International Journal of Engineering & Technology, 7 (4.36) (2018) 780-783

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Image Enhancement of Complex Document Images Using Histogram of Gradient Features

Sajan A. Jain[1*], N. Shobha Rani[2], N. Chandan[3]

[1]*Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.*
[2]*Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.*
[3]*Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.*

Abstract

Enhancement of document images is an interesting research challenge in the process of character recognition. It is quite significant to have a document with uniform illumination gradient to achieve higher recognition accuracies through a document processing system like Optical Character Recognition (OCR). Complex document images are one of the varied image categories that are difficult to process compared to other types of images. It is the quality of document that decides the precision of a character recognition system. Hence transforming the complex document images to a uniform illumination gradient is foreseen. In the proposed research, ancient document images of UMIACS Tobacco 800 database are considered for removal of marginal noise. The proposed technique carries out the block wise interpretation of document contents to remove the marginal noise that is present usually at the borders of images. Further, Hu moment's features are computed for the detection of marginal noise in every block. An empirical analysis is carried out for classification of blocks into noisy or non-noisy and the outcomes produced by algorithm are satisfactory and feasible for subsequent analysis.

Keywords: *Document images, pre-processing, marginal noise removal, Hu moment's, optical character recognition.*

1. Introduction

Removal of marginal noise from document image is one of the challenging research problem. Marginal noise usually exists at the borders of document image and results in formation of a non-uniform illumination gradient. The direct processing of complex documents through OCR for extraction of text reduces its accuracy due to its erroneous intermediate processing stages like segmentation. Marginal noise can occur in a document mainly due to scanning of broader documents or skewed orientations. It typically appears at the borders or corners or margins as large and dark regions in the document image. Marginal noise sometimes wraps up the meaningful objects in the document which introduces obstructions in performing the segmentation and recognition of those objects. It is very significant for removal of marginal noise in order to increase the accuracy of OCR. In the proposed work, the marginal noise removal involves marginal noise detection and marginal noise deletion. The proposed work removes the dark borders rectifies the illumination gradient of the images sufficing the smooth recognition process. Fig 1 depicts some of the sample document image with the above explained characteristics, collected from UMIACS Tobacco800 complex Document Image Database of university of Illinois institute of technology. The summary of the existing experimentations on the same is as follows.

Faisal shafait et al [1] had described the marginal noise that appear along the border of the pages. Including the effects produced by neighbor pages. The algorithm consists of techniques for removing textual and non-textual noise using a method of projection profile analysis. The datasets used for method is from university of Washington. Rajeev N. Verma et al [2] had contributed a method for obtaining a shading free image. Initially document images are subjected to detection of shading, dark borders and skew defect and then removed from the images as they cause many obstructions in the document recognition process. The methods applied inlcude skew removal, principal component analysis, Hough transform, border detection, convex hull. Faisal shafait et al [3] had devised an adaptive binarization algorithm which is similar to binarization but takes advantage of time which runs close to that of global thresholding method. The algorithm consists of combination of statistical constraints of sauvola method with the integral images where the mean and variance of local window is calculated. Mudit agrawal et al [4] had proposed a stroke-like pattern noise removal algorithm for double side texted document images. In the first step text component features are computed using supervised classification. This method is effective on rule-line degradation, clutter residues, marks, and degraded background. The classification of component is obtained using SVM with an RBF kernel. This technique does not aim on script or character recognition in order to carry out text extraction. Mudit agrawal et al [5] had devised an algorithm for clutter detection and removal method using SVM classifier. As removal is restrictive, the text closer to the clutter is not deleted in the procedure. This technique was tested on a collection of degraded, noisy, machine-printed and on handwritten documents with irregular and non-periodic clutter noise. Atena farahmand et al [6] had described the noises present in the scanned document that may reduce the accuracy of OCR system. This technique is applied on the marginal noise typically present in the large dark region around document image and can be textual or non-textual. This technique consists of mathematical morphology based, Hough transform to extract text features and projection profile to estimate lines. Anshul gupta et al [7] had presented an algorithm for iterative classification which automatically assigns noise labels to bounding boxes using rule based classifier. This technique is illustrated using spatial distribution and geometry which does not require dedicated image processing algorithm and also language

agnostic. Rafael Dueire et al [8] had devised an approach for monochromatic document images, which helps in removing noisy border, corrects the image orientation, calculates the skew angle and finally compress the image to required format. Big Batch, an algorithm used for various kinds of document images is used to process the digitized documents fed by production line scanner. Stamatopoulos et al [9] had presented a novel system for the enhancement of document images to automatically detect the borders in the document, cuts the noisy black borders and also noisy text from the neighboring page. The technique is applied by the combination of projection profile and connected component labeling. The noisy text region is detected using signal cross-correlation method. Thus resulting in the document images free from noisy black borders and noisy text from neighbor page. Faisal Shafait et al [10] had contributed a technique to detect the page frame of the document to clean-up the image using geometric matching algorithm helping in detection of actual page contents area and ignore the marginal noise. The algorithm is applied on UW-III database, which mainly focuses on textual noise. But, the output of algorithm still consists of some textual noise that results in an undesired text that need to be removed later. This approach shows the accuracy by removing typeset outside the computed page frame. Thai et al [11] had presented a method for removal of edge noise present in the graphical document image that needs to be removed from the document image for its accurate analysis and recognition. Kuo et al [12] had devised a novel technique to remove marginal noise comprising of marginal noise detection and marginal noise deletion. Marginal noise detection method reduces original image to smaller document image and finds the noisy region according to shape length and location of the split blocks. After the detection various removal techniques are performed such as local thresholding is applied to remove marginal noise for gray-scale images and region growing method for binary image. The accuracy is improved by removing the marginal noise accurately without destroying meaningful data. Pratiksha et al [13][14] had devised a binarization technique to remove the noise using Otsu's thresholding. This technique is also compared with Niblack and Sauvola thresholding for getting better result. The image still contains some salt and pepper noise at its margin.

It is observed that many of the existing work are based on projection profile analysis working on non-textual noise, stroke-like pattern removal algorithm on double-sided documents, and principle component analysis and are widely working on gray scale images rather on binary images.

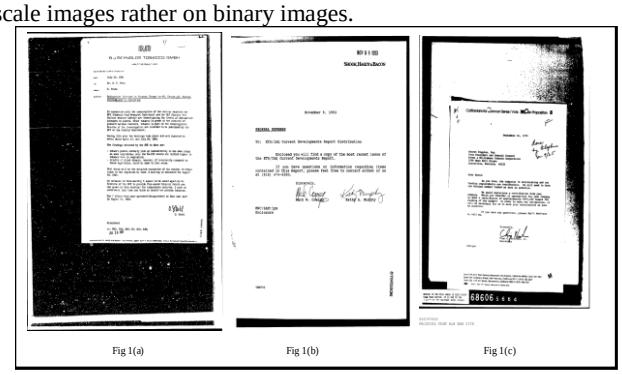

Fig 1: Input document image instances from UMIACS Tobacco800 database

2. Proposed methodology

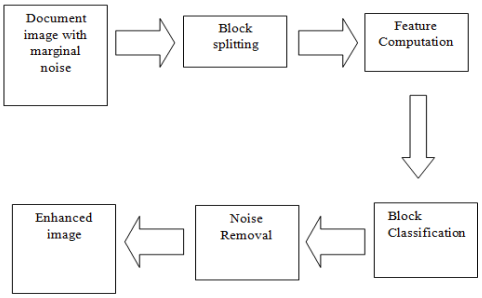

Fig. 2: Block diagram of marginal noise removal system

The proposed method for marginal noise removal in the document image is accomplished in four stages. Initially, a document image with marginal noise is assumed as input and then it is forwarded for block splitting. Further, the document is split into small rectangles of equal size and then interpreted for detection of marginal noise. Hu moments features are employed for classification of blocks with or without noise. Finally, the detected blocks are subject for noise removal to obtain an enhanced image. Fig 2 depicts the block diagram for proposed methodology.

3. Marginal noise detection

Marginal noise detection [15] includes four main steps. They are (a) Image resizing (b) Image segmentation (c) moment feature extraction and (d) block identification[17][18]. Image resizing is performed initially to reduce the size of an image to decrease the image processing time. The resized image is then segmented into small parts called blocks for easy detection and removal of noise. Moment feature extraction is performed on each obtained segment. Finally block identification is performed through inference based technique.

3.1. Image resizing

Image resizing is a technique used to reduce or increase the image dimensions suitably to segment the blocks into dimensions of 25x25 as the images present in the database are of different sizes. Usually, the image is resized to dimensions of nearest even integers in terms of rows and columns if the dimensions of image are in odd numbers.

3.2. Block splitting

Block splitting is a technique which splits a particular image into blocks. It consists of marginal noise blocks and non-marginal noise block. The features from each segmented block are computed through Hu moments feature extraction technique and further detailed empirical study is carried out to identify differential characteristics between noisy and non-noisy blocks. Fig. 3 depicts the image after applying block splitting technique on a particular image.

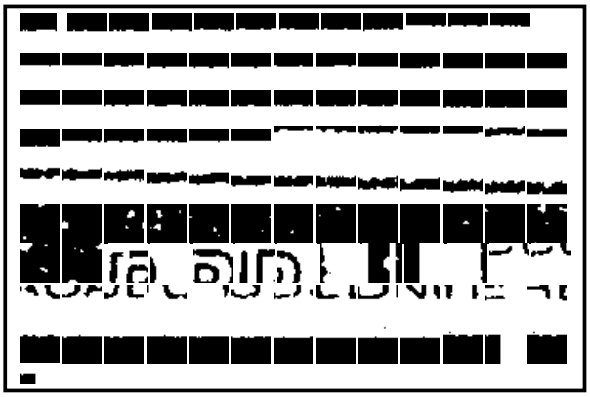

Fig. 3: Block splitting of a particular image

3.3.Block identification

After the detailed empirical study of features extracted, the Moment feature variants appears to be indefinite quantity for a marginal noise block and a definite integer in case of non-noisy blocks. Therefore, Hu moment's helps in detection of of marginal noise blocks.

4. Marginal Noise Deletion

Marginal noise deletion is the method which reverts the intensity of identified marginal noise blocks to back ground intensity. A region containing marginal noise (black pixel) and its neighbor blocks are automatically converted into background pixel i.e. white pixel. The process of transforming noisy to non-noisy blocks is carried out in multiple iterations of block identification and deletion module so as to achieve the required enhancement in document.

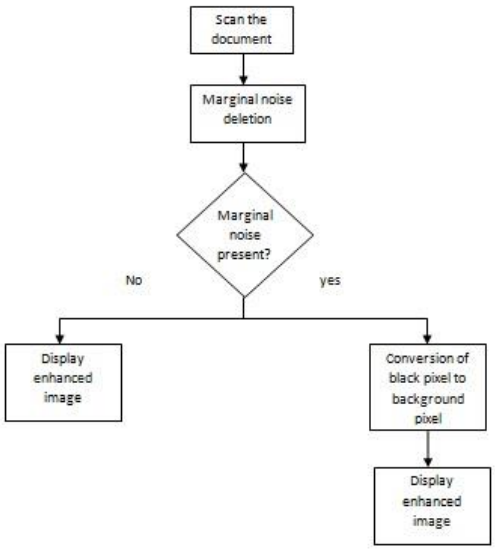

Fig. 4: decision tree for deciding the block noise removal

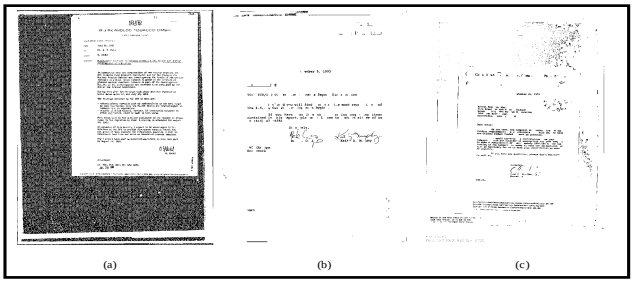

Fig. 5: Document image after first iteration

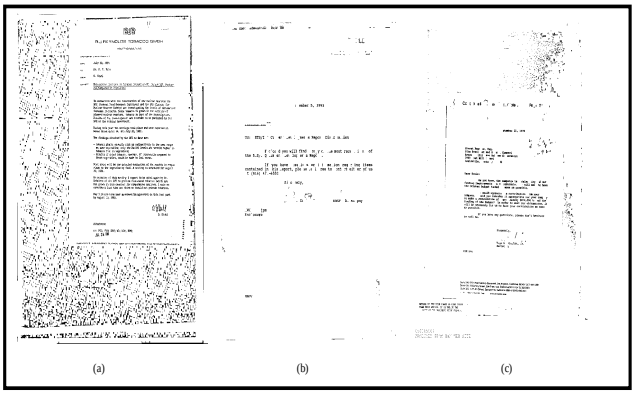

Fig. 6: Document images after second iteration

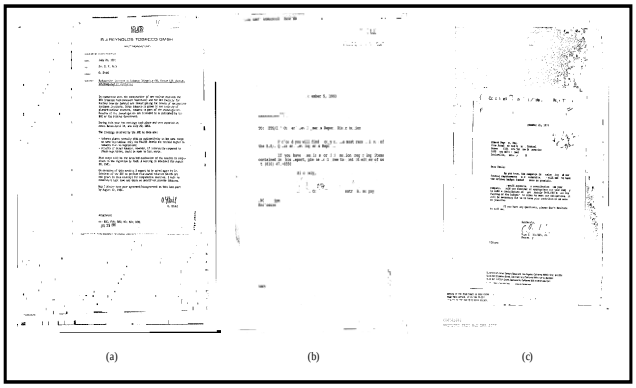

Fig. 7: Document image after third iteration

5. Experimental Analysis

The algorithm performance is analyzed by testing it on the 500 document images. Many of the existing system on non-textual noise, stroke-like pattern removal algorithm on double-sided documents, and principal component analysis have used gray scale images computations rather than binary images. In the proposed method, block splitting is used for identifying the marginal noise blocks using Hu moments and later the blocks containing marginal noise is deleted by converting it into the background pixel. The performance of the algorithm is determined subjectively in the proposed system. The result obtained after removal of marginal noise is as presented in the fig 7.

6. Conclusion

The high accuracies of OCR system can be realized only through removing the unwanted noise from the document image, which improves the accuracy of feature extraction and classification stages. The good outcome is achieved by removing the marginal noise from the borders of the image. In the proposed technique, marginal noise is detected and deleted by block identification using Hu moments method and converting the neighbor pixel to the background pixel consisting of three iteration.

References

[1] Shafait F &Breuel TM, "A simple and effective approach for border noise removal from document images", *IEEE 13th International Conference on Multi topic*, (2009), pp. 1-5.

[2] Verma RN & Malik LG, "Review of illumination and skew correction techniques for scanned documents", *Procedia Computer Science*, Vol.45, (2015), pp.322-327.

[3] Shafait F, Keysers D & Breuel TM, "Efficient implementation of local adaptive thresholding techniques using integral images", *International Society for Optics and Photonics Electronic Imaging*, (2008), pp.681510-681510.

[4] Agrawal M & Doermann D, "Stroke-like pattern noise removal in binary document images. *IEEE International Conference on Document Analysis and Recognition*, (2011), pp.17-21.

[5] Agrawal M & Doermann D, "Clutter noise removal in binary document images", *IEEE 10th International Conference on Document Analysis and Recognition*, (2009), pp.556-560.

[6] Farahmand A, Sarrafzadeh A & Shanbehzadeh J, "Document image noises and removal methods", *International Multi Conference of Engineers and Computer Scientists*, (2013), pp.1-5.

[7] Gupta A, Gutierrez-Osuna R, Christy M, Capitanu B, Auvil L, Grumbach L & Mandell L, "Automatic Assessment of OCR Quality in Historical Documents". *AAAI*, (2015), pp.1735-1741.

[8] Lins RD, Ávila BT & De Araújo Formiga A, "Big Batch–an environment for processing monochromatic documents", *International Conference Image Analysis and Recognition*, (2006), pp.886-896.

[9] Stamatopoulos N, Gatos B & Kesidis A, "Automatic borders detection of camera document images", *2nd International Workshop on Camera-Based Document Analysis and Recognition*, (2007), pp.71-78.

[10] Shafait F, Van Beusekom J, Keysers D & Breuel TM, "Document cleanup using page frame detection", *International Journal of Document Analysis and Recognition*, Vol.11, No.2, (2008), pp.81-96.

[11] Hoang TV, Smith EHB & Tabbone S, "Sparsity-based edge noise removal from bilevel graphical document images", *International Journal on Document Analysis and Recognition*, Vol.17, No.2, (2014), pp.161-179.

[12] Fan KC, Wang YK & Lay TR, "Marginal noise removal of document images", *Pattern Recognition*, Vol.35, No.11, (2002), pp.2593-2611

[13] Jadhav PD, Jadhav DR, Gite SS & Mulik V, "Enhancement of Old Degraded Documents Using Phase Base Binarization by Dip Techniqe", *International Journal of Engineering Science*, Vol.4225, (2016).

[14] Lewis D, Agam G, Argamon S, Frieder O, Grossman D & Heard J, "Building a test collection for complex document information processing", *Annual Int. ACM SIGIR Conference*, (2006), pp. 665-C666.

[15] Agam G, Argamon S, Frieder O, Grossman D & Lewis D, "The Complex Document Image Processing (CDIP) test collection", *Illinois Institute of Technology*, (2006).

[16] The Legacy Tobacco Document Library (LTDL), University of California, San Francisco, (2007).

[17] Rani NS & Vasudev T, "An Efficient Technique for Detection and Removal of Lines with Text Stroke Crossings in Document Images", *International Conference on Cognition and Recognition*, (2018), pp.83-97.

[18] Rani DANS, Vineeth P & Ajith D, "Detection and removal of graphical components in pre-printed documents", *International Journal of Applied Engineering Research*, Vol.11, No.7, (2016), pp.4849-4856.