Preprocessing Framework for Document Image Analysis

Umesh. D. Dixit
Department of Electronics & Communication Engineering,
B.L.D.E.A's V.P. Dr. P.G.Halakatti College of Engineering & Technology,Vijayapur, Karnataka – 586103.
Email: uddixit@rediffmail.com
M. S. Shirdhonkar
Department of Computer Science & Engineering
B.L.D.E.A's V.P. Dr. P.G.Halakatti College of Engineering & Technology,Vijayapur, Karnataka – 586103.
Email: ms_shirdhonkar@rediffmail.com

ABSTRACT

Preprocessing is the first step used in all the document image analysis algorithms. A well organized preprocessing could lead to better results of the analysis. This paper proposes a framework for preprocessing of document image for analysis. The frame work uses four steps such as color image to grayscale conversion, enhancement of grayscale image, binarizing the grayscale image and finally removal of clutter-noise. Horizontal and vertical projections are used to detect possible locations of clutter noise in this work. Then foreground pixels are replaced by background colored pixels based on the run length. The frame work provided better results for test images.

Keywords - : Analysis, Clutter noise, Noise removal, Preprocessing

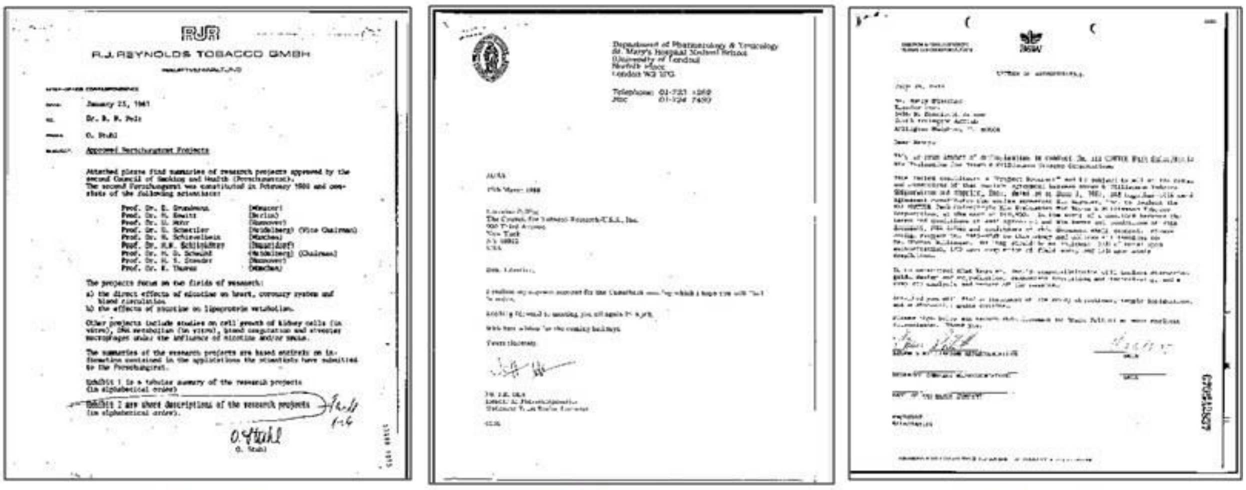
Date of Submission: Dec 10, 2018 Date of Acceptance: Jan 30, 2019

I. INTRODUCTION

Document image analysis is promising area of research. Since two decades, it has attracted lot of researchers to provide solutions to automation of document processing. A successful research in this area could lead to paperless office. Document image analysis mainly deals with recognition and extraction of textual as well as graphical components. Text to speech, signature verification, script identification, detection of name, address, pin-code etc., are the applications of document image analysis [1].

Duty noise, poor ink quality, large ink blobs present a great challenge to document image analysis and demand the well organized preprocessing steps. This paper proposes a framework for preprocessing steps. The framework includes conversions of document image, enhancement and removal of clutter noise. Generally clutter noise is an unwanted set of foreground pixels that appear in margin space of the document images. The Fig. 1 shows sample of document images with clutter noise.

The main contribution of this paper is to propose a simple and an efficient technique for removal of clutter noise from the document image. We used horizontal and vertical projections of document image to detect possible locations of the clutter noise. Then in the possible location, the foreground pixels are replaced with background pixels based on the run length. Rest of the paper is organized as follows: Section 2 provides literature review, Section 3 details the proposed framework, Section 4 discusses the results and finally section 5 concludes the work.



(a) Document 1 (b) Document 2 (c) Document 3

Fig. 1 Sample document images with clutter noise

II. LITERATURE REVIEW

Farahmand et al. [2] discuss about different types document image noise and provide review of noise removal methods. Wang et al. [3] presented fast text detection algorithm for scanned document images. They used low pass filter in their technique to remove high frequency noise, morphological operations and edge detection schemes. Then to enhance the text pixels a clipping operation is applied on L and C channels of LCH space.

Khan and Puri [4] presented a study of text detection techniques for printed documents. Peng et al. [5] delivered a method to estimate the quality of document using sparse representation. [6] presents the use of sparse representation for restoration of noisy document images. Lelore and Bouchara [7] proposed an algorithm for document image restoration using double thresholded edge detection method.

Banerjee et al. [8] employed probabilistic context model for restoration of document images. They integrated restoration and super-resolution into a single frame-work. Sgarbi et al. [9] used morphological color operators for color text restoration. A genetic algorithm for document image restoration is developed and presented in [10].

Yagoubi et al. [11] presented nonlinear anisotropic diffusion approach for document image enhancement. This method helped for better distinguish of foreground and background of text characters.

A super-resolution technique for text images is proposed by [12]. The teager filter is used to obtain super-resolution text. Document image enhancement for Farsi text documents using logarithm domain is presented in [13]. Tan ans Shen [14] proposed wavelet based technique for restoration of archived documents. They developed a method for recovering contents of handwritten documents by interfering handwriting on the reverse side caused by seeping of ink.

Balamurugan et al. [15] proposed steerable filters based on fuzzy unsharp masking. This technique introduces fuzzy based theory for unsharp masking. Image enhancement techniques for printed bangle documents are experimented by Islam et al. [16]. Masood et al. [17] detailed survey of features that have been used for content based image retrieval. Nagabhushana et al. [18] present a method for image reconstruction using wavelets. They propose a new multi-level wavelet decomposition for better reconstruction of the images.

Thus many techniques and methods for document image enhancement and restoration are reported in the literature. In this work, we propose a general frame work for preprocessing of the document for analysis. The frame work includes image conversion, enhancement, binarization and removal of clutter noise. As a case study we tested the framework for noisy and degraded document images of Tobacco 800 database [19].

III. PROPOSED FRAME WORK

The Algorithm 1 shows the steps used for document image noise removal and enhancement. The important operations required are color image to gray scale conversion, image enhancement, gray to binary conversion, clutter noise removal for further processing. These steps are briefly explained below.

Algorithm -1

```
Input: Noisy document image
Output: Noise free – enhanced document image

BEGIN

Step 1: if input document image is color image then
            Convert the color image to gray-scale image
        Endif

Step 2: Enhancement of gray-scale image

        Convolve the grayscale image with the mask
        shown in fig. 1 and calculate PSNR by using
        low-pass, median and wiener filters. Let PSNR1,
        PSNR2 and PSNR3 are PSNR of these filters
        respectively.

              If PSNR1>PSNR2 and PSNR1>PSNR3
                 Apply Low-pass Filter
              Elseif PSNR2>PSNR1 and PSNR2>PSNR3
                 Apply Median Filter
              Else
                 Apply Wiener Filter
              Endif

Step 3: Convert grayscale image into binary image.

Step 4: Removal of clutter noise

        Find the region of interest for removing clutter
        noise using horizontal and vertical projection of
        the document image. Then perform the
        following in region of interest.

        Step 4.1: Remove clutter in horizontal direction.

                  Count number of continuous black
                  pixels in horizontal direction.

                  If count>threshold
                         Replace black pixels with
                         background color
                  Endif

       Step 4.2: Remove clutter noise in vertical
       direction
```

```
                Count number of continuous black
                pixels in vertical direction.

                If count>threshold
                    Replace black pixels with
                    background color
                Endif
        END
```

3.1 Color image to grayscale conversion: The captured document images may be in color format or gray scale format. In this step, if the input image is a color image, it is converted to gray scale using weighted sum of Red, Green and Blue components. The equation (1) shows the method for conversion.

$$g(x,y) = 0.2989 \times R + 0.5878 \times G + 0.1140 \times B \quad (1)$$

Where R, G, B are Red, Blue and Green components of a color image.

3.2 Enhancement of the grayscale image: Various techniques for grayscale image enhancement are reported in the literature. In the proposed method, to enhance the converted grayscale document image g(x,y), an averaging(smoothing) operation is applied followed by unsharp filtering. To obtain unsharp filtered image, the mask shown in Fig. 1 is convolved with g(x,y). This results in enhanced details of the edges.

1	-1	-1
-1	8	-1
-1	-1	-1

Fig. 1 Mask for unsharp filter

Then to improve quality of the image, the convolved output is passed through a suitable filter. The Peak Signal to Noise Ratio (PSNR) is used as a parameter to decide the suitability of filter. PSNR is a quality measure of the image. Higher the PSNR, better is the quality. Hence, in this step, the filter that provides highest PSNR is selected. We used three filters namely: a low pass filter, median filter and weiner filter in the proposed method.

Low pass filter: The low pass filter restores every pixel of the image with mean of its neighborhood pixels [20].

Median filter: The median filter restores every pixel with median of its neighborhood pixels. It helps in reducing pepper and salt noise [20].

Wiener filter: The wiener filter aims to restore the image such that mean square error between the original and restored image should be less [20].

The PSNR for an image can be computed using equation (2).

$$PSNR = 10 \times Log_{10} \frac{R^2}{MSE} \quad (2)$$

Where "R" is the maximum gray level value of an image and MSE is Mean Square Error between original image and enhanced image. If I_1 and I_2 are the original and enhanced images of size M×N, the MSE between these images can be computed using equation (3).

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[I_{1(i,j)} - I_{2(i,j)} \right]^2 \quad (3)$$

3.3 Grayscale to Binary Conversion: Most of the document image analysis techniques require a binary image for processing purpose. Three thresholding methods namely global, local and adaptive are used for binarizing the image. Global thresholding depends only on gray levels of the image. Local thresholding depends on both gray levels and local properties of the image. However adaptive thresholding depends on gray levels, local properties as well as spatial coordinates of the pixels.

From the literature, it is learnt that Otsu[] method provided better results for grayscale to binary conversion. Therefore, we used an Otsu [21] method for converting enhanced gray scale image of the previous step to binary. The Otsu method, computes the global threshold level to be used for converting intensity image into binary. Threshold value lies in the range of 0 to 1, and this value can be used for binarizing the image.

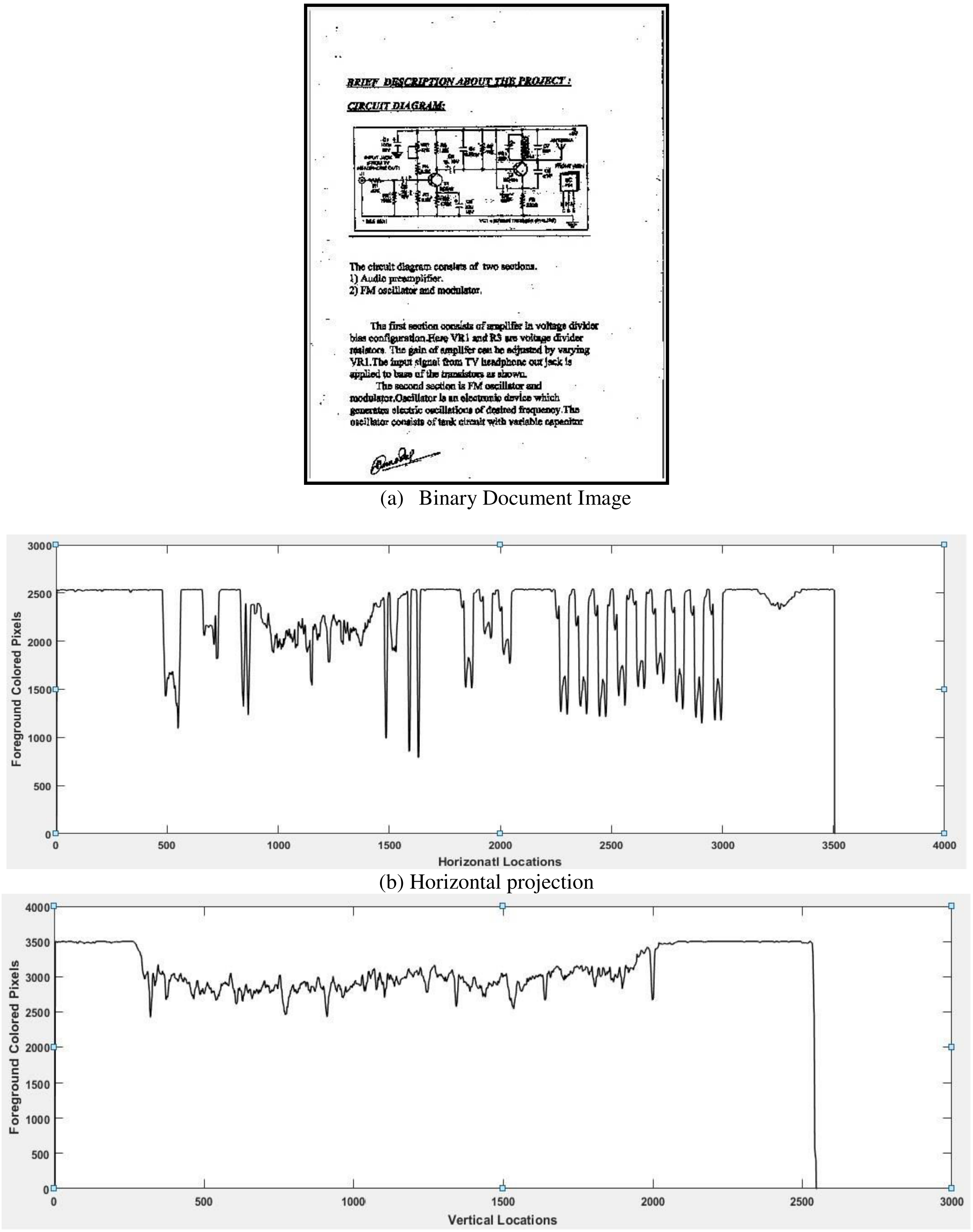
3.4 Removal of clutter noise: Clutter is a general term used to represent noisy or unwanted part of document image, whose size will be larger than the size of characters in document image. Clutter noise in document image is incurred during scanning process. It may be due to gap between gutter and scanner or distance between paper edges and bed of the scanner.

As the clutter noise mainly appears at margin area of the document, we first detect the area of the document occupied by the text. This helps in finding the region of interest for clutter removal algorithm to work. To detect region of interest, the horizontal and vertical projection of the document for row "I" and column "j" is computed using equation (4) and (5).

$$H - Projection(i) = \sum_{j=1}^{M} D(i,j) \quad (4)$$

$$V - Projection(j) = \sum_{i=1}^{N} D(i,j) \quad (5)$$

Where H and V represent horizontal and vertical projections of the document image D for i^{th} row and j^{th} column, whose size is M×N. Fig. 2 and Fig. 3 shows horizontal and vertical projection for a sample document image. Fig. 3 shows textual area and the region of interest, that is possible location where clutter noise can incur.



BRIEF DESCRIPTION ABOUT THE PROJECT :

CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifier in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

(a) Binary Document Image

(b) Horizontal projection

(b) Vertical projection

Fig. 2 Document Image, Horizontal and Vertical projections

Using horizontal projection and vertical projection, the text region of the document is detected and this area is excluded; assuming that it is free from clutter noise. The continuous sequence of foreground pixels in both horizontal and vertical direction represents clutter noise. Hence, we detected such sequence by counting the number of successive foreground pixels in both the directions. If this count is greater than the threshold value, they are replaced with background color of the document. Threshold value is chosen empirically by conducting the experiments. In this method we found that threshold value can be approximately 1.5 times the size of the largest character in the document. This helps in preserving loss of details from the printed characters. Equations (6) and (7) represent the idea of removing clutter noise.

$$If\ H_{count} > Threshold \begin{cases} g(x,y) = background\ color \\ g(x,y) = foreground\ color \end{cases} \tag{6}$$

$$If\ V_{count} > Threshold \begin{cases} g(x,y) = background\ color \\ g(x,y) = foreground\ color \end{cases} \tag{7}$$

Where H_{count} and V_{count} are count of successive foreground colored pixels representing clutter noise in the document image and g(x,y) is restored pixel at location (x,y).

IV. EXPERIMENTAL RESULTS

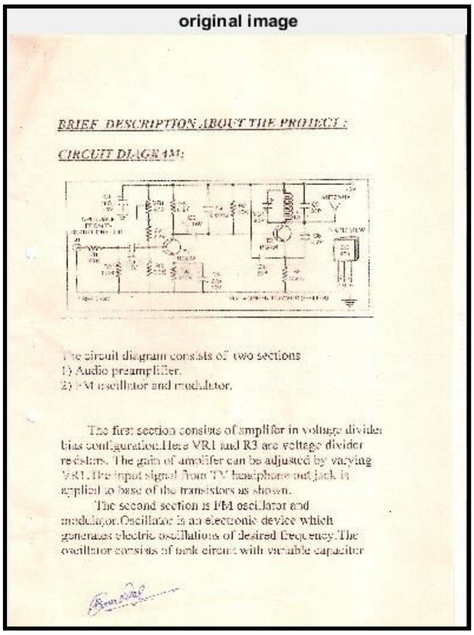
The proposed method uses document images of publicly available database Tobacco 800 [19] for testing puropose. This database is subset of IIT CDIP and has millions of document images, which is released by Tobacco companies. In our work, the document images affected by clutter noise are considered for testing. We also used some scanned document images from our own database to test the proposed work.

Results of the algorithm are shown step by step. Fig.3 and Fig. 4 shows input color document image and its grayscale converted image. Fig. 5 is the output of un-sharp filter. It can be observed that, the edges of the characters and other objects are enhanced. Fig. 6, Fig. 7 and Fig. 8 are the results of applying low pass, median and weiner filter to the unsharpened image. Table 1 shows PSNR values with the three filters.

Table 1. MSE & PSNR with different filters

Sl.No	Filter	MSE	PSNR
1	Lowpass	13.7318	36.7875
2	Median	2.2431	44.6563
3	Wiener	1.3907	46.7326

It is found that wiener filter provides lowest MSE of 1.3907 and highest PSNR of 46.7326. Hence the wiener filter is chosen by the proposed method for filtering. Fig. 9 shows the binarized version of the image obtained using Otsu method. Finally Fig. 10 shows the document image after removal of clutter noise. Fig. 11 shows some more results with only input and output document images skipping intermediate steps.


original image

BRIEF DESCRIPTION ABOUT THE PROJECT :

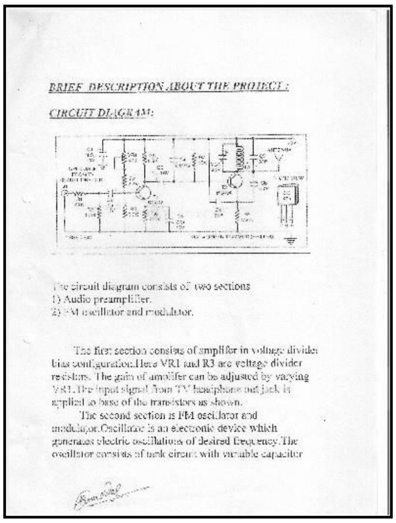
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 3 Input color document image


BRIEF DESCRIPTION ABOUT THE PROJECT :

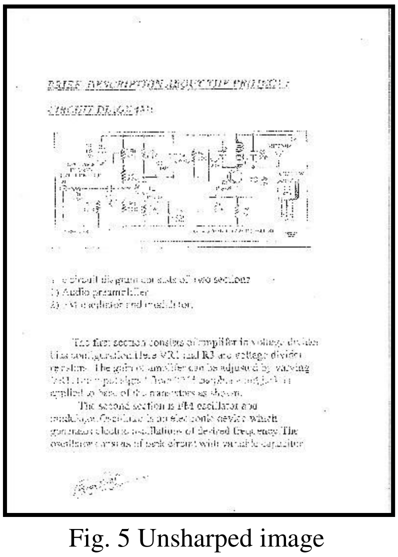
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 4 Gray scale image


BRIEF DESCRIPTION ABOUT THE PROJECT :

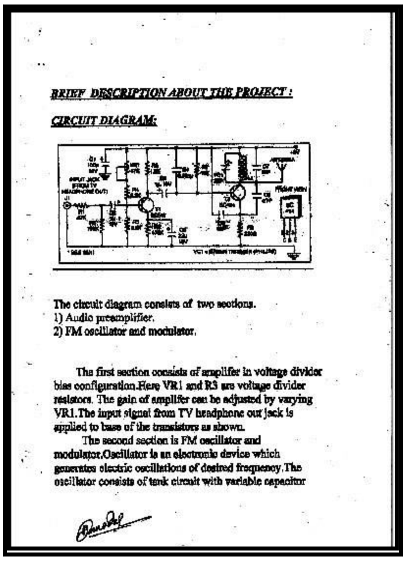
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 5 Unsharped image



BRIEF DESCRIPTION ABOUT THE PROJECT :

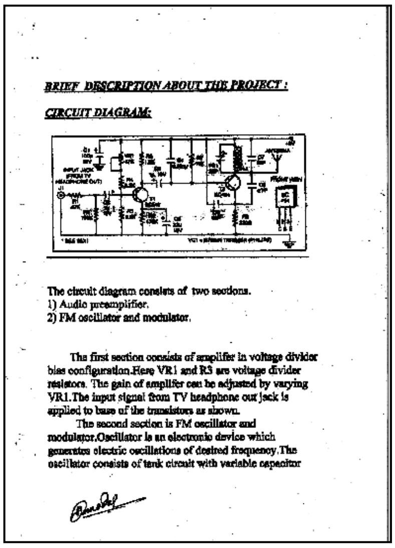
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 6 Low-pass filtered image



BRIEF DESCRIPTION ABOUT THE PROJECT :

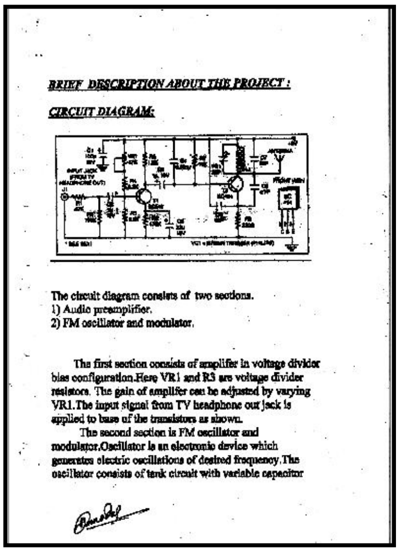
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 7 Median filtered image



BRIEF DESCRIPTION ABOUT THE PROJECT :

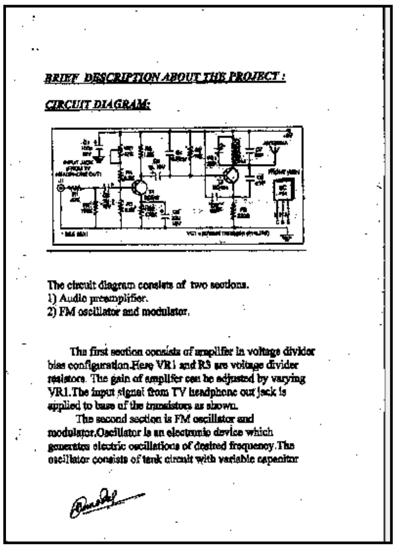
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 8 Wiener filtered image



BRIEF DESCRIPTION ABOUT THE PROJECT :

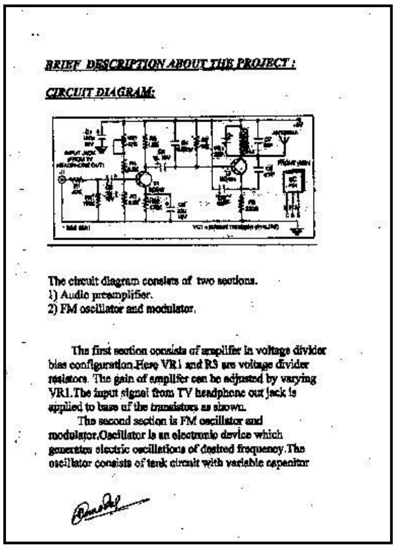
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 9 Binarizaed image



BRIEF DESCRIPTION ABOUT THE PROJECT :

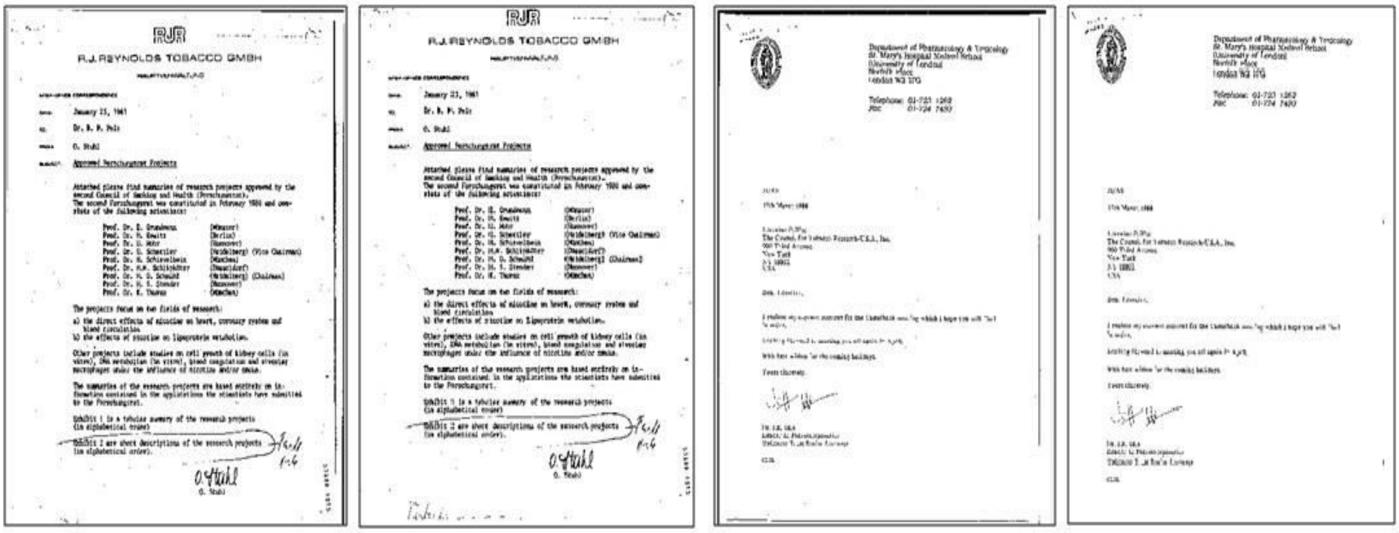
CIRCUIT DIAGRAM:

The circuit diagram consists of two sections.
1) Audio preamplifier.
2) FM oscillator and modulator.

The first section consists of amplifer in voltage divider bias configuration.Here VR1 and R3 are voltage divider resistors. The gain of amplifer can be adjusted by varying VR1.The input signal from TV headphone out jack is applied to base of the transistors as shown.

The second section is FM oscillator and modulator.Oscillator is an electronic device which generates electric oscillations of desired frequency.The oscillator consists of tank circuit with variable capacitor

Fig. 10 Image after noise removal



(a) Input Document Image1 (b) Output Document Image1 (c) Input Document Image2 (d) Output Document Image2

Fig. 11 Sample results for Tobacco 800 document images

From the results it can be observed that, the document is free from clutter noise and also its quality is enhanced. An enhanced document with no noise will help to achieve better results during document image analysis. Thus proposed frame work provides a simple approach for preprocessing of the document images.

V. CONCLUSION

This paper presents a preprocessing framework for document image analysis. The major steps such as color to gray scale conversion, image enhancement, binarization and an idea for removal of clutter noise are included. The frame work is tested on document images of publicly available database Tobacco 800. The results of the preprocessing are promising and encouraging.

REFERENCE

[1] Umesh D. Dixit and M.S.Shirdhonkar, “A Survey on Document Image Analysis and Retrieval System”, *International Journal on Cybernetics & Informatics,* 4(2), 2015, 259-270.

[2] Atena Farahmand, Abdolhossein Sarrafzadeh, Jamshid Shanbehzadeh, “Document Image Noises and Removal Methods”, *Proc. of International Conf. of Engineers and Computer Scientists 2013*, 1-5.

[3] Yue Wang, Jobin J Mathew, Eli Saber, David Larson, Peter Bauer, George Kerby, Jerry Wagner, *“*Scanned Document Enhancement based on Fast Text detection”, *Proc. of IEEE ICASSP 2016*, 1961-1965.

[4] Nouman Khna and Shalini Puri, “A study on text detection techniques of printed documents”, *Proc. of IEEE WiSPNET 2016*, 2478-2482.

[5] Xujun Peng, Huaigu Cao, Prem Natarajan, “Document image quality assessment using discriminative sparse representation”, *Proc. of 12th IAPR Workshop on Document Analysis Systems 2016*, 227-232.

[6] Vijay Kumar, Amit Bansal, Goutam Hari Tulsiyan, Anand Mishra, Anoop Namboodiri and C. V. Jawahar, “Sparse document image coding for restoration”, *Proc. of 12th IEEE Conf. on Document Analysis and Recognition 2013,*713-717.

[7] Thibault Lelore and Frederic Bouchara, “FAIR: A Fast Algorithm for Document Image Restoration”, *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 35(8), 2013, 2039-2048.

[8] J. Banerjee, A. M. Namboodiri, C. V. Jawahar, "Contextual Restoration of Severely Degraded Document Images," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 517-524.

[9] E. M. Sgarbi, W. A. Della Mura, N. Moya, J. Facon and H. A. L. Ayala, "Restoration of Old Document Images Using Different Color Spaces", *Proc.of International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014, 82-88.

[10] H. Deborah and A. M. Arymurthy, "Image Enhancement and Image Restoration for Old Document Image Using Genetic Algorithm", *Proc. of Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 2010, 108-112.

[11] M. R. Yagoubi, A. Serir, A. Beghdadi, "Blind Document Image Enhancement Based on Diffusion Process", *Proc. of 5th European Workshop on Visual Information Processing (EUVIP)*, 2014, 1-6.

[12] A. Rajwade, A. Rangarajan and A. Banerjee, "Image Denoising Using the Higher Order Singular Value Decomposition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4), 2013, 849-862.

[13] Hossein Nezamabadi-pour and Saeid Saryazdi, “An Efficient Method for Document Image Enhancement”, *Proc. of International Symposium on Telecommunications*, 2005, 175-180.

[14] Chew Lim Tan, R. Cao, Peiyi Shen, "Restoration of Archival Documents Using a Wavelet Technique",

IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(10), 2002, 1399-1404.

[15] E. Balamurugan, P. Sengottuvelan, K. Sangeetha, "Document image restoration using steerable filters based fuzzy unsharp masking", *International Journal of Soft Computing*, 9(2), 2014, 88-94.

[16] Anika Binte Islam, Fahim Salam Chowdhury, Fariha Nusrat, Kazi Lutful Kabir, Hasan Sarwar, "A Study on Image Enhancement Method for Printed Bangla Document Images", *Proc. of 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016, 725-730.

[17] Anum Masood, Muhammad Alyas Shahid, Muhammad Sharif, "Content-Based Image Retrieval Features: A Survey", *Int. J. Advanced Networking and Applications,* 10(1), 2018, 3741-3757.

[18] Nagabhushana, Aravinda T V, A V Radhika, "Image Reconstruction Using Wavelet Method", *Int. J. Advanced Networking and Applications*, 10(2), 2018, 3804-3807.

[19] G.Zhu and D. Doermann, Tobacco-800 Complex Document Image Database and Ground truth. online, 2008. http://lampsrv01.umiacs.umd.edu/projdb/edit/project.php?id=52.

[20] Lim, Jae S., *Two-Dimensional Signal and Image Processing*, Englewood Cliffs, NJ, Prentice Hall, 1990.

[21] Otsu N., "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man, and Cybernetics,* 9(1), 1979, 62-66.

AUTHORS' PROFILE

Mr. Umesh D. Dixit has received B.E (E&C) and M.Tech (CSE) from Visvesvaraya Technological University, Belagavi. He is working as Asst. Professor in the dept. of Electronics & Communication Engineering, B.L.D.E.A's V. P. Dr. P. G. Halakatti CET, Vijayapur (Karnataka), India since 14 years. His areas of interest are Image processing, Dcoument image analysis & retrieval and embedded systems. Currently he is pursuing Ph.D under Visvesvaraya Technological University. Belagavi.

Dr. M. S. Shirdhonkar is a Professor in the Department of Computer Science and Engineering, B.L.D.E.A's V.P.Dr.P.G.Halakatti CET, Vijayapur (Karnataka) India. He has received B.E (CSE), M.E (CSE) from Shivaji University, Kolhapur, Maharashtra and his Doctorate from Swami Ramanand Teerth, Marathwada University, Nanded, Maharashtra. He has 21 years of experience. His area of interest is Image processing, Document image retrieval and analysis, Pattern recognition and Speech recognition. He has published more than 26 papers in National, International conferences and journals. He is life member of professional bodies like IEI, CSI and ISTE.