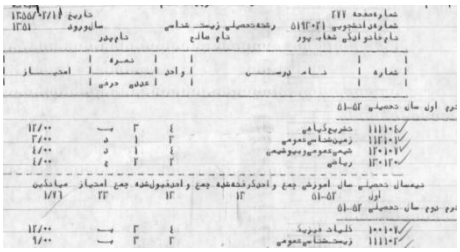
Isolated salt noise can be removed by simple filters like median. In 2007 [16], a morphological-based method was proposed. This method solved one of the most important problems of morphology-based approaches by using a learning phase for finding the parameters of a suitable structuring element. After that, a dilation operator is used to fill places where there is a lack of ink. This method experienced some problems such as a high execution time because of the learning phase, and produced undesirable connections between some characters, particularly in a situation where the fonts were very thick.

## VII. BACKGROUND NOISE

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 (see Fig. 2). These problems cause challenges similar to those in an OCR system. Such degradations can destroy the blank spaces between lines and words. There are many methods in the literature to enhance background degradations in document images, therefore, we have divided the methods into five major groups:



(a)



(b)

Fig. 2. Examples of background degradation

### A. Binarization and Thresholding Based Methods

One of the methods to enhance background quality of gray scale images employs thresholding and binarization techniques. Some resources divide thresholding techniques into two major groups. The methods in the first group use global algorithms which employ global image features to determine appropriate thresholds to divide image pixels into object or background classes. 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 [17]. However, the methods of the second group are much

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

slower than the first, but their accuracy is higher.

### B. Fuzzy Logic Based Methods

Enhancing image quality using fuzzy logic operators is based on mapping gray levels of image to fuzzy space, and we know that defining an appropriate membership function requires experience and prior knowledge. Enhancement with fuzzy operators employs weighting features proportional to some image features, like average intensity to increased contrast.

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 will be improved.

Later, in 2006, a method was proposed to improve Tizhoosh's algorithm by using an intensification operator on the first and second type of IFSs, which is defined as follows [18]:

Let X be a non-empty set. An IFS A in X is defined as an object of the form $A = \{\langle x, \mu_A(x), \gamma_A(x)\rangle : x \in X\}$ where the fuzzy sets $\mu_A : X \to [0,1]$ and $\gamma_A : X \to [0,1]$ denote the membership and non-membership functions of A respectively, and $0 \le \mu_A + \gamma_A \le 1$ for each $x \in X$.

### C. Histogram Based Methods

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. Histogram-based methods solve most of the fuzzy logic-based method's problems.

In 2001, POSHE (Partially Overlapped Sub-Block Histogram Equalization) was proposed [19]. In this method, the image is divided into blocks, then in each block histogram equalization is done. This method achieves better performance in contrast enhancement than former methods because of using local feature extraction.

In 2005, Leung et al. used POSHE with generalized fuzzy operators [20]. Using GFO alone, no improvement in image contrast is achieved since there is no significant difference of gray level in the image. Hence, this method uses a pre-processing technique to enhance the objects of interest so that the background can be significantly distinguished from the objects of interest. This method proposes two methods of pre-processing. The first one is histogram equalization and the second is POSHE. A GFO operator is then used to enhance background quality.

### D. Morphology Based Methods

Mathematical morphology is a powerful methodology for enhancing uneven backgrounds. The operators are powerful tools for processing and analyzing shapes with structural features like borders, area etc. Methods in this group search for noise patterns, which appear as shadows in the background, with defined structuring elements. Then, in one or more steps, morphological operators like thickening and pruning…remove shadows. Some algorithms in this group start with a pre-processing stage.

In 2009 [21], the Shadow Location and Lightening (SL*L) method was proposed. This method uses thickening to highlight features that cause shadows in images, then uses pruning to remove the shadows. With an even background without noise, binarization can also be done using higher quality or even global methods like Otsu which will produce better results.

In 2007 [22], a method that uses mathematical morphology and a Wiener filter was proposed. This method has two steps: First, a pre-processing phase is carried out by using a Wiener filter. Wiener is a low-pass filter which smoothes image in an adaptive manner; it uses a standard deviation of intensities to decide the amount of smoothness. So, despite edges, the background becomes smooth and the difference between the text and the background increases. In the second step, text patterns to be removed are found in the image by using mathematical morphology operators. This process results in an estimation of the background and, by subtracting it from the original image, an enhanced image is obtained.

### E. 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. In order to improve quality, well-known filters such as Fourier transform, Gabor filters, and wavelet transforms can be used. However, it is difficult for a single filtering technique to deal with a variety of degradations. To solve similar problems, Nagao et al. [23, 24] used GAs to construct an optimal sequence of image processing filters to extract characters from different sources. In 2006, Kohmura [25] 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.

In 2010 [26] 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. In general, the more we know about the degradation function and the additive noise term, the better we are able to restore the image [27].

## REFERENCES

[1] J. Said, M. Cheriet, and C. Suen, "Dynamical morphological processing: a fast method for base line extraction," ICDAR, pages 8–12, 1996.

[2] Lei Xu, E. Oja, and P. Kultanen, "A New Curve Detection Method: Randomized Hough Transform (RHT)," Pattern Recognition Letters, Vol.11, pp331-338, 1990.
[3] H. Cao, R. Prasad, and P. Natarajan, "A stroke regeneration method for cleaning rule lines in handwritten document images," MOCR '09: Proceedings of the International Workshop on Multilingual OCR, pages 1–10, New York, NY, USA, 2009.
[4] Zhixin Shi, Srirangaraj Setlur, and Venu Govindaraju," Removing Rule-Lines from Binary Handwritten Arabic Document Images Using Directional Local Profile," ICPR 2010: pp. 1916-1919.
[5] Zheng Zhang and Chew Lim Tan, "Recovery of Distorted Document Images from Bound Volumes," IEEE, 2001, pp. 429-433.
[6] Kuo-Chin Fan, Yuan-Kai Wang, and Tsann Ran Lay, "Marginal noise removal of document images," Pattern Recognition Society, Elsevier Science, 2002, pp. 2593-2611.
[7] W. Peerawit and A. Kawtrakul, "Marginal Noise Removal from Document Images Using Edge Density," Proceedings of Fourth Information and Computer Eng. Postgraduate Workshop, Janaury 2004.
[8] Faisal Shafait, Joost van Beusekom, Daniel Keysers, and Thomas M. Breuel, "Document cleanup using page frame detection," IJDAR, Vol. 11(2), pp. 81-96, 2008.
[9] F. Shafait and T.M. Breuel, "A Simple and Effective Approach for Border Noise Removal from Document Images," Proceedings of. 13th IEEE International Multi-Topic Conference, December 2009, pp. 126-137.
[10] K.-C. Fan, Y.-K. Wang, and T.-R. Lay, "Marginal noise removal of document images," Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pp. 317–321, 2001.
[11] M. Agarwal and D. Doermann, "Clutter noise removal in binary document images," in Proceedings of International Conference on Document Analysis and Recognition, pp. 556–560, 2009.
[12] Mudit Agrawal, David S. Doermann, "Stroke-Like Pattern Noise Removal in Binary Document Images," ICDAR 2011: pp. 17-21
[13] G. Story, L. O'Gorman, D. Fox, L. Schaper and H. Jagadish, "The rightpages image-based electronic library for alerting and browsing," Computer, vol. 25, no. 9, pp. 17– 26, September 1992.
[14] N. Premchaiswadi, S. Yimgnagm and W. Premchaiswadi, "A scheme for salt and pepper noise reduction and its application for ocr systems," W. Trans. onComp., vol. 9, pp. 351–360, April 2010.
[15] J. Serra, "Image Analysis and Mathematical Morphology," 3rd ed. Academic Press, 1983.
[16] H. Grayloo, H. Kabir, "Fixing undesirable fragmentations in Persian printed documents using morphological dilation," 12th International Conference of Computer Society of Iran, 2007.F. Shafait and T.M. Breuel, "A Simple and Effective Approach for Border Noise Removal from Document Images," Proc. 13th IEEE Int'l Multi-Topic Conf., Dec. 2009.
[17] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition, (2000), Vol. 33 Issue 2, pp. 225-236.
[18] R. Parvathi, S. Jayanthi, N. Palaniappan, and S. Devi, "Intuitionistic Fuzzy Approach to Enhance Text Documents," Proceedings 3rd IEEE International Conference on Intelligent Systems (IEEE IS '06), 2006, pp. 733-737.
[19] Kim, J.-Y., L.-S. Kim, et al., "An advanced contrast enhancement using partially overlapped sub-block histogram equalization," IEEE Trans. Cir. and Sys. for Video Technol. Vol. 11, pp. 475–484.
[20] C. Leung, K.-S. Chan, H. Chan, W. Tsui, "A new approach for image enhancement applied to low-contrast–low-illumination IC and document images," Pattern Recognition Letters, vol. 26 (6) (2005), pp. 769–778.
[21] S. Nomura, K. Yamanaka, T. Shiose, H. Kawakami, O. Katai, "Morphological preprocessing method to thresholding degraded word images," Pattern Recognition Letters, vol. 30(8), 2009, pp. 729–744. 30 (8) (2009) 729–744.
[22] H. Rajaee, "Enhancement of document images by morphological operators," 12th International Conference of Computer Society of Iran, 2007.
[23] S. Masunaga, T. Nagao, "Automatic construction of image transformation processes using genetic algorithm," Proceedings of International Conference on Image Processing, 1996, Vol.3, pp. 731 – 736.
[24] S. Aoki, and T. Nagao, "Automatic construction of tree structural image transformations using genetic programming," Proceedings of 10th Conference on Image Analysis and Processing, 1999, pp. 276 – 279.
[25] H. Kohmura, T. Wakahara, "Determining optimal filters for binarization of degraded characters in color using genetic algorithms," Proceedings of 18th International Conference on Pattern Recognition, 2006, Vol. 3, pp. 661–664.
[26] H. Deborah and A. M. Arymurthy, "Image Enhancement and Image Restoration for Old Document Image using Genetic Algorithm," Proceedings of Second International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT 2010), pp. 108-112, 2010.
[27] R. C. Gonzales, R. E. Woods, "Digital Image Processing 2nd Edition," New Jersey: Prentice-Hall, 2002.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)