INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)
www.vmsindia.org

Comparison of segmentation techniques in document image analysis

Manish T. Wanjari[1], Jageshvar K. Keche[2], Mahendra P. Dhore[3]
[1]Department of Computer Science, SSESA, Science College, Congress Nagar, Nagpur, (MS)-India, mwanjari9@gmail.com
[2]Department of Computer Science, SSESA, Science College, Congress Nagar, Nagpur, (MS)-India, jkkeche@gmail.com
[3]Department of Electronics & Computer Science, RTM Nagpur University, Campus Nagpur, (MS)-India, mpdhore@rediffmail.com

Abstract –
Document Image Segmentation (DIS) is the process of extraction and recognition of information from document images. It is the process of partitioning a document image into multiple segments like set of pixels. DIS is an important aspect of many document images and computer vision applications. Over last few decades, many segmentation techniques have been proposed. This paper deals with the performance of various segmentation techniques such as Edge detection, Region based, K-means clustering, Texture based and Gabor filter technique. In this paper we have implemented the existing document image segmentation techniques and a detailed comparative analysis is carried out with reference to the performance of the segmentation techniques.

Keyword- Document Image Analysis (DIA), Document Image Segmentation (DIS), Edge Detection, Region Growing, Clustering, Otsu, Boundary Detection.

I. INTRODUCTION

Segmentation subdivides a document image into its constituent regions or objects. The level to which the subdivision is carried out depends on the problem being solved. Segmentation should stop when an object of interest in an application have been isolated. DIS simplifies and changes the representation of a document image. i.e., the document image is transferred into something that is more meaningful and easier to analyze. It is typically used to locate objects of interest and boundaries such as lines, curves in a document image [1].

DIS algorithms generally are based on one of two basic properties of intensity values such as discontinuity and similarity. In the first category, the approach is to partition a document image based on abrupt changes in intensity, such as edges in a document images. The principal approaches in the second category are based on partitioning a document image into regions that are similar according to a set of predefine criteria. We discuss a number of approaches in the first category mentioned. We begin the development with methods suitable for detecting gray level discontinuities such as points, lines and edges. Particularly Edge detection has been a staple for many years. In addition to edge detection, we also discuss methods for connecting edge segments and for 'assembling' edges into region boundaries [2].

Edge detection techniques are generally used for finding discontinuities in gray level images. Edge detection is the most common approach for detecting meaningful discontinuities in the gray level. Document Image segmentation methods for detecting discontinuities are boundary based methods. Edges are local changes in the document image intensity. Edges typically occur on the boundary between two regions. Important features can be extracted from the edges of a document image (e.g., corners, lines, curves). Edge detection is an important feature for DIA. These features are used by higher-level computer vision algorithms (e.g., recognition). Edge detection is used for object detection which serves various applications like document image processing, biometrics etc. Edge detection is an active area of research as it facilitates higher level document image analysis. There are three different types of discontinuities in the grey level like point, line and edges. Spatial masks can be used to detect all the three types of discontinuities in a document images. And also we discuss various operators uses the first and second order derivatives of edge detection techniques. [3]

The objective of segmentation is to partition a document image into regions. We approached the problem by finding boundaries between region based on discontinuities in gray levels, whereas segmentation was accomplished via thresholds based on the distribution of pixel properties, such as gray-level values or color. Document Image segmentation is useful in many applications. It can identify the regions of interest in a scene or annotate the data. We categorize the existing segmentation method into region-based segmentation [4].

The goal of clustering is to partition a data set into several disjoint subsets in a manner that elements in a subset are more similar to each

other than to elements in other subsets. Clustering refers to the process of grouping samples so that the samples are similar within each group these groups are called clusters. Clustering is DIA; clustering approaches were one of the first techniques used for the segmentation of document images [5]. In partitioned clustering, the aim is to create one set of clusters that partitions the data in to similar groups. In our work we have used K-means clustering approach for performing document image segmentation using matlab.

Texture segmentation has been an important task in image processing. Basically the goal at segmenting a textured document image into several regions having the similar patterns. An effective and efficient texture segmentation technique will be very useful in applications like the analysis of document images. Other segmentation problems, the segmentation of textures required the choice of proper texture specific features with good discriminative power [6].

In image processing a gabor filter is a linear filter used for edge detection. Frequency and orientation representations of gabor filters are similar to those of the human visual system and they have been found to be particularly appropriate for texture representation and discrimination. The document image analysis by the gabor functions is similar to perception in the human visual system [7].

II. DOCUMENT IMAGE ANALYSIS

Document image analysis can be important when the original document is produced by computer. Document image analysis refers to algorithm and techniques that are applied to images of documents to obtain a computer readable description from pixel data. A well known document image analysis product is the Optical Character Recognition (OCR) software that recognizes character in a scanned document. The objective of document image analysis is to recognize the text and graphics components in images of documents and to extract the related information [8].

III. DOCUMENT IMAGE SEGMENTATION

All image processing operations generally aim at a better recognition of objects of interest, i.e., at finding suitable local features that can be distinguished from other objects and from the background. The next step is to check each individual pixel to see whether it belongs to an object of interest or not, this operation is called segmentation and produces a binary document image. A pixel has the value one if it belongs to the object; otherwise it is zero. Segmentation is the operation at the threshold between low-level image processing and document image analysis. After segmentation, it is known that which pixel belongs to which object. The document image is partitioned into regions and we get the discontinuities as the boundaries between the regions [9].

IV. DOCUMENT IMAGE SEGMENTATION TECHNIQUES

The document image segmentation techniques can be discussed are as follows:

A. Edge detection technique

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 image segmentation. Most of the used discontinuity based edge detection methods are reviewed in this paper. Those detectors are Prewitt, Sobel, Canny, Roberts, Zerocross, Laplacian of Gaussian and Marr- Hildreth [10].

a. *Prewitt:* The prewitt edge detector is an appropriate way to estimate the magnitude and orientation of an edge.
b. *Roberts:* The Robert cross algorithm the horizontal and vertical edges bring out individually and then they put together for the resulting edge detection.
c. *Sobel:* The sobel edge detection technique is similar to that of the Robert cross algorithm. The design of Sobel and Robert are common, the basic difference is the kernels that each uses to obtain the document image is different.
d. *Canny:* The famous edge detection algorithm canny first presented in 1986. The canny edge detector is regarded as one of the best edge detector recently used. Canny edge detector ensures good noise immunity and at the same time detects true edge points with minimum error.
e. *Zerocross*: It uses second derivative and it includes Laplacian operator. It is having fixed characteristics in all directions and sensitive to noise.
f. *Laplacian of Gaussian:* It was invented by Marr and Hildreth. (1980) the gausian filtering is combined with Laplacian to break down the document image where the intensity varies to detect the edges effectively [11-12].

B. Region based technique

The objective of segmentation is to partition a document image into regions. Detection of discontinuities and boundary detection approached this problem by finding boundaries

between regions based on discontinuities in gray levels, whereas in thresholding segmentation was accomplished via thresholds based on the distribution of pixel properties, such as gray level value or color. We discuss segmentation techniques that are based on finding the regions directly.

a. *Region Growing:* region growing is a procedure that groups pixels or sub regions into larger regions based on predefined criteria. The basic approach is to start with a set of "seed" points and from these grow regions by appending to each seed those neighboring pixels that have properties similar to the seed.

b. *Region splitting and merging:* Split-and-merge segmentation is based on a quad tree partition of a document image. It is sometimes called quad tree segmentation. This method starts at the root of the tree that represents the whole document image. If it is found non-uniform (not homogeneous), then it is split into four son-squares (the splitting process), and so on so forth. Conversely, if four son-squares are homogeneous, they can be merged as several connected components (the merging process). The node in the tree is a segmented node. This process continues recursively until no further splits or merges are possible [13-14]. When a special data structure is involved in the implementation of the algorithm of the method, its time complexity can reach $O(n \log n)$, an optimal algorithm of the method[15-16].

C. K-mean clustering technique

K mean is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K cluster) fixed apriori. The main idea is to define k centers one for each clusters. These centers should be placed in a cunning way because of different location causes different result. So the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop, we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more [17].

D. Texture based technique

Texture segmentation is the process of partitioning a document image into regions with different textures containing similar group of pixels. Texture has long been an active area of research in the Image processing community. Document image texture analysis methods use different descriptors of texture. The descriptors used capture a different part of the intuitive understanding of texture.

Texture segmentation is simply document image segmentation where the labeling of pixels is based on some measure of texture. Texture segmentation algorithm, then, differ on the actual form of texture used in the measurement [18-20].

E. Gabor filter technique

Gabor filters applied to extract features aligned at particular orientations or angles. It posses optimal localization properties in both spatial and frequency domains and they have been successfully used in many pattern recognition applications [21]. Gabor filters are not optimal when objective is to achieve broad spectral information with maximum spatial localization.

The document image features are extracted by filtering the given document image using a filter bank consisting of a number of gabor filters with various frequencies, resolutions and orientations. The gabor filter is based on a multi-channel filtering theory for processing visual information in the early stages of the human visual systems. In image processing a gabor filter is a linear filter used for edge detection. Frequency and orientation representations of gabor filters are similar to those of the human visual system and they have been found to be particularly appropriate for texture representation and discrimination. The document image analysis by the gabor functions is similar to perception in the human visual system [22].

V. COMPARISON OF DOCUMENT IMAGE SEGMENTATION TECHNIQUES

Document image segmentation is one of the most challenging problems in documentation area and has been studied extensively in the last few decades. One of the common behaviors of document images is that they are inherently fuzzy in most of the cases and do not exhibit discrete boundaries posing major challenge for clear segmentation of desired structure within the document image.

The first technique is edge detection; Edge detection is a challenging problem. The Prewitt and Sobel edge detector is a simple and

appropriate way to estimate the magnitude and orientation of an edge and it is well known edge detectors. Canny edge detector is one of the best, standard and powerful edge detectors; its results are best as compared to other edge detectors or the methods [23].

The Second technique is Region based; Document image segmentation can never be perfect there is an extra and missing region. Corrected the result of segmentation by removing the extra region or merge region with others or splitting regions into more regions. We have implemented the binary image by thresholding; correct thresholding leads to the better [24-25].

The Third technique is K-mean clustering; K-mean is one of the simplest unsupervised learning algorithms that solve the well known clustering problems. K-mean is a simple and easy way to classify a given data set through a certain number of clusters a fixed apriori. K-mean is a simple algorithm that has been adapted to many problem domains [26].

The fourth technique is Texture based; Texture classification algorithm can be evaluated and compared quantitatively. Using the computed textures over a number of textured document images a model of the input classes formed. The texture representation or a texture classification method, the percentage of pixels that are correctly classified may be compared with other known techniques [27-28].

The fifth technique is Gabor filter, Gabor filter enables to capture the local as well as global information in a texture. The properties of gabor filters are exploited for unsupervised texture segmentation. However the use of document image specific knowledge together with gabor features should produce significantly better results for texture classification [29].

VI. EXPERIMENTAL RESULTS

The Segmentation techniques was implemented by using (MATLAB 2011b) and the performance of the segmentation techniques was tested with a number of various document images from ICDAR-UK , NIST and other Databases as follows using edge, region based, k-mean clustering, texture and gabor filter segmentation techniques, and they are compared with each other.

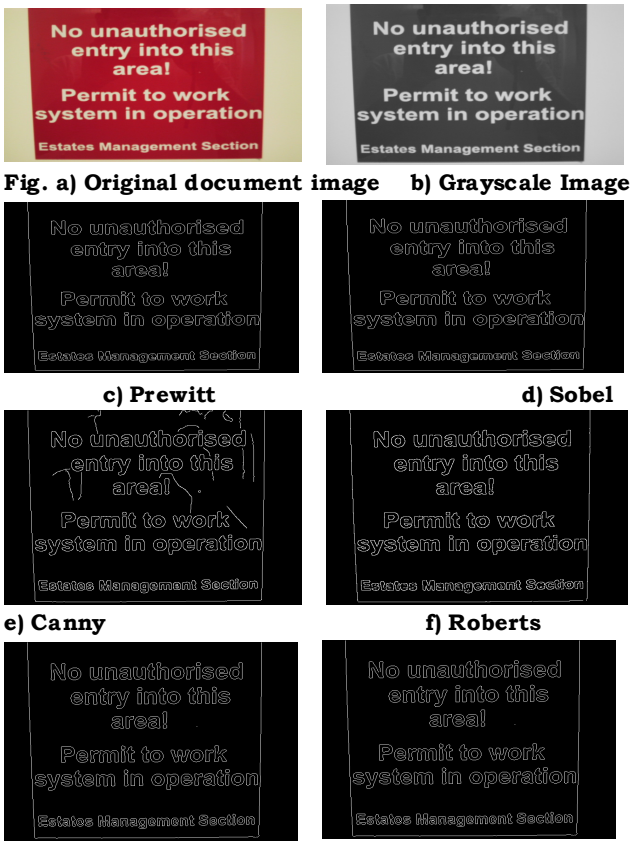


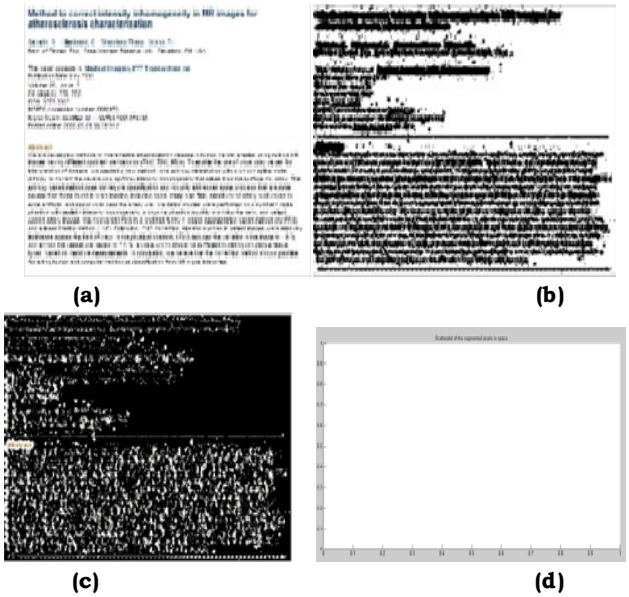
Fig. a) Original document image b) Grayscale Image

c) Prewitt d) Sobel

e) Canny f) Roberts

g) Zerocross h) Laplacian of Gaussian

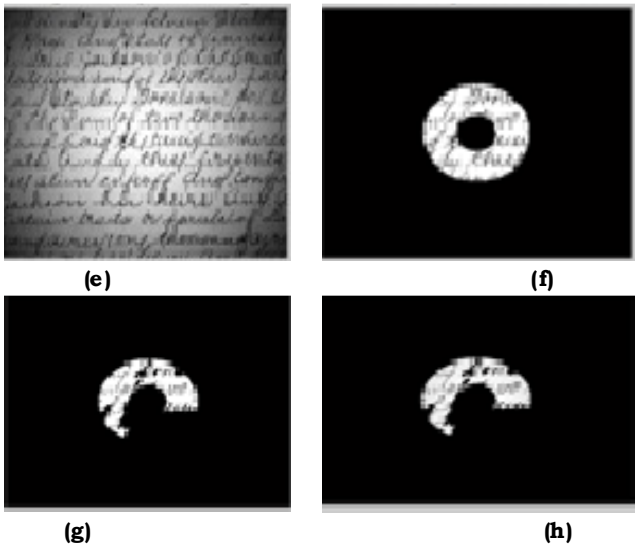
Fig.1 Edge Detection Technique



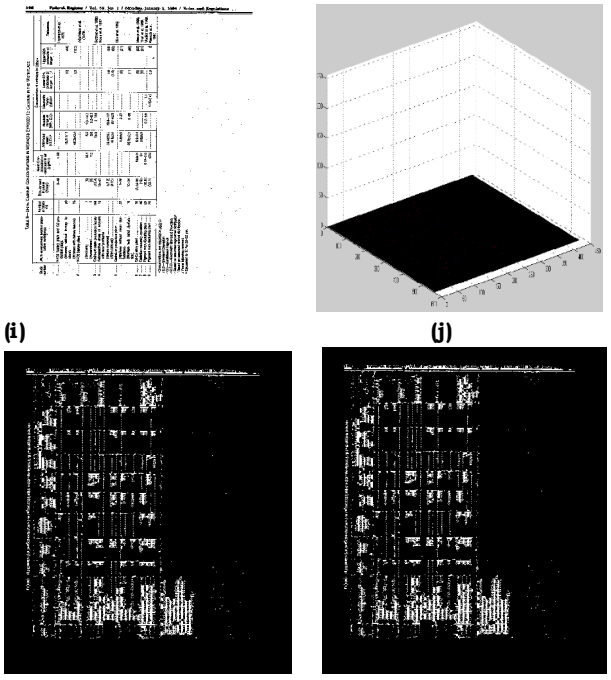
(a) (b)

(c) (d)

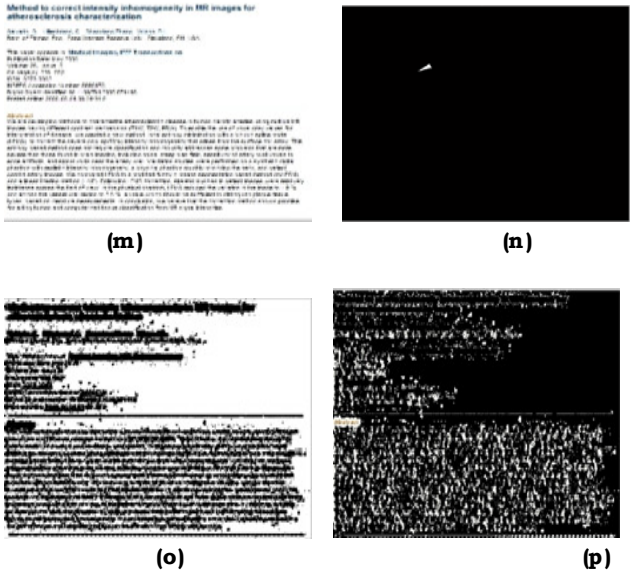
a) Original grayscale image b) & c) Segmented Region based document image, d) Scatterplot of the segmented pixels in 'a*b*' space , xlabel 'a* values' and ylabel 'b*' values.



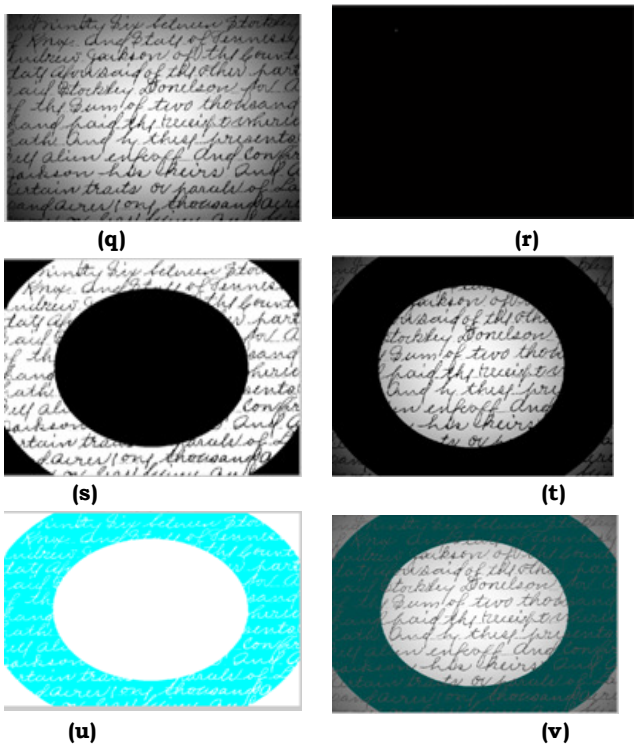
e) Original grayscale image, f) Binary Image within a/-10 Gray levels of 241, g) Magic wand reconstructed binary image h) Magic wand grayscale image.



By using Otsu method i) Original grayscale image, j) The Background Approximation as a Surface, k) Subtract the Background Image from the Original Image, l) Binary image by thresholding.



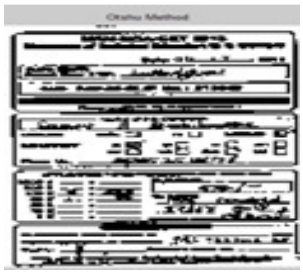
m) Original document image, n) sample region, o) & p) Segmented image.



Problem in seeded region growing algorithm q) Original document image, r) Initial seed image, s) Thresholding the absolute difference between original image & seed value, t) final result, u) Result of region growing process & v) Final result superimposed on original image.

Fig. 2. Region based Technique

(a)



(b)

(a) Original Grayscale image, (b) Otsu Method

(c)

(d)

(e)

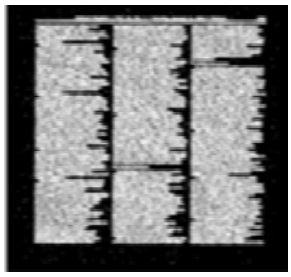
(f)

(g)

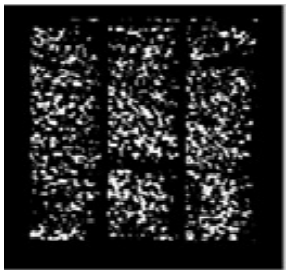
(c) Label every pixel in the document image using the result from K-means clustering (d) Objects in cluster-I (e) Objects in cluster-II (f) Objects in cluster-III (g) Segmenting the nuclei into a separate document image.

Fig. 3. K-means clustering

a) Original Image



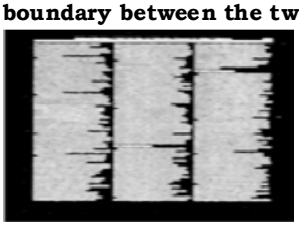
b) Texture Image

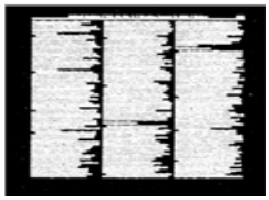


c) & d) Rough mask for bottom texture

e) & f) Rough mask of top texture

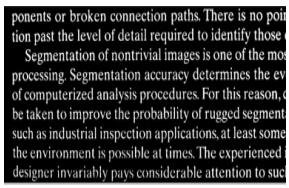
g) Threshold using graythresh h) Outline the boundary between the two texture



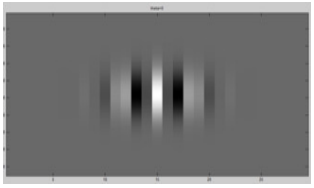


i) & j) Texture filter

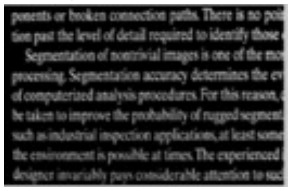
Fig. 4. Texture based technique



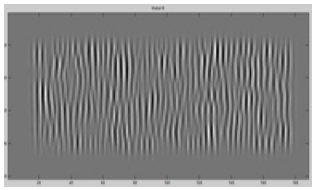
a) Original image



b) Theta = 0



c)Theta = 0



d) Theta = 0

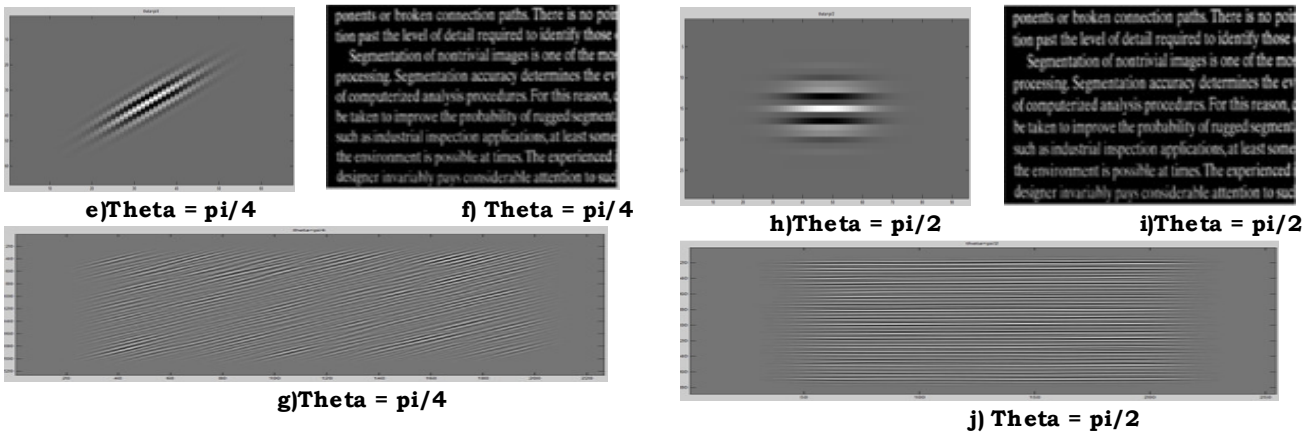


Fig. 5. Gabor filter technique

In the above implemented results by using MATLAB (2008a), figure 1. Shows, we considered a grayscale or a binary document image as input, and returns a binary document image of the same size, with 1's where the function finds edges in image and 0's elsewhere. Figure shows (a) Original document image (b) Grayscale or a binary document image (c) The Prewitt method finds edges using the prewitt approximation to the derivative. It returns edges at those points where the gradient of document image is maximum. (d) The Sobel method finds edges using the sobel approximation to the derivative. It returns edges at those points where the gradient of document image is maximum. (e) the most powerful edge detection method that edge provides is the canny method. The Canny method finds edges by looking for local maxima of the gradient of a document image. The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds, to detect strong and weak edges, and includes the weak edges in the output if they are connected to strong edges. This method is therefore less likely than the others to be fooled by noise, and to detect true weak edges. This method's result is better than all other methods of edge detection technique. (f) The Roberts method finds edges using the Roberts approximation to the derivative. It returns edges at those points where the gradient of document image is maximum. (g) The Zerocross method finds edges by looking for zero crossing after filtering document image with a filter specify. (h) The Laplacian of Gaussian method finds edges by looking for zero crossing after filtering document image with a Laplacian of Gaussian filter.

The existing region based segmentation method is applied for the document images, figure 2 shows a) Original Document Image b) & c) implemented the segmented region based document image and also getting the experimental result in the form of d) scatterplot diagram of segmented pixels in 'a*b*' space values for document image. We can see how well the nearest neighbor classification separated the different pixel populations by plotting the 'a*' and 'b*' values of pixels that were classified into separate pixels. For display purposes, label each point with its pixel label.

In e) original grayscale document image f) Binary Document Image within a/-10 Gray levels of 241, g) Magic wand reconstructed binary document image h) Magic wand grayscale document image in this way we have getting the different outputs. By using Otsu method i) Original grayscale image and implemented in j) Using Morphological Opening to Estimate the Background, The Background Approximation as a Surface, k) Subtract the Background Image from the Original Image, l) Binary image by thresholding**,** removing the background noise, in this way we have getting the required output.

In m) original document image n) From original document image display the sample region and o) & p) display the segmented region from original document image. In this figure shows the better result of segmented region in region based segmentation techniques. Problem in seed region growing algorithm such as q) Original document image r) the initial seed region document image and s) thresholding the absolute difference between original document image t), u) and v) seed value Displaying the better result of region growing process and final result superimposed on original document image.

Figure 3 shows a) the original grayscale document image and in b) result shows the segmenting image by using Otsu method applying on the document image, getting the different clustering.

In c) For every object in original input document image, k-means returns an index corresponding to a cluster. The cluster center result from k-means used, Label every pixel in the document

image with its cluster index. By using pixel label can separate object in original document image by which shows result in d), e) & f). g) Programmatically determine the index of the cluster containing the objects because k-means will not return the same cluster index value every time. Using the cluster center value which contains the mean a* & b* value for each cluster and segmenting the nuclei into a separate document image.

Figure 4 shows, Texture segmentation is the identification of regions based on their texture. The goal is to segment two kinds of fabric in a document image using texture filters. a) Original document image, Threshold the rescaled document image to segment the textures. A threshold value of 0.8 is selected because it is roughly the intensity value of pixels along the boundary between the textures. b) Texture Image, c) & d) Rough mask for bottom texture, Compare the binary document image rough Mask to the original document image. Notice the mask for the bottom texture is not perfect because the mask does not extend to the bottom of the document image. However, we can use rough Mask to segment the top texture. e) & f) Rough mask for top texture, g) Threshold using graythresh h) Outline the boundary between the two texture, i) & j) texture filter.

Figure 5, in Gabor filter technique shows a) Original document image b), c) & d) Theta = 0, e), f) & g) Theta = pi/4, h), i) & j) Theta = pi/2.

VII. CONCLUSION

In this paper a comparison of document image segmentation techniques has been performed on the existing methods.

The first technique is Edge detection; it provides more information about the document image being detected. The Sobel edge detector or operator has been good edge detector as shown in result. The canny edge detector is one of the best edge detection methods for the better performance, standard and more powerful as compare to other edge detection methods.

The second technique is region based document image segmentation; Region-Based Segmentation Methods are an important means of document image segmentation. Otsu method is more proper for document images where objects are distinguished from their background. Document image segmentation can never be perfect there is an extra and missing region.

The Third technique is K-mean clustering, the benefit of K-means clustering that its simplicity, speed and efficient which allows us to run on large amount of datasets. Otsu method is more proper for document images that their objects are distinguished from their background and for document images.

The fourth technique is texture based, an objective way to measure the accuracy of texture segmentation algorithms independent of texture classification was developed and used to compare different texture segmentation algorithm.

The fifth technique is Gabor filter, the use of document image specific knowledge together with gabor features should produce significantly better results for texture classification and being a multiresolution approach takes care difference in size and distribution of texture elements.

REFERENCES

[1] Sukhmanpreet Singh, Deepa Verma, Arun Kumar, Rekha, "Image Segmentation Using Soft Computing".

[2] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", University of Tennessee, MedData Interactve, 2005.

[3] Punam Thakare, "A Study of Image Segmentation and Edge Detection Techniques", International Journal on Computer Science and Engineering (IJCSE).

[4] Yu-Hsiang Wang, "Segmentation", Graduate Institute of Communication Engineering National Taiwan University, Taipei, Taiwan, RO.

[5] S. P. Lloyd, —"Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. IT-28, no. 2, pp. 129–136, Mar.1982.

[6] B.B. Chaudhuri and N. Sarkar N, "Texture Segmentation Using Fractal Dimension", IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(1), 1995, pp. 72-77.

[7] Hans G. Feichtinger, Thomas Strohmer: "Gabor Analysis and Algorithms", Birkhäuser, 1998; ISBN 0-8176-3959-4.

[8] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins "Digital Image Processing using MATLAB" Second Impression, 2007.

[9] Y.S. Pratyusha, N.S. Murthy Dr. K. Sri RamaKrishna, "An Efficient Technique for Segmentation of Characters of Vehicle Identification Number Using Watershed Algorithm", Y.S. Pratyusha et al. / (IJAEST) International Journal of advanced engineering sciences and technologoes Vol No. 5, Issue No. 2, 187 – 194.

[10] 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.

[11] R. Muthukrishnan, M. Radha, "Edge detection techniques for image segmentation", International jJournal of computer Science & Information Technology (IJCSIT) Vol. 3, No 6, Dec. 2011.

[12] Manish T. Wanjari, Mahendra P. Dhore, "Segmenting document images using Edge detection technique", National Conference on Recent Advances in Electronic & Computer science (NCRAECS-2014), ISBN: 978-81-906457-5-1.

[13] S.L. Horowitz and T. Pavlidis, "Picture Segmentation by a Directed Split and Merge Procedure", Proc. ICPR, 1974, Denmark, pp.424-433.

[14] S.L. Horowitz and T. Pavlidis, "Picture Segmentation by a Tree Traversal Algorithm", Journal of the ACM, 23 (1976), pp. 368-388.

[15] L. Chen, "The lambda-connected segmentation and the optimal algorithm for split-and-merge segmentation", Chinese J. Computers, 14(1991), pp 321-331.

[16] M. P. Dhore, Dr. V.M. Thakare, "Morphological Segmentation in Document Image Analysis for Text Document Images", BioInfo Publications International Journal of Computational Intelligence Techniques, ISSN: 0976–0466 & E-ISSN: 0976–0474, vol. 2, Issue 2, 2011, pp-35-43 (Impact Factor 4.47).

[17] Tapas Kanungo, David M. Mount, Nathan S. Netynyahu, Chritine D. Piatko, Ruth Silverman and Angela Y. Wu., " An Efficient K- means Clustering Algorithm: Analysis and Implementation".

[18] R.M. Haralick, "Statistical and structural approaches to texture," Proc. IEEE, vol. 67, no. 5, pp. 786–804, May 1979.

[19] X. Huang, Y. Akiri, and M. Jack, Hidden Markov Models for Speech Recognition, Edinburgh Univ. Press, Edinburgh, 1990.

[20] V. Lakshmanan, V. DeBrunner, R. Rabin, "An Unsupervised, Agglomerative, Spatially Aware Texture Segmentation Technique", IEEE Transactions on Image Processing.

[21] Zheng, D., Zhao,Y., Wang, J.,(2004a) "Features Extraction Using A Gabor Filter Family", Proceedings of the sixth LASTED International conference, Signal and Image processing, Hawaii, USA.

[22] Hans G. Feichtinger, Thomas Strohmer: "Gabor Analysis and Algorithms", Birkhäuser, 1998; ISBN 0-8176-3959-4.

[23] Manish T. Wanjari, Mahendra P. Dhore, "Segmenting document images using Edge detection technique", National Conference on Recent Advances in Electronics and Computer Science-NCRAECS-2014 ISBN 978-81-906457-5-1.

[24] Harikrishna Rai G.N, 2 T.R.Gopalakrishnan Nair, " Gradient Based Seeded Region Grow method for CT Angiographic Image Segmentation", 1Research Associate, RIIC, D S Institutions, SET labs Infosys, Bangalore, hk_rai@yahoo.com 2Director – Research and Industry, Senior Member IEEE, trgnair@ieee.org Dayananda Sagar Institutions, Bangalore.

[25] Manish T. Wanjari, Keshao D. Kalaskar, Dr. Mahendra P. Dhore, "Document Image Segmentation using Region Based Methods" International Journal of Computing Science and Information Technology, ISSN: 2278-9669, July 2015 (http://ijcsit.org) Vol.03, Issue 03, pp. 01-08.

[26] Jose F. Costa and Jackson G. de Souza, "Image Segmentation through clustering Based on Natural computing techniques", Federal University of Rio grande do Norte Brazil.

[27] P.P. Ohanian and R.C. Dubes, "Performance Evaluation for Four Classes of Textural Features," Pattern Recognition, vol. 25, pp. 819–833, 1992.

[28] V. Manian, R. Vasquez, and P. Katiyar, "Texture classification using logical operators," IEEE Trans. on Image Proc., 2000.

[29] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," Pattern Recog., vol. 24, pp. 1167-1186, 1991.

*****