GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

DOCUMENT IMAGE SEGMENTATION USING HYBRID METHOD

Dr. Manish T. Wanjari*[1], Dr. Keshao D. Kalaskar [2] & Dr. Mahendra P. Dhore[3]

[*1] Department of Computer Science, SSESA's, Science College, Congress Nagar, Nagpur (M.H.), India
[2]Associate Professor, Department of Computer Science, Dr. Ambedkar College, Chandrapur (M.H.), India
[3]Associate Professor, Department of Electronics & Computer Science,
RTMNU, Nagpur (M.H.), Nagpur

ABSTRACT

One of the most important operations in Document Image Analysis (DIA) is segmentation. Document Image Segmentation is a challenging problem in DIA and has been studied extensively in the last few decades. This research paper presents an efficient segmentation technique for Document Images. In this paper we have implemented the Edge detection using canny operator and clustering technique using Otsu's technique. We have proposed a hybrid Document Image Segmentation Technique, combining edge detection using canny operator and clustering using Otsu's technique. The performance metrics MSE, RMSE, PSNR are used for evaluation of images from Tobacco800 and NIST document images databases.

Keywords: *Document Image Analysis (DIA), Document Image Segmentation, Edge Detection, Canny Operator, Clustering.*

I. INTRODUCTION

Document Image Analysis addresses the problem of Separation of text and graphics and their identification and recognition. The methods used for solving this problem generally make use of the differences in the properties of textual and image regions within the document. Many researchers have contributed in solving the problem of segmentation of document images with widely recognized successful results. As such there is no unique ground-truth segmentation of a document image against which the output of an algorithm may be compared. Segmentation is the process of extracting and representing information from the document image. It is the process of partitioning a digital image into multiple segments that means set of pixels. Document image analysis is concerned with quantitative measurement from a document image to produce a description. Document Image Segmentation (DIS) contains transformation of any information presented on paper document into an equivalent symbolic representation accessible to computer information processing. The aim of DIA and understanding is to automatically recognize and extract textual or graphical contents from digitized documents [1, 2].

Segmentation partitions a document image into its constituent parts or objects. In general, autonomous segmentation is one of the most difficult tasks in Digital Image Processing (DIP). A segmentation procedure brings the process a long way toward successful solution of imaging problems that require objects to be identified individually. On other hand, weak or erratic segmentation algorithms almost always guarantee eventual failure. In general, the more accurate the segmentation, the more likely recognition is to succeed.

Segmentation is the process of finding the boundary between foreground and background of a document image [3]. Document Image segmentation process may also partition the scene into more than two mutually exclusive and collectively exhaustive regions. Segmentation can be achieved in spatial domain or grey scale domain. Spatial segmentation is a geometrical boundary between the objects present in the scene based on operations like edge detection, boundary identification etc. The second approach, grey scale thresholding divides the pixels into foreground and background based on a threshold grey value. The pixels on one side of the threshold value are the foreground pixels and the other are identified as the background pixels. This process is called thresholding. Thresholding algorithms may be broadly classified into global, local or adaptive techniques depending on the work

[4]. Algorithms compute thresholds which optimize some objective functions. A noisy document usually undergoes filtering and binarization to separate the characters from the background.

Edge detection is a set of mathematical methods which aim at identifying points in a digital document image at which the document image brightness changes sharply or, more formally has, discontinuities. The points at which document image brightness changes sharply are typically organized into a set of curved line segments called edges [5]. In edge detection technique used canny operator because canny is one of the best operator.

Clustering is the process of grouping samples so that the samples are similar within each group, such groups are called clusters. Clustering approaches were one of the first techniques used for the segmentation of document images [6]. 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 DIS using MATLAB.

II. EDGE DETECTION TECHNIQUE

Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection. Edge-detection method is used for performing DIS tasks. Edge detection is one of the most important element in document image analysis and image processing in computer vision because they apply sound significant role in many applications of image processing particular for machine vision. Edge detection is a method of determining the discontinuities in gray level or binary document images. Edge detection method is the 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 includes various applications such as image processing, biometrics etc. Edge detection is an active area of research as it facilitates higher level DIA. 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. Various operators use the first and second order derivatives of edge detection methods such as canny, Sobel, Roberts, Prewitt, zero cross and Laplacian of guassian [7].

Canny

The Canny edge detector is regarded as one of the best and standard edge detectors. It 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 who considered the mathematical problem of deriving an optimal smoothing filter given the criteria of detection, localization and minimizing multiple responses to a single edge. He showed that the optimal filter given these assumptions is a sum of four exponential terms. He also showed that this filter can be well approximated by first-order derivatives of Gaussians. Canny also introduced the notion of non-maximum suppression, which means that given the pre-smoothing filters, edge points are defined as points where the gradient magnitude assumes a local maximum in the gradient direction.

Looking for the zero crossing of the 2nd derivative along the gradient direction was first proposed by Haralick [8]. It took less than two decades to find a modern geometric variational meaning for that operator that links it to the Marr-Hildreth (zero crossing of the Laplacian) edge detector [9]. Although his work was done in the early days of computer vision, the Canny edge detector (including its variations) is still a state-of-the-art edge detector [10]. Unless the preconditions are particularly suitable, it is hard to find an edge detector that performs significantly better than the Canny edge detector. The Canny-Deriche detector was derived from mathematical criteria as the Canny edge detector, although starting from a discrete viewpoint and then leading to a set of recursive filters for document image smoothing instead of exponential filters or Gaussian filters [11].

III. CLUSTERING TECHNIQUE

A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. Document Image Segmentation is the basis of document image analysis and understanding and a crucial part and an oldest and hardest problem of image processing [12].

Clustering is the classification of objects into different groups. The partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common idea often proximity according to some defined distance measure. Data clustering is the common technique for statistical data analysis, which is used in various fields, including machine learning, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as **k-clustering**.

Otsu

Otsu method is one of the most successful methods for document image thresholding. The objective function of Otsu method is equivalent to that of k-means method in multilevel thresholding. They are both based on a same criterion that minimizes the within-class variance. However, Otsu method is an exhaustive algorithm of searching the global optimal threshold, while K-means is a local optimal method. Moreover, K-means does not require computing a gray-level histogram before running, but Otsu method needs to compute a gray-level histogram firstly. Therefore, K-means can be more efficiently extended to multilevel thresholding method, two-dimensional thresholding method and three dimensional methods than Otsu method.

IV. DESIGN OF PROPOSED DOCUMENT SEGMENTATION TECHNIQUE

In Segmentation approach we implemented two proposed methods. First, combination of Edge detection and clustering using Otsu method. The block diagram of proposed segmentation techniques implementation is as given below.

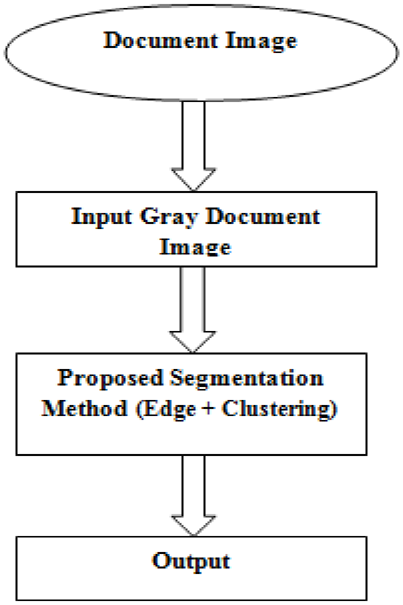

Document Image
Input Gray Document Image
Proposed Segmentation Method (Edge + Clustering)
Output

Figure 1: Block Diagram of Proposed Segmentation techniques in Document Images

The proposed document image segmentation technique is expressed in the form of following algorithm.

Segmentation Algorithm of Proposed Method:

```
Begin
Step - 1 Read the Input Document Image          // Gray Document Image
Step - 2 Preprocess
Step - 3 Combine the Proposed Method            // Edge & Otsu
Step - 4 Calculate the MSE, RMSE and PSNR values
Step - 5 Display the segmentation result
Step - 6 Stop
```

The above document image segmentation technique is implemented on various document images from the Tobacco800 is a public subset of the complex document image processing (CDIP) test collection constructed by Illinois Institute of Technology, assembled from 42 million pages of documents (in 7 million multi-page TIFF images) released by tobacco companies under the Master Settlement Agreement and originally hosted at UCSF and NIST Special Database 25 – volume 1 (NISTIR 6245). This database contains 20 daily Federal Register (FR) issues for the entire month of January 1994. A graphical user interface is developed using MATLAB.

V. PERFORMANCE EVALUATION

The performance of DIA system is measured on the basis on its Mean, Standard Deviation, MSE, RMSE & PSNR for segmentation. The performance metrics used are as follows:

Mean

$$m = \sum_{i=0}^{L-1} zi\, p(zi) \qquad \text{..... (1)}$$

The mean or average that is used to derive the central tendency of the data, it is determined by adding all the data points in a population and then dividing the total by number of points. The resulting number is called as the mean or average.

Standard Deviation (STD)

$$\boldsymbol{\sigma} = \sqrt{\text{⍰}2}\,(z) = \sqrt{\text{⍰}2} \qquad \text{..... (2)}$$

The standard deviation is a numerical value used to indicate how widely individuals in a group mean, the standard deviation big and vice versa. It is important to distinguish between the standard deviation of a sample. They have different notation and they are computed differently. The standard deviation of a population is denoted by σ.

MSE (Mean Square Error)

$$\text{MSE=mean (mean ((im-imf1).^2))} \quad \text{..... (3)}$$

The mean squared error (MSE) of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator.

RMSE (Root Mean Square Error)

$$\text{RMSE = sqrt(MSE)} \quad \text{..... (4)}$$

MSE has the same units of measurement as the square of the quantity being estimated. In an analogy to standard deviation, taking the square root of MSE yields the root-mean-square error (RMSE), which has the same units as the quantity being estimated; for an unbiased estimator, the RMSE is the square root of the variance, known as the standard deviation.

PSNR(Peak Signal-to-Noise Ratio)

PSNR = 10*log10((255^2)/MSE) ….. (5)

The term peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. Because many signals have a very wide dynamic range, (ratio between the largest and smallest possible values of a changeable quantity) the PSNR is usually expressed in terms of the logarithmic decibel scale.

VI. RESULT ANALYSIS

Segmentation results of sample document images

The segmentation results for Proposed Segmentation Techniques using Tobacco800 & NIST database are as below.

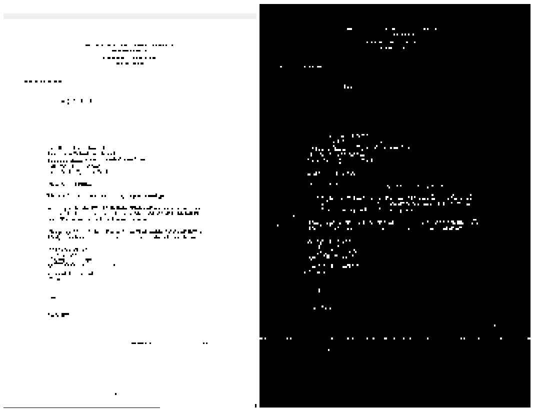


Figure 2: Original Document Image & Proposed Segmentation Techniques

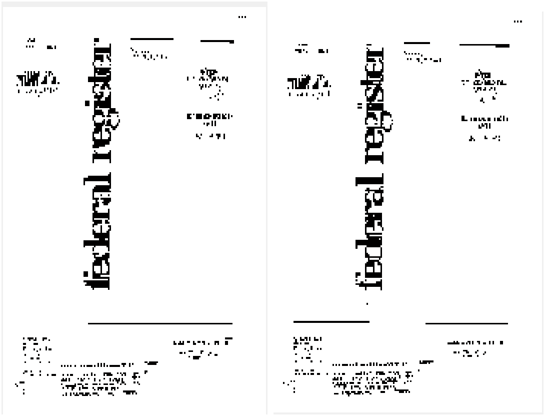


Figure 3: Original Document Image & Proposed Segmentation Techniques

The table 1 & 2 shows the results calculated on Tobacco800 & NIST database. The results calculated on 10 images and find MSE, RMSE & PSNR for document images using proposed technique. From the table, it is clear that the results obtained are the good.

Table 1: Tobacco800 Document Images results for Proposed Segmentation Technique

Document Image	Original document image		Resultant document image		Resultant document image for proposed segmentation technique		
	Mean	STD	Mean	STD	MSE	RMSE	PSNR
Img_001	0.9915	0.0916	0.0049	0.0701	**0.0026**	0.0513	**73.9233**
Img_002	0.9869	0.1136	0.0065	0.0805	0.0030	0.0594	72.6563

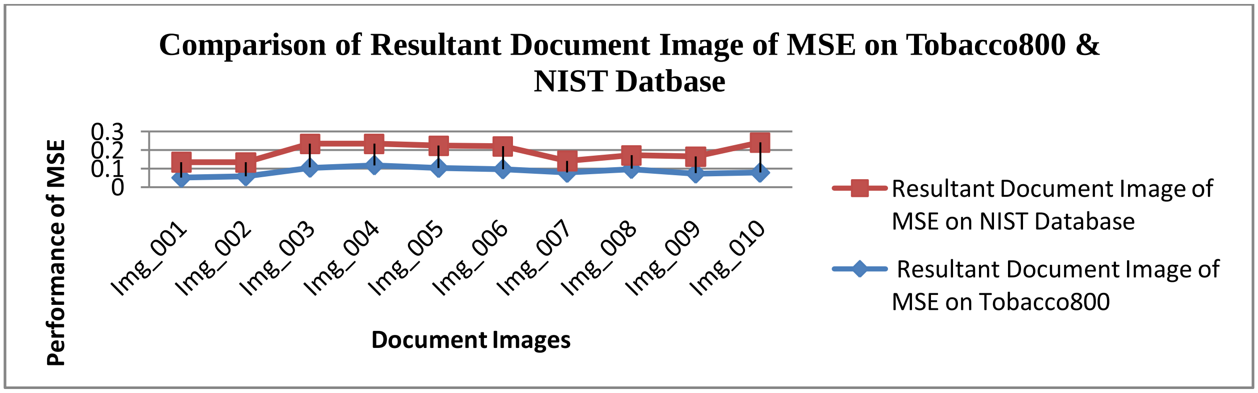
Img_003	0.9710	0.1678	0.0200	0.1399	0.0111	0.1051	67.6967
Img_004	0.9668	0.1791	0.0251	0.1564	0.0141	0.1186	66.6486
Img_005	0.9752	0.1554	0.0187	0.1355	0.0109	0.1042	67.7705
Img_006	0.9638	0.1868	0.0171	0.1297	0.0094	0.0971	68.3846
Img_007	0.9594	0.1973	0.0113	0.1058	0.0061	0.0781	70.2799
Img_008	0.9818	0.1338	0.0170	0.1293	0.0098	0.0988	68.2356
Img_009	0.9784	0.1455	0.0103	0.1011	0.0055	0.0742	70.7260
Img_010	0.9695	0.1720	0.0110	0.1044	0.0062	0.0785	70.2307

Table 2: NIST Document Images results for Proposed Segmentation Technique

Document Image	**Original document image**		**Resultant document image**		**Resultant document image for proposed segmentation technique**		
	Mean	**STD**	**Mean**	**STD**	**MSE**	**RMSE**	**PSNR**
Img_001	0.9528	0.2120	0.0122	0.1096	0.0072	0.08459	69.5456
Img_002	0.9516	0.2147	0.0094	0.0966	0.0055	0.0748	70.6519
Img_003	0.9080	0.2890	0.0303	0.1715	0.0166	0.1290	65.9195
Img_004	0.9223	0.2678	0.0254	0.0157	0.0137	0.1172	66.7531
Img_005	0.9270	0.2602	0.0269	0.0617	0.0146	0.1208	66.4891
Img_006	0.9139	0.2805	0.0285	0.1665	0.0154	0.1242	66.2489
Img_007	0.9784	0.1452	0.0075	0.0863	**0.0042**	0.0645	**71.9387**
Img_008	0.9732	0.1615	0.0098	0.0986	0.0053	0.0728	70.8868
Img_009	0.9505	0.2169	0.0152	0.1223	0.0082	0.0908	68.9686
Img_010	0.8347	0.3714	0.0492	0.2162	0.0264	0.1626	68.9106

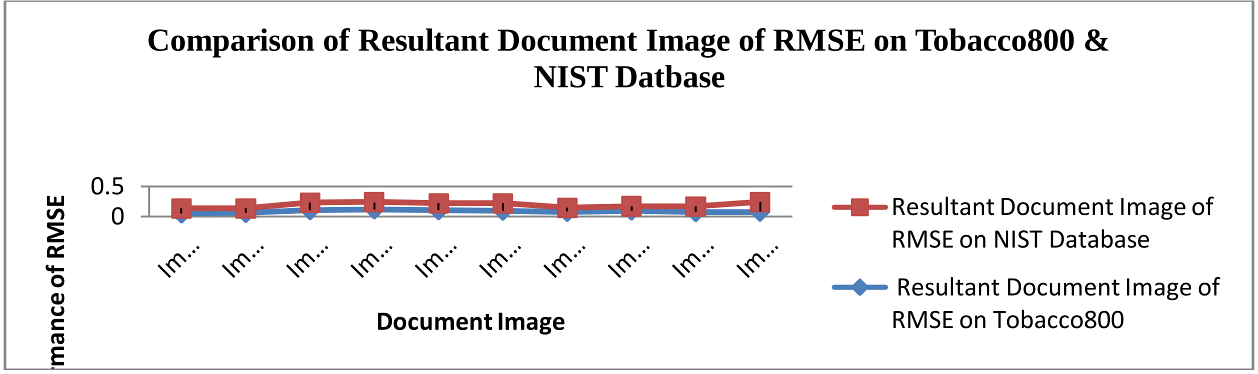
The proposed segmentation techniques were implemented using 10 document images from Tobacco800 & NIST database. The results obtained for the 10 document images are presented in the above table. The table consists of details such as document image, mean and standard deviation of original document image as well as resultant document image. It also consists of values such as MSE, RMSE and PSNR for proposed segmentation technique.

Following graphs shows the comparison of MSE, RMSE and PSNR of resultant document image for the proposed segmentation technique on Tobacco800 & NIST database.

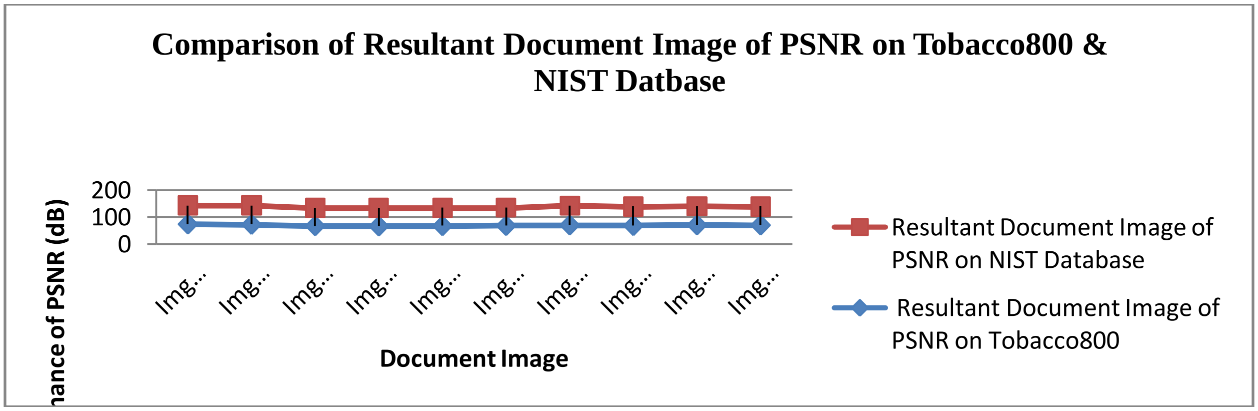


Graph 1: Comparison of Resultant document image of MSE on Tobacco800 & NIST Database for proposed segmentation technique.

ISSN 2348 – 8034
Impact Factor- 5.070



Graph 2: Comparison of Resultant document image of RMSE on Tobacco800 & NIST Database for proposed segmentation technique.



Graph 2: Comparison of Resultant document image of PSNR on Tobacco800 & NIST Database for proposed segmentation technique.

Graphs 1, 2 & 3 show the comparison of MSE, RMSE & PSNR for proposed segmentation technique for document images of Tobacco800 & NIST database. From above performance of the resultant document image of Tobacco800 database results is much better than NIST document image database as MSE values are very low. The lower is the value of MSE, the lesser is the error. PSNR value is almost more than 67 dB.

VII. CONCLUSION

Segmentation techniques are an important task in the implementation of the document image analysis system. In this paper we dealt with document Image segmentation which refers to the process of segments and then represent or recognize object from the document image. Document image analyses recognize the text and graphics components in images of documents.

The Proposed segmentation technique implemented the document image on Tobacco800 & NIST database. The experiments were carried out through Edge detection using canny method, Clustering using Otsu method from document images. We used Mathematical and Statistical metrics such as Mean, STD, MSE, RMSE & PSNR.

The proposed segmentation technique of Tobacco800 & NIST database achieved better performance of MSE, RMSE & PSNR are **0.0026**, 0.0513, **73.9233** and **0.0042**, 0.0645, **71.9387**.The above Performance of the proposed segmentation technique on Tobacco800 database is better than NIST Database, because MSE value is low that means error is lesser and PSNR value is high that means quality of document image is good

REFERENCES

1. J. Freixenet, X. Mu˜noz, D. Raba, J. Mart´ı, and X. Cuf´ "Yet Another Survey on Image Segmentation: Region and Boundary Information Integration" University of Girona. Institute of Informatics and Applications. Campus de Montilivi s/n. 17071. Girona, Spain.
2. S. C. Hinds et al, "A document skew detection method using run-length encoding and the Hough transform", Proc. 10th Int. Conf. Patt. Recogn., 464-468, 1990.
3. Casey, R. G., Wong, K.Y.,(July 1990) "Document Analysis Systems and Techniques, Image Analysis Applications", Image Analysis Applications, pp.1-35.
4. Claude Faure, Nicorevincent, "Document Image Analysis for Active Reading", International Workshop on Semantically Aware Document Processing and Indexing, ISBN 978-1-59593-668-4, pp 7-14, 2007.
5. Gonzalez, Rafael C. Woods, Richard E. and Eddins, Steven L. 2007, "Digital Image Processing using MATLAB', Second Impression.
6. Ms. Chinki Chandhok, Mrs. Soni Chaturvedi, Dr. A. A. Khurashid, "An Approach to Image Segmentation using K-Means Clustering Algorithm", International Journal of Information Technology (IJIT), Volume – 1, Issue – 1, August 2012 ISSN 2279-008X.
7. Ms. Priyanka S. Chikkali, Prof. K. Prabhushetty (2011), "FPGA based image Edge detection and Segmentation",International Journal of Advance Engineering science and Technologies, Vol. No. 9 issue No. 2. , 187-192.
8. R. Haralick, "Digital step edges from zero crossing of second directional derivatives", IEEE Trans. on Pattern Analysis and Machine Intelligence, 6(1):58–68, 1984.
9. R. Kimmel and A.M. Bruckstein, "On regularized Laplacian zero crossings and other optimal edge integrators",International Journal of Computer Vision, 53(3) pages 225-243, 2003.
10. Shapiro L.G. & Stockman G.C., "Computer Vision", London etc.: Prentice Hall, Page 326, 2001.
11. R. Deriche, "Using Canny's criteria to derive an optimal edge detector recursively implemented",Int. J. Computer Vision, vol. 1, pages 167–187, 1987.
12. Wang, Xiao-song; Huang,Xin-yuan and Fu,Hui"The study of color free image segmentation", In: Second International Workshop, Computer Science and Engineering WCSE 09, Sch of Inf.,Beijing Forestry Univ.,Beijing,China (2009).