Text Extraction from Document Images Using Gabor, Wavelet and Hough Technique: A Novel Approach

Vikas K. Yeotikar, Manish T. Wanjari, Dr. Mahendra P. Dhore

***Abstract*— Text extraction in document images has been an important research area. Extraction of the information in the form of text involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given document image. A large number of techniques have been proposed to address this problem. In this paper a novel method is proposed by using three features extraction techniques i.e. Gabor, Wavelet and Hough to detect text objects from document images. The performance of the proposed method is tested on NIST document Image dataset.**

***Ketwords*— Document Image Analysis (DIA), Text Extraction, Text Detection, Text Localization, Text Enhancement, Gabor, Wavelet, Hough and Canny.**

I. Introduction

It has been an ancient dream to achieve human functions like reading, recognizing and thinking through machines. Much of our interaction with environment requires recognition of `things' such as sounds, smells, shapes in a scene (text characters, faces, flowers, plants), etc. Analysis of document images for information extraction has gained immense importance in recent past. Wide variety of information, which has been conventionally stored on paper, is now being converted into electronic form for better storage and intelligent processing. This needs processing of documents using image analysis algorithms. Locating text image blocks and tables, and defining appropriate algorithm is the major challenge in document image analysis [1, 2].

Purpose of texture analysis:

- To identify different textured and non-textured regions in a document image.
- To classify/segment different text regions in an image.
- To extract boundaries between major texture regions.
- To describe the Text unit. [3]

A number of approaches have been proposed by researchers. In this paper we have implemented the same using Gabor Filter, wavelet and Hough Transform.

The proposed model can capture the characteristics of characters and the structure of text objects simultaneously; three new features are used to describe the inherent properties of characters. Our method is robust to the font, size, color, and orientation of text and can discriminate text objects from others effectively.

However, our method works only when text is composed of two or more isolated characters which are placed in an orderly manner. It is designed to detect single character and word since there are typically many such candidates that can only be accurately labeled as text after an optical character recognition (OCR). It is also not appropriate for languages in which many characters in a word are connected together or a single composite character is formed by disjoint character strokes.

In Section 2, we present review of the related work. In Section 3, we present our proposed text detection method. The performance of our method is evaluated and discussed in Section 4 using NIST document image datasets. Finally, we conclude the paper in Section 5.

II. Related Work

A robust approach to segment text from color images was put forth by Y. Zhan et.al [4]. The proposed algorithm uses the multiscale wavelet features and the structural information to locate candidate text lines. Then a SVM classifier was used to identify true text from the candidate text lines. This approach mainly included four stages. In preprocessing step text blocks were enhanced by using cubic interpolation to rescale the input text blocks and a Gaussian filter to smooth the text blocks and remove noises. These image blocks were split into connected components and non-text connected components were eliminated by a component filtering procedure. The left connected components were merged using K-means clustering algorithm into several text layers, and a set of appropriate constraints were applied to find the real text layer. Finally, the text layer was refined through a post-processing step.

Thai et.al [5] described an approach for effective text extraction from graphical document images. The algorithm used Morphological Component Analysis (MCA) algorithm, an advancement of sparse representation

Manuscript received June 12, 2015.

Vikas K. Yeotikar, Department of Computer Science, SSESA's, Science College, Congress Nagar, Nagpur (MH), India, 9405436996.,

Manish T. Wanjari, Department of Computer Science, SSESA's, Science College, Congress Nagar, Nagpur (MH), India, 9028413175.,

Dr. Mahendra P. Dhore, Department of Electronics & Computer Science, RTM Nagpur University Campus, Nagpur (MH), India, 9423103043,

framework with two appropriately chosen discriminative over complete dictionaries. Two discriminative dictionaries were based on undecimated wavelet transform and curvelet transform. This method overcame the problem of touching between text and graphics and also insensitive to different font styles, sizes, and orientations.

S. Audithan et.al [6] formulated an efficient and computationally fast method to extract text regions from documents. They proposed Haar discrete wavelet transform to detect edges of candidate text regions. Non-text edges were removed using thresholding technique. They used morphological dilation operator to connect the isolated candidate text edge and then a line feature vector graph was generated based on the edge map. This method exploited an improved canny edge detector to detect text pixels. The stroke information was extracted the spatial distribution of edge pixels. Finally text regions were generated and filtered according to line features.

Grover et.al [7] described an approach to detect text from documents in which text was embedded in complex colored document images. They proposed a simple edge based feature to perform this task. The image was converted to gray scale by forming a weighted sum of the R, G, and B components. Then edge detection was performed on the gray-scale image by convolving the image with Sobel masks, separately for horizontal and vertical edges. Convolution was followed by elimination of non-maxima and thresholding of weak edges.

Next, the edge image was divided into small non overlapping blocks of m x m pixels, where m depends on the image resolution. They performed block classification using pre-defined threshold which would differentiate the text from the image.

P. Nagabhushan et.al [8] proposed a novel approach to extract the text in complex background color document images. The proposed method used canny edge detector to detect edges. When dilation operation was performed on edge image, it created holes in most of the connected components that corresponds to character strings. Connected components without hole(s) were eliminated. Other non-text components were eliminated by computing and analyzing the standard deviation of each connected component. An unsupervised local thresholding was devised to perform fore-ground segmentation in detected text regions. Finally the noisy text regions were identified and reprocessed to further enhance the quality of retrieved foreground.

A robust and efficient algorithm for automatic text extraction from colored book and journal cover sheets was proposed by Davod et.al[9] based on wavelet transform. A dynamic threshold was used to detect edges from detail wavelet coefficient. Further effective edges were obtained by blurring approximate coefficients with alternative heuristic thresholding. Region of Interest (ROI) technique was applied and finally text was extracted. They evaluated the performance of their algorithm on 80 pictures collected from internet.

Another algorithm for Automatic text location and identification on colored book and journal covers was proposed by Karin et.al[10]. The number of colors was reduced by applying a clustering algorithm. Text candidates were located using a top-down analysis based on successive splitting in horizontal and vertical direction. A bottom-up analysis detected homogeneous regions using a region growing method; grouping step was applied to find subsets of region. Finally text regions and non-text regions were distinguished.

Zhixin Shi et.al[11] proposed an extraction algorithm based on connectivity features for a complex handwritten historical document. This paper presented an algorithm using adaptive local connectivity map (ALCM) technique. Thresholding the gray scale image discloses clear text-line patterns as connected components. Grouping algorithm was used to group the connected components into location masks for each text line. Text line was extracted by mapping the location masks back onto the binary image to collect the text line components.

Splitting algorithm overcame the problem of components touching multiple lines. This method dealt with fluctuating or skewed text lines and used for other types of images such as binary images, machine printed or even mixed script.

Syed Saqibet.al[12] described an approach for curled textline information extraction from grayscale camera-captured document images. The grayscale textline was enhanced by using multi-oriented multi-scale anisotropic Gaussian smoothing. Detection of central lines of curled textlines was found using ridges. This approach was based on differential geometry, which used local direction of gradients and second derivatives as the measure of curvature. Hessian matrix was used for finding direction of gradients and derivatives. By using this information, ridges were detected by finding the zero-crossing of the appropriate directional derivatives of smoothed image. Modified coupled snakes model was used for estimating x-line and baseline pairs from detected textlines. Their approach was robust against high degrees of curl and requires no post-processing.

Wafa et.al [13] suggested a new enhanced text extraction algorithm from degraded document images of both color and grayscale type on the basis of the probabilistic models. Color document image was converted to YIQ colors space image and operate on Y luminance channel. Initial estimates and their corresponding mean and standard deviation vectors for expectation maximization (EM) algorithm were calculated using k-means clustering method. The EM algorithm was used to estimate and improve the parameters of the mixtures of densities recursively. The maximum likelihood (ML) segmentation method estimates the probability that a pixel belongs to text or background.

III. Methodology

Document images are acquired by scanning journal, printed document, degraded document images, handwritten historical document, and book cover etc. The text may appear in a virtually unlimited number of fonts, style, alignment, size, shapes, colors, etc. Extraction of text from text document images and from complex color background is difficult due to complexity of the background and mix up of colors of fore-ground text with colors of background. In this section, we present the main ideas and details of the proposed algorithm.

Implementation of any system needs the study of features, it may be symbolic, numerical or both. An example of a symbolic feature is color; an example of numerical feature is weight. Features may also result from applying a text extraction algorithm or operator to the input data. The related problems of feature selection and feature extraction must be addressed at the outset of any text recognition system design. The key is to choose and to extract features that are computationally feasible and reduce the problem data into a manageable amount of information without discarding valuable information.

Different methods used for text extraction from document images (as shown in fig. 1) include:

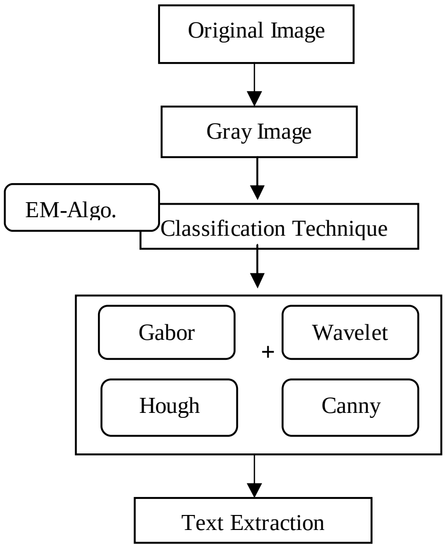


Fig. 1. Schematic Overview of proposed Text Extraction from Document Images

A. Feature Extraction

Feature extraction involves the extracting the meaningful information from the document image. The features are classified in to Global features and Local features. Features that are extracted from whole image are known as the global features and the features that are extracted from blocks identified during segmentation or from subdivision of the document are known s local features. They can be divided into several categories: textural, geometric, component, structural and content based. The extractions of global and local features provide input to classification algorithm/techniques.

B. Color Extraction (RGB)

There are three types of images which include Binary image, gray scale and RGB(color) image. Binary image consists of Logical array containing only 0s and 1s, interpreted as black and white, respectively. Grayscale image is also known as an intensity, gray scale, or gray level image. Array of class uint8, uint16, int16, single, or double whose pixel values specify intensity values. For single or double arrays, values range from [0, 1]. For uint8, values range from [0,255]. For uint16, values range from [0, 65535]. For int16, values range from [-32768, 32767]. True color image is also known as an RGB image. A true color image is an image in which each pixel is specified by three values one each for the red, blue, and green components.

C. Classification EM-Algorithm

The EM algorithm is an iterative algorithm for calculating the maximum-likelihood or maximum-a-posterior estimates when the observations can be viewed as incomplete data. Each iteration of the algorithm consists of an expectation step followed by a maximization step.

We now define the EM algorithm, starting with cases that have strong restrictions on the complete-data specification $f(x \mid \Phi)$, then presenting more general definitions applicable when these restrictions are partially removed in two stages. The simplicity of description and computational procedure, and thus the appeal and usefulness, of the EM algorithm are greater at the more restricted levels. [19]

Suppose first that $f(x \mid \Phi)$ has the regular exponential-family form:

$f(x \mid \Phi) = b(x) \exp (\Phi t(x)^T) / a(\Phi)$,

D. Gabor Filter Method

The Gabor Transform is also referred to as the Short Time Fourier Transform (STFT). Filtering the time-frequency content of a signal is indeed one of the main applications of Gabor multipliers. Gabor analysis is an essential part of time-frequency analysis, initiated by the seminar paper of Denis Gabor in 1946.

Some properties of Gabor filters:

- A tunable bandpass filter
- Similar to a STFT or windowed Fourier transform
- Satisfies the lower-most bound of the time- spectrum resolution (uncertainty principle)
- It’s a multi-scale, multi-resolution filter
- Has selectivity for orientation, spectral bandwidth and spatial extent.
- Has response similar to that of the Human visual cortex (first few layers of brain cells)
- Computational cost often high, due to the necessity of using a large bank of filters in most applications [14]

E. Wavelet Transform Method

Wavelets are functions that satisfy certain mathematical requirements and are used in presenting data or other functions, similar to sines and cosines in the Fourier transform. However, it represents data at different scales or resolutions, which distinguishes it from the Fourier transform. Wavelet transform is an increasingly popular tool in computer vision and image processing. Many applications, such as compression, detection, recognition, image retrieval have been investigated. Wavelet transform has nice features of space-frequency localization and multi-resolutions. The wavelet transform of a 1-D signal *f(x)* is defined as:

$$(W_a f)(b) = \int f(x)\, \Psi a,b\,(x) dx$$

With

$$\Psi a,b\,(x) = \frac{1}{\sqrt{a}}\, \Psi \left(\frac{x-b}{a}\right)$$

The mother wavelet Ψ has to satisfy the admissibility criterion to ensure that it is a localized zero-mean function. [15]

F. Hough Transform Method

Hough Transform technique is used to extract text from document images. Hough Transform (HT) is recognized as a powerful tool for graphic element extraction from images due to its global vision and robustness in noisy or degraded environment. The method herein proposed detects text lines on document images which may include either lines oriented in several directions, erasures, or annotations between main lines. At each stage of the process, the best text-line hypothesis is generated in the Hough Transform domain.

The Hough transform is a feature extraction technique used in image analysis, computer vision, and digital image processing.[16]

G. Edge Detection Method (Canny)

The edge representation of a document image significantly reduces the quantity of data to be processed; it retains necessary information regarding the shape of character in document image. There are many edge detection methods in the literature for document images. Most of the used discontinuity based edge detection methods are reviewed. Those methods are Prewitt, Sobel, Canny, Roberts, Zerocross and Laplacian of Gaussian. [17] The Canny edge detector is regarded as one of the best and standard edge detectors recently in use; Canny's edge detector ensures good noise immunity and at the same time detects true edge points with minimum error. The Canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in document images. It is developed by John Canny considered the mathematical problem of deriving an optimal smoothing filter given the criteria of detection, localization and minimizing multiple responses to a single edge.

Algorithm of Proposed Method:

```
Input : Image Recognition f(x,y)        $1 \le x \le M$ ,  $i \le y \le N$
Output : Recognize Characters Array $C_R(i)$       $0 \le i \le NC$
Method:
Step 1: Read an Input Image f(x,y)
   img = imread (FileName)       // image file name
Step 2: if isrgb (img)
   img1 = rgb2gray (img)       // If img is RGB then convert to Gray
   else
   img1 = img ;
   end
Step 3 : if isrgb (img)            // Extract color features
   R = colorfeature (img);
   G = colorfeature (img);
   B = colorfeature (img);
          else
          end
Step 4 : Texture = TextureFeature (img1);// Texture Feature Extraction
Step 5 : NOC = Number of classes;
   var (i) = variance of class i
   p (i) = probability of class i
   IC = ExpMax (img1, NOC, var(i), p(i) );
Step 6 : Create Gabor filter bank           // Gabor & Wavelet Algorithm
   Calculate Gabor-Wavelet Features.
Step 7 : Apply Hough Transform for Line Detection.
Step 8 : Combine Image Classification, Gabor-Wavelet features, Hough Transform and Canny to
          Extract Characters.
Step 9 : Extract Characters from the image and display each one by one.
Step 10 : Calculate Tp, fp, fn, Precision, Recall and F-measure and display results.
Step 11 : STOP
```

IV. Implementation and Results

The goal of our proposed method is text extraction that achieves the highest recognition accuracy and fast performance. In this section, we demonstrate the efficiency of the proposed method on 20 Document Images of NIST database. For that we used following Mathematical and Statistical methods:

a. F_N (False Negative)

False Negatives (FN) / Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

b. F_P (False Positive)

False Positives (FP) / False ala**r**ms are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as text.

c. T_P (Total Positive)

Total Positive (T_P) is the correctly detected characters.

d. Precision

Precision rate (P) is defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives.

Eq(1)… $$P = \frac{TP}{TP + FP}$$

e. Recall

Recall rate (R) is defined as the ratio of the correctly detected characters to sum of correctly detected characters plus false negatives.

Eq(2)… $$P = \frac{TP}{TP+FN}$$

f. F-Measure

F-Measure is the harmonic mean of recall and precision rates.

Eq(3)… $$F = \frac{2*PR}{P+R}$$

Nist Database

The National Institute of Standards and Technology, Technology Administration, U. S. database of document images has been used in order to carry out experimental work. The image document database used is NIST DATABASE which consist of 4711 document images, Federal Register Document Image Database. NIST Special Database 25 – volume 1.(NISTIR 6245).[18]

Sample Results on NIST Database

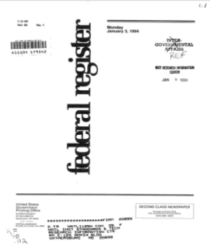


Fig2a: Original Image

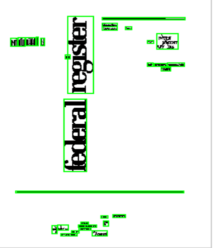


Fig2b: Detected Text of 2a Image

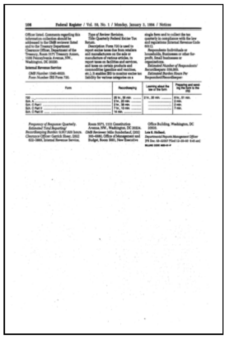


Fig3a: Original Image

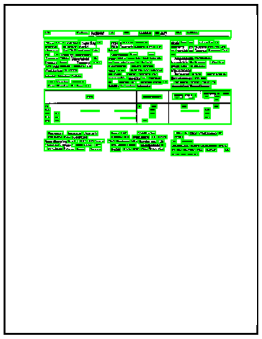


Fig3b: Detected Text of 3a Image

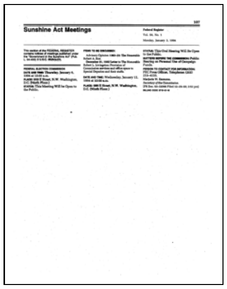


Fig4a: Original Image

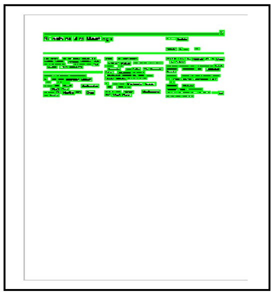


Fig4b: Detected Text of 4a Image

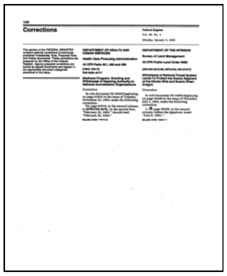


Fig5a: Original Image

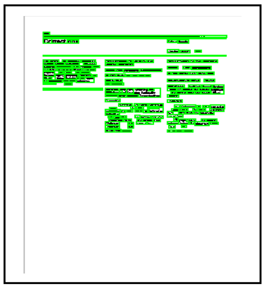
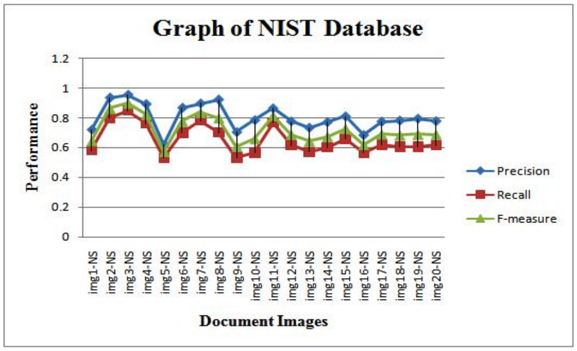


Fig5b: Detected Text of 5a Image

Table 1: Results Calculated on NIST Database

F-Name	Precision	Recall	F-measure
img1-NS	0.720	0.581	0.643
img2-NS	0.935	0.797	0.860
img3-NS	0.954	0.845	0.896
img4-NS	0.892	0.761	0.821
img5-NS	0.616	0.529	0.570
img6-NS	0.867	0.698	0.773
img7-NS	0.896	0.779	0.833
img8-NS	0.922	0.701	0.797
img9-NS	0.703	0.531	0.605
img10-NS	0.784	0.564	0.656
img11-NS	0.863	0.767	0.812
img12-NS	0.775	0.614	0.685
img13-NS	0.733	0.569	0.640
img14-NS	0.770	0.599	0.673
img15-NS	0.809	0.657	0.725
img16-NS	0.683	0.563	0.617
img17-NS	0.774	0.618	0.688
img18-NS	0.781	0.604	0.681
img19-NS	0.794	0.605	0.687
img20-NS	0.777	0.614	0.686



Graph 1. Performance on NIST Database

Table2: Performance Analysis with different Methods

S.N	Author	Year	Accuracy
1	Davod et.al [9]	2011	81.20%
2	Thai et.al[5]	2010	94.76%
3	Nagabhushan et.al [8]	2010	87.12%
4	Grover et.al [7]	2009	92%
5	Syed et.al [12]	2009	91%
6	Wafa et. al. [13]	2009	76%
7	S.Audithan et.al [6]	2009	84.80%
8	Zhan et al. [4]	2006	84.3%.
9	Zhixin et.al [11]	2005	85%
10	Karin et.al [10]	1999	Promising Results
11	OUR METHOD	2015	88.46%

The performance evaluation of our method on NIST dataset is listed in Table 2 along with the best algorithms reported and the algorithms proposed. We can see that the proposed text detection method achieved better performance than most of the listed algorithms/techniques.

V. CONCLUSION

In this paper, we have presented a novel approach for detection and recognition of text from Document Images. Text detection and recognition are accomplished concurrently with exactly the same features and classification scheme. A text-image-analysis is needed to enable a text information extraction system to be used for any type of document images. The proposed method is evaluated on the dataset (NIST). It is shown that the proposed method outperforms all the compared state-of-the-art and baseline algorithms, which illustrates the robustness of the proposed method.

REFERENCES

[1] Shuichi Tsujimoto And Haruo Asada. Invited Paper .Major Components of a Complete Text Reading System. Proceedings of the IEEE, Vol. 80, No. 7, pp.1133-1149, July 1992.

[2] Gaurav Harit,Santanu Chaudhari, Gupta P., Vohra N., Joshi S. D. .A Model Guided Document Image Analysis Scheme. proceedings of IEEE pp. 1137-1141, 2001.

[3] Haralick, R.M. 1979. Statistical and Structural Approaches to Texture. Proceedings of the IEEE, 67:786-804; (also 1973, IEEE-T-SMC.

[4] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-And-Merge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[5] Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation"in *Proc. Das*, pp 143–150. International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.4, August 2012.

[6] S. Audithan,, R.M.Chandrasekaran (2009), "Document Text Extraction From Document Images Using Haar Discrete Wavelet Transform",*European Journal Of Scientific Research*, Vol.36 No.4 , pp.502-512.

[7] Sachin, Grover, Kushal Arora,,Suman K. Mitra(2009),"Text Extraction From Document Images Using Edge Information",*IEEE India Council Conference*.

[8] P. Nagabhushan, S. Nirmala(2009) ,"Text Extraction In Complex Color Document Images For Enhanced Readability",*Intelligent Information Management*, pp: 120-133.

[9] Davod Zaravi, Habib Rostami, Alireza Malahzaheh, S.S Mortazavi(2011)," Journals Subheadlines Text Extraction Using Wavelet Thresholding And New Projection Profile", *World Academy Of Science, Engineering And Technology* .Issue 73.

[10] Karin Sobottka, Horst Bunke and Heino Kronenberg(2009), "Identification Of Text On Colored Book And Journal Covers", *ICDAR*.

[11] Zhixin Shi, Srirangaraj Setlur And Venu Govindaraju(2005), "Text Extraction From Gray Scale Historical Document Image Using Adaptive Local Connectivity Map", *Proceeding Of The Eighth International Conference On Document Analysis And Recognition*, Vol. 2, pp: 794–798.

[12] Syed Saqib Bukhari , Thomas M. Breuel,Faisal Shafait(2009), "Textline Information Extraction From Grayscale Camera-Captured Document Images ", *ICIP Proceedings Of The 16th IEEE International Conference On Image Processing*, pp: 2013 – 2016.

[13] Wafa , Aymen Bougacha, Abderrazak Zahour, Haikal El Abed, Adel Alimi(2009) ,"Enhanced Text Extraction From Arabic Degraded Document Images Using Em Algorithm", *10th International Conference On Document Analysis And Recognition*.

[14] Simona E. Grigorescu, Nicolai Petkov, and Peter Kruizinga; Comparison of Texture Features Based on Gabor Filters; IEEE Transactions on Image Processing, Vol. 11, No. 10, OCT. 2002; pp 1160-1167.

[15] E. J. Stollnitz, T. D. DeRose and D. H. Salesin" Wavelets for computer graphics: a primer, part I,"IEEE Computer Graphics and Applications, vol.15, No. 3, pp. 76-84, May 1995.

[16] Duda R. O. and P. E. Hart, "Use of the Hough Transformation to detect Lines and Curves in Pictures," Comm. ACM, Vol. 15, pp. 11-15 , Jan-1972.

[17] Muthukrishnan.R and M.Radha (Dec. 2011). Edge Detection Techniques For Image Segmentation. *International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6.*

[18] Manual of NIST DATABASE, Federal Register Document Image Database. NIST Special Database 25 – volume 1.(NISTIR 6245).

[19] A. P. DEMPSTERN, M. LAIRDa nd D. B. RubIN, "Maximum Likelihood from Incomplete Data via the EM Algorithm" *Journal of the Royal Statistical Society*. *Series B (Methodological)*, Vol. 39, No. 1. (1977), pp. 1-38.

Vikas K. Yeotikar is a Research Scholar Pursuing Ph. D. in Computer Science. He is currently working as Lecturer in Department of Computer Science, SSESA's, Science College, Congress Nagar Nagpur. Email Id-vkyeotikar@gmail.com

Manish T. Wanjari is a Research Scholar pursuing Ph. D. in Computer Science. He is currently working as Project Fellow under UGC Sponsored Major Research Project in the Department of Computer Science, SSESA's, Science College, Congress Nagar Nagpur. Email Id-mwanjari9@gmail.com

Dr. Mahendra P. Dhore is Associate Professor in Computer Science, Department of Electronics & Computer Science, RTM Nagpur University, Nagpur. He is having teaching experience of more than 18 years at UG & PG level. His research areas include Digital Image Processing, Document Image Analysis, Mobile Computing & Cloud Computing. He is Member of IEEE, IAENG, IACSIT, IETE, and ISCA.
Email Id–mpdhore@ieee.org, mpdhore@rediffmail.com