International Journal of Computing Science and Information Technology, **2015, Vol.3, Iss. 3, 01-08**
ISSN: 2278-9669, July 2015 (http://ijcsit.org)

Document Image Segmentation using Region Based Methods

[1]Manish T. Wanjari, [2]Keshao D. Kalaskar, [3]Dr. Mahendra P. Dhore
[1]Department of Computer Science, SSESA's, Science College Congress Nagar, Nagpur (MH), India
[2]Department of Computer Science, Dr. Ambedkar College, Chandrapur (MH), India
[3]Department of Electronics & Computer Science, RTM Nagpur University, Nagpur (MH), India
[1]mwanjari9@gmail.com, [2]keshao_kalaskar@yahoo.co.in, [3]mpdhore@rediffmail.com

Abstract

Document Image Segmentation subdivides a document image into its constituent regions or objects. The level to which the subdivision is carried depends on the problem being solved. The region based segmentation is partitioning of a document image into homogenous areas of connected pixels through the application of homogeneity criteria among candidate sets of pixels. This paper discusses few segmentation techniques that are based on finding the regions. Region based Segmentation techniques can be divided in two categories, which include the Region transformation based such as region growing, region splitting and merging, watershed region and Threshold based such as Global, Local and Dynamic Thresholding. The performance of the region based segmentation is tested with a number of various document images using region based methods, threshold and Otsu method. The performance of experimental results is also discussed in the paper.

Keywords: Image Segmentation, Document Image Segmentation, Region Growing, Region Splitting, Region Merging, Thresholding.

1. Introduction

The objective of segmentation is to partition a document image into regions. We approached this 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. [1] 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. [2]

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. [3] Region based segmentation methods are categorized into two categories such as Region Transformation based such as region growing, region splitting and merging and Threshold based such as global, local and dynamic thresholding are discussed.

Region growing is a simple region-based document image segmentation method. It is classified as a pixel-based document image segmentation method since it includes the selection of initial seed points. The first method was the seeded region growing method. This method takes a set of seeds as input along with the document image. The seeds mark each of the objects to be segmented. The regions are iteratively grown by comparing all unallocated neighboring pixels to the regions. The difference between a pixel's intensity value and the region's mean, δ, is used as a measure of similarity. The pixel with the smallest difference measured this way is allocated to the respective region. This process continues until all pixels are allocated to a region.

The segmentation of regions is an important first step for a variety of document image analysis and visualization tasks. There is a wide range of document image segmentation techniques in the literature; some considered general purpose and some designed for a specific class of document images. Conventional segmentation techniques for monochromatic document images can be categorized into two distinct approaches [4, 5]. One is region based, which relies on the homogeneity of spatially localized features, whereas the other is based on boundary finding, using discontinuity measures. The two methods exploit two different definitions of a region which should ideally yield identical results. Homogeneity is the characteristic of a region and no homogeneity or discontinuity is the characteristic of the boundary of a region. Based on one or both of these properties, diverse approaches to image segmentation exhibiting different characteristics have been suggested. [6, 7, 8]

The complement of the boundary –based approach is to work with the region. Region based method is rely on the postulate that neighboring pixels within the one region have similar value. This leads to the class of algorithm called as region growing of which the "split and merge" technique is probably the known. The regions growing approaches is the opposite of split and merge approach. [9, 10]

Thresholding is the simplest way to perform segmentation, and it is used in extensively in many document image processing applications. Thresholding is based on the notion that regions corresponds to the different regions can be classified by using range function applied to the intensity values of document image pixels.

In [27], Asma Ouji et al. have presented an efficient color segmentation system for noisy document images. The proposed system is generic, since it is applicable on any document structure. All the parameters are automatically computed using novel stroke thickness estimation. The authors have introduced a new measure of pseudo-saturation to detect chromatic pixels and got rid of the saturation noise. Within the chromatic layer, the segmentation is achieved using a double classification using local and global Hue histograms. Similarly, luminance histograms are used to separate the black & white and the gray layers. Such a decomposition allows high quality and local (targeted) binarization avoiding any penalizing loss of information. The resulting layers made the text detection easy and efficient. The acquired color and text information is used to detect ads in press images. Such an issue is innovative as it is the first one to handle ads in complex document images.

In [28], Lacerda et al. have presented an algorithm for segmentation of connected handwritten digits based on the selection of feature points, through a skeletonization process, and the clustering of the touching region via Self-Organizing Maps. The segmentation points are then found, leading to the final segmentation. The method can deal with several types of connection between the digits, having also the ability to map multiple touching.

In [29], the author has proposed new approach to the waterflow algorithm for text line segmentation. the algorithm is applied to documents from left to right and vice versa. The wetted and unwetted areas are established. These areas separate the text and nontext areas from the text lines which represent important control areas for segmentation.

In [30], the authors presented a methodology for detecting and extracting the text lines of images from complex handwritten historical documents. The proposed line segmentation algorithm is based on computing a binary transition map of the document and then extracting and refining the corresponding line regions through skeletonization. To improve the accuracy of line segmentation, a new graph-based splitting method to separate the touching lines is introduced. Once text lines have been segmented, algorithm based on mathematical morphology operators and position heuristics is used to extract the component words on each text line.

2. Region Based Segmentation Method

The objective of segmentation is to partition a document image into regions. Segmentations were accomplished via thresholds based on the distribution of pixel properties, such as gray-level values or color. Region based method are based on continuity. The region based segmentation is partitioning of a document image into homogenous areas of connected pixels through the application of homogeneity criteria among candidate sets of pixels. Each of the pixels in a region is similar with respect to some characteristics or computed property such as color, intensity or texture. Failure to adjust the homogeneity criteria accordingly will produce undesirable results. The following are some of them:

i) The segmented region might be smaller or larger than the actual

ii) Over or under-segmentation of the image (arising of pseudo objects or missing objects)

iii) Fragmentation

Region growing is a simple region-based document image segmentation method. It is classified as a pixel-based document image segmentation method since it includes the selection of initial seed points. This approach to segmentation examines neighboring pixels of initial "seed points" and determines whether the pixel neighbors should be added to the region. The process is iterated on, in the same sequence as general data clustering algorithms. Region-growing approaches exploit the important fact that pixels which are close together have similar gray values**.** [26]

The Region based Segmentation methods are categorized into two categories such as Region

Transformation based and Threshold based is as follows.

2.1. Region Transformation Based

Segmentation consists on partitioning a document image into a set of connected regions.

2.1.1. Region Growing Method

The goal of region growing is to map the input document image data into sets of connected pixels, called regions, according to a prescribed criterion which generally examines the properties of local groups of pixels. The growing starts from a pixel in the approximate of the seed point initially selected by the user. The pixel can be chosen based on either its distance from the seed point or the statistical properties of the neighborhood. Then each of the 4 or 8 neighbours of that pixel are visited to determine if they belong to the same region. This growing expands further by visiting the neighbours of each of these 4 or 8 neighbor pixels. This recursive process continues until either some termination criterion is met or all pixels in the image are examined. The output is a set of connected pixels determined to be located within the region of interest. [11]

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 asset of "seed" points and from these grow regions by appending to each seed those neighboring pixels that have properties similar to the seed (such as specific ranges of gray level or color).

Homogeneity of regions is used as the main segmentation criterion in region growing. The criteria for homogeneity:

- gray level
- color
- texture
- shape
- model

Seeded region growing requires seeds as additional input. The segmentation results are dependent on the choice of seeds. Noise in the image can cause the seeds to be poorly placed. Unseeded region growing is a modified algorithm that doesn't require explicit seeds. It starts off with a single region A_1– the pixel chosen here does not significantly influence final segmentation. At each iteration it considers the neighboring pixels in the same way as seeded region growing. It differs from seeded region growing in that if the minimum δ is less than a predefined threshold T then it is added to the respective region A_j. If not, then the pixel is considered significantly different from all current regions A_i and a new region A_{n+1} is created with this pixel.

One variant of this technique, proposed by Haralick and Shapiro (1985), [12] is based on pixel intensities. The mean and scatter of the region and the intensity of the candidate pixel is used to compute a test statistic. If the test statistic is sufficiently small, the pixel is added to the region, and the region's mean and scatter are recomputed. Otherwise, the pixel is rejected, and is used to form a new region.

A special region-growing method is called λ-connected segmentation. It is based on pixel intensities and neighborhood-linking paths. A degree of connectivity (connectedness) will be calculated based on a path that is formed by pixels. For a certain value ofλ, two pixels are called λ-connected if there is a path linking those two pixels and the connectedness of this path is at least λ . λ -connectedness is an equivalence relation. [13]

Popular fast color document image segmentation is SRM (Statistical Region Merging) with codes publicly available in Java, Matlab or Python [14].

The region growing techniques took on a variety of aspects the block diagram following the potential sequences of processes that can lead to segmentation using region growing.

Region growing approach is the opposite of the split and merges approach:

- An initial set of small areas is iteratively merged according to similarity constraints.
- Start by choosing an arbitrary seed pixel and compare it with neighboring pixels.
- Region is *grown* from the seed pixel by adding in neighboring pixels that are similar, increasing the size of the region.
- When the growth of one region stops we simply choose another seed pixel which does not yet belong to any region and start again.
- This whole process is continued until all pixels belong to some region.
- A *bottom up* method.

Region growing methods often give very well Segmentations that correspond well to the observed edges.

2.1.2. Region Splitting

The basic idea of region splitting is to break the document image into a set of disjoint regions, which are coherent within themselves:

- Initially take the document image as a whole to be the area of interest.
- The area of interest and decide if all pixels contained in the region satisfy some similarity constraint.
- If TRUE then the area of interest corresponds to an entire region in the document image.
- If FALSE split the area of interest (usually into four equal subareas) and consider each of the sub-areas as the area of interest in turn.
- This process continues until no further splitting occurs. In the worst case this happens when the areas are just one pixel in size.

If only a splitting schedule is used then the final segmentation would probably contain many neighboring regions that have identical or similar properties. We need to merge these regions.

2.1.3. Region Merging

The region merging usually depends on the order in which regions are merged. The simplest methods begin merging by starting the segmentation using regions of 2x2, 4x4 or 8x8 pixels. Region descriptions are then based on their statistical gray level properties. A region description is compared with the description of an adjacent region; if they match, they are merged into a larger region and a new region description is computed. Otherwise regions are marked as non-matching. Merging of adjacent regions continues between all neighbors, including newly formed ones. If a region cannot be merged with any of its neighbors, it is marked 'final' and the merging process stops when all document image regions are so marked.

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.[15, 16] 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.[17]

2.1.4. Watershed Region

Watershed region is defined as the region over all point of flows 'downhill' to common point. In geography, a watershed is the ridge that divides areas drained by different river systems. A catchment basin is the geographical area draining into a river or reservoir. The watershed transform applies these ideas to gray scale image processing in a way that can be used to solve a variety of document image segmentation problems. [18]

The watershed is applied of the document image gradient and the watershed lines separate homogeneous regions and giving the desired segmentation results. The gradient document image for the transform is often found using the morphological gradient. Over segmentation is significant problem for most watershed algorithm, which was addressed in various literatures. [19, 20, 21] Conventionally, watershed transform is mostly designed for the objective of document image segmentation.

2.2. Threshold Based

The threshold technique is simplest in segmenting method. Threshold techniques are document image segmentation techniques based on image-space region. [22, 23, 24] It is the existing properties and simplicity of Implementation, image thresholding enjoys a central position in applications of image segmentation. We discuss the ways of choosing the threshold value automatically and consider a method for varying the threshold according to the properties of local image neighborhoods.

2.2.1. Global Threshold

In general of all thresholding techniques is to partition the image histogram by using a single global threshold T. Segmentation is accomplished by scanning the image pixel by pixel and labeling each pixel as object or background, depending on whether the gray level of that pixel is greater or less than the value of T. When T depends only on f(x, y) (i.e. only on gray level values) the threshold is called global threshold.

For choosing a threshold automatically, Gonzalez and Woods describe some of the following procedures

i) Select an initial estimate for T.

ii) Segment the image using T. This will produce two groups of pixels: G1, consisting of all pixels with intensity values > T, and G2, consisting of pixels with values < T.

iii) Compute the average intensity values $\mu 1$ and $\mu 2$ for the pixels in the regions G1 and G2.

iv) Compute a new threshold value

$$T = \frac{1}{2} (\mu_1 + \mu_2)$$

v) Repeat steps 2 through 4 until the difference in T in successive iterations is smaller than a predefined parameter T_0.

2.2.2. Local Threshold

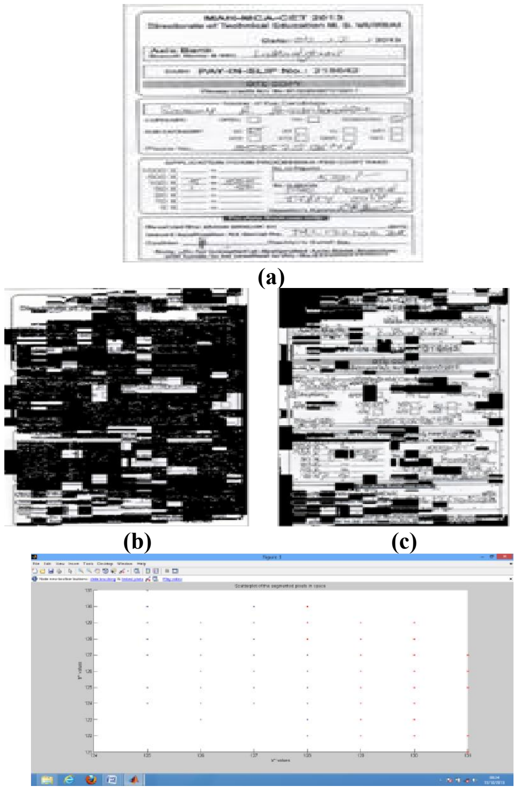
Global thresholding method can fail, when the background illumination is uneven, as illustrated in figure. The improve thresholding result was computed by applying a morphological top-hat operator and then using graythresh on the result. If T depends on f(x, y) and p(x, y), the threshold is called local threshold.

2.2.3. Dynamic Threshold

If, in addition, T depends on the spatial coordinates x and y, the threshold is called dynamic or adoptive threshold. [25]

3. Experimental Result

The performance of the region based segmentation was tested with a number of various document images using region based methods, threshold and Otsu method, and they are compared with the various document images.

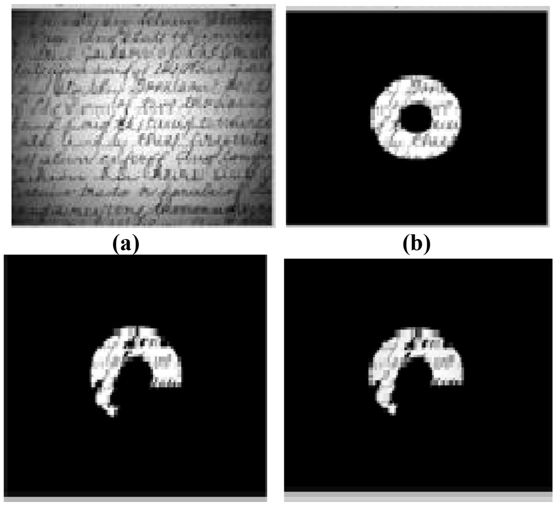


(a)

(b) (c)

(d)

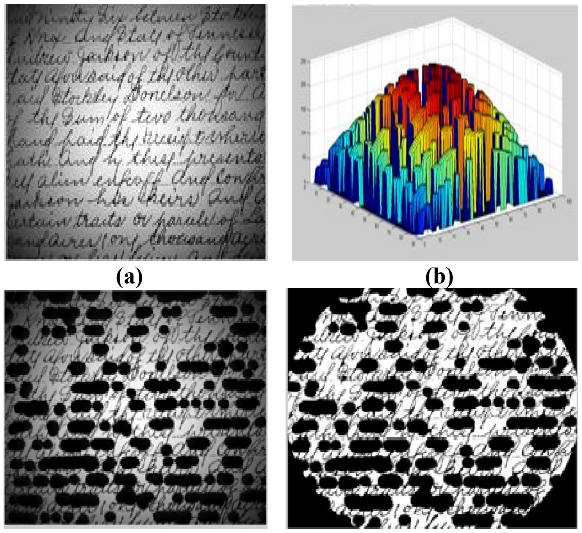
Fig.1.a) Original grayscale image b) & c) Segmented Region based document image, d) Scatterplot of the segmented pixels in 'a*b*' space , xlabel 'a* values' and ylabel 'b*' values.



(a) (b)

(c) (d)

Fig.2. a) Original grayscale image, b) Binary Image within a/-10 Gray levels of 241, c) Magic wand reconstructed binary image d) Magic wand grayscale image.



(a) (b)

(c) (d)

Fig.3 By using Otsu method a) Original grayscale image, b) The Background Approximation as a Surface, c) Subtract the Background Image from the Original Image, d) Binary image by thresholding.

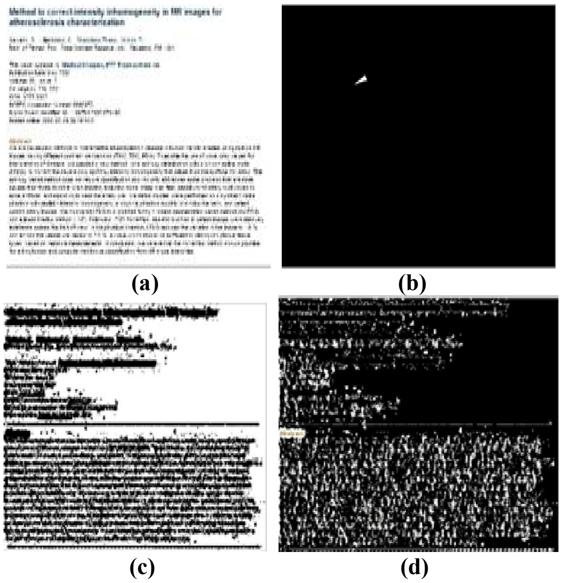


Fig.4. a) Original document image, b) sample region, c) & d) Segmented image.

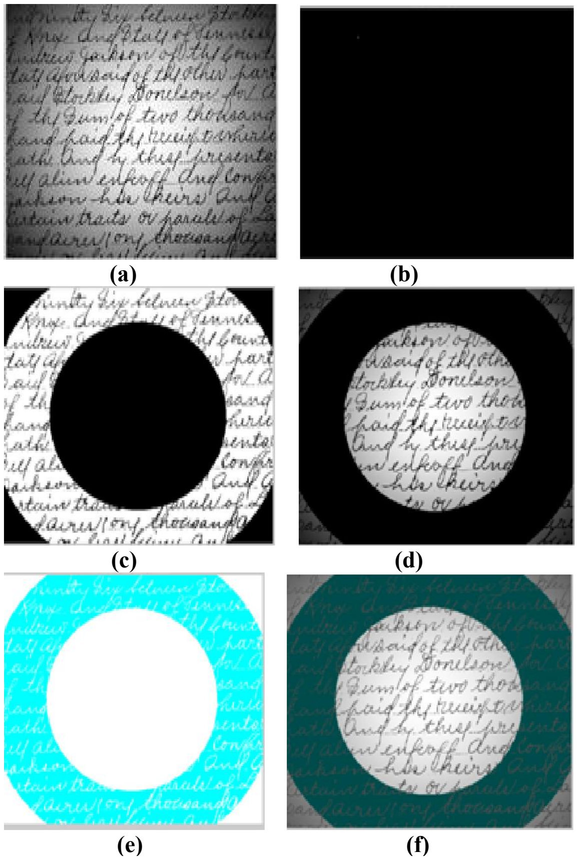


Fig.5 Problem in seeded region growing algorithm a) Original document image, b) Initial seed image, c) Thresholding the absolute difference between original image & seed value, d) final result, e) Result of region growing process & f) Final result superimposed on original image.

The existing region based segmentation method is applied for the document images, In figure 1 shows a) Original Document Image b) & c) implemented the segmented region based document image and also getting the experimental result in the form of 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 figure 2 shows a) original grayscale document image b) Binary Document Image within a/-10 Gray levels of 241, c) Magic wand reconstructed binary document image d) Magic wand grayscale document image in this way we have getting the different outputs.

In figure 3 shows By using Otsu method a) Original grayscale image and implemented in b) Using Morphological Opening to Estimate the Background, The Background Approximation as a Surface, c) Subtract the Background Image from the Original Image, d) Binary image by thresholding, removing the background noise, in this way we have getting the required output.

In figure 4 shows a) original document image b) From original document image display the sample region and c) & d) display the segmented region from original document image. In this figure shows the better result of segmented region in region based segmentation techniques.

In figure 5 shows problem in seed region growing algorithm such as a) Original document image b) the initial seed region document image and c) thresholding the absolute difference between original document image d), e) and f) seed value Displaying the better result of region growing process and final result superimposed on original document image.

5. Conclusion

We have implemented the region based segmentation as applied to gray scale document images. Region-Based Segmentation Methods are an important means of document image segmentation. Otsu method is more proper for images where objects are distinguished from their background. Document image segmentation can never be perfect there is an extra and missing region. Corrected the results 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.

6. Acknowledgements

The authors are thankful to the University Grant Commission, New Delhi for supporting this work as a part of Major Research Project.

REFERENCES

[1] R. C. Gonzalez, R. E. Woods, "Digital Image Processing", *University of Tennessee & Metadata Interactive.*

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

[3] Harikrishna Rai G.N, 2 T.R.Gopalakrishnan Nair, " Gradient Based Seeded Region Grow method for CT Angiographic Image Segmentation", *Research 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.*

[4] S. A. Hojjatoleslami and J. Kittler, "Region Growing: A New Approach" *Ieee Transactions On Image Processing, Vol. 7, No. 7, July 1998.*

[5] D. H. Ballard and C. Brown, "Computer Vision*", Berlin, Germany: Springer Verlag, 1982.*

[6] A. J. Abrantes and J. S. Marques, "A class of constrained clustering algorithms for object boundary extraction," *IEEE Trans. Image Processing, vol. 5, pp. 1507–1521, 1996.*

[7] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans.Pattern Anal. Machine Intell., vol. 16, pp. 641–647, 1994.*

[8] R. M. Haralick and L. G. Shapiro, "Survey: Image segmentation techniques," *CVGIP, vol. 29, pp. 100–132, 1985.*

[9] S. W. Zucker, "Region growing: childhood and adolescence", *Comput Graph Image process, vol. 5, pp 382-399, 1976.*

[10] S. C. Horwitz and T. Pavids, "Picture Segmentation by a directed split-and-merge procedure," *prac. 2nd Int Joint Conf. Pattern Recognition 1974, pp 424-433.*

[11] Jianping Fan, Guihua Zeng , Mathurin Body, Mohand-Said Hacid, "Seeded region growing: an extensive and comparative study", *Pattern Recognition Letters 26 (2005), 1139–1156.*

[12] Linda G. Shapiro and George C. Stockman (2001): "Computer Vision", *pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13- 030796-3.*

[13] L. Chen, H.D. Cheng, and J. Zhang, "Fuzzy subfiber and its application to seismic lithology classification", *Information Sciences: Applications, Vol 1, No 2, pp 77-95, 1994.*

[14] Richard Nock, Frank Nielsen,"Statistical Region Merging", *IEEE Trans. Pattern Anal. Mach. Intell. 26(11): 1452-1458 (2004).*

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

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

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

[18] M. T. Wanjari, K. D. Kalaskar and M. P. Dhore, "Wavelet and Watershed Transform based Document Image Segmentation"*, Proceedinhg of International conference TechEd-2015.*

[19] K. Haris, S. Efstratiadis, "Hybrid Image Segmentation Using Watersheds and Fast Region Merging", *[J].IEEE Transaction on Image Processing, 7, No.12, pp.1684-l699, 1998 on, National Central University, Computer Science and Information Engineering.*

[20] S. Beucher, "The Watershed Transform Applied to Image Segmentation", *Proceedings of the Pfefferkorn Conference on Signal and Image Processing in Microscopy and Microanalysis, pp. 299–314, September 1991.*

[21] A. Bala, "An Improved Watershed Image Segmentation Technique using MATLAB", *International Journal of Scientific & Engineering Research Volume 3, Issue 6, June-2012 1 ISSN 2229-5518.*

[22] W. Malina, S. Ablameyko, W. Pawlak. "Fundamental Methods of Digital Image Processing", *(In Polish), 2002.*

[23] E. A. Savakis., "Adaptive Document Image Thresholding Using Foreground and Background clustering", *published in proceeding of International Conference on Image Processing ICIP, 98.*

[24] L. Spirkovsk., "A Summary of Image Segmentation Techniques", *AmesResearch Center, Moffett Field, California, 1993.*

[25] Rafael C. Gonzalez, Richard E. Woods & Steven L. Eddins, "Digital Image Processing Using MATLAB" , *University of Tennessee, Metadata Interactive & The Mathworks, Inc.*

[26] Shilpa Kamdi, R.K.Krishna, " Image Segmentation and Region Growing Algorithm ", *ISSN 2249-6343 International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 1.*

[27] Ouji, A., Leydier, Y., LeBourgeois, F., "A hierarchical and scalable model for contemporary document image segmentation", *(2013) Pattern Analysis and Applications, 16 (4), pp. 679-693.*

[28] Lacerda, E.B., Mello, C.A.B., "Segmentation of connected handwritten digits using Self-Organizing Maps", *(2013) Expert Systems with Applications, 40 (15), pp. 5867-5877.*

[29] Brodić, D., "Extended approach to water flow algorithm for text line segmentation", *(2012) Journal of Computer Science and echnology, 27 (1), pp. 187-194.*

[30] Sánchez, A., Mello, C.A.B., Suárez, P.D., Lopes, A., "Automatic line and word segmentation applied to densely line-skewed historical handwritten document images", *(2011) Integrated Computer-AidedEngineering,18(2),pp.125-142.*