Automated Colon Cancer Detection Using Kernel Sparse Representation Based Classifier

Seena Thomas, Anjali Vijayan

Abstract— Colon cancer causes deaths of about half a million people every year. Common method of its detection is histopathological tissue analysis, which correlated to the tiredness, experience, and workload of the pathologist. Researchers have been working since decades to get rid of manual inspection, and to develop trustworthy systems for detecting colon cancer. Lesion detection can be difficult due to low contrast between lesions and normal anatomical structures. Lesion characterization is also challenging due to similar spatial characteristics between the tumor and abnormal nodes. To tackle this problem, Gabor wavelet filter algorithm is proposed. The detection of cancerous tissue in tissue image is divided into three main stages. The feature extraction and selection using the Gabor algorithm plays a critical role in the performance of the classifier. Higher accuracy of the classifier can be also achieved by the selection of optimum feature set. Features like the time (spatial) and frequency information can be extracted by using t-test algorithm and the tunable kernel size allows it to perform multi-resolution analysis.

Index Terms— Feature Extraction and Selection, Graph Cut Segmentation, Gabor Filter.

I. INTRODUCTION

Cancer is a class of diseases characterized by out of control cell growth. There are over 100 different types of cancer; each is classified by the type of cell that is initially affected. Colon cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death in men and women combined. In 2014, colon cancer caused 694000 deaths worldwide. The colon and rectum together make up the large intestine, part of the body's digestive system. The colon is a large muscular tube that collects and stores waste then passes into rectum. Tumors can develop within the walls of the colon and/or rectum tissue which are called polyps. These tumors can either be malignant or benign. One of the biggest steps that must be taken in order to prevent colon cancer is to increase screening. Colon cancer starts in the inner lining of the colon, slowly growing through some or all of its layers. Adenocarcinoma is the most common type of colon cancer [2]. For example, colon adenocarcinoma originates from epithelial cells and leads to deformations in the morphology and composition of gland structures formed of the epithelial cells (Figure.1). In the early stages of the disease, colon cancer symptoms may be minimal, or not present at all. As the disease progresses, symptoms may increase in quantity and degree of severity [3]. Screening and tests that can find both colon. When several tests are involved, the ultimate diagnosis may be difficult to obtain even for medical expert. Digital pathology provides a digital environment for the management and interpretation of pathology. Although digital pathology systems are implemented for different purposes, including segmentation, retrieval and tissue image classification. Major advantages of digitals pathology are that slides can be viewed via computer monitors, achieved and retrieved easily and most importantly analyses through software algorithms rather than manual analysis.

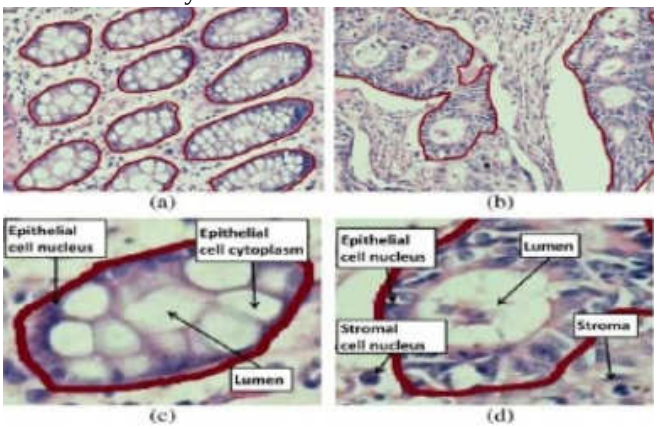


Figure 1: Colon adenocarcinoma changes the morphology and composition of colon glands. In this figure (a) and (b) are normal and cancerous tissue image, (c) and (d) normal and cancerous gland image.

Many computer aided (CAD) systems based on tissue images have been developed to reduce the mortality rate of colon cancer. Image segmentation is considered as most critical stage of data processing among the all the stages of image processing, because of a good classification is dependent on the features extracted from the segmented images [1]. The aim of this study is to analyses digitalized tissue images of colon by applying computer image techniques to enhance tissue images to reduce the unnecessary biopsies. In this work we proposing to use the technique called Gabor Wavelet Filter algorithm. The feature extraction and selection from an image plays a critical role in the performance of the classifier. Higher accuracy of the classifier can be achieved by the selection of optimum feature set. Features like time (spatial) and frequency information are extracted by using t-test algorithm and the tunable kernel size allows it to perform multi-resolution analysis. These features can be passed to a classifier for discrimination for the images to test whether they are low, moderate or high grade samples.

II. METHODOLOGY

Aim of this work is to develop an automated system for colon cancer detection and grading. The cancer diagnosis consists of four main computational steps: pre-processing, segmentation, feature extraction, and classification. Gabor wavelet algorithm used for feature extraction and t-test

Revised Version Manuscript Received on August 29, 2015.

Seena Thomas, Assistant Professor, Department of Computer Science and Engineering, Kerala University, Trivandrum, India.

Anjali Vijayan, Department of Computer Science and Engineering, Kerala University, Trivandrum, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

algorithm used for accurate feature set selection. Kernel Sparse classifier is chosen for final classification. It starts with training stage in which different colon conditions trained to the system. The inputted colon tissue images are processed and at last classified according to the information in the trained set. The figure 2 shows the block diagram of entire system.

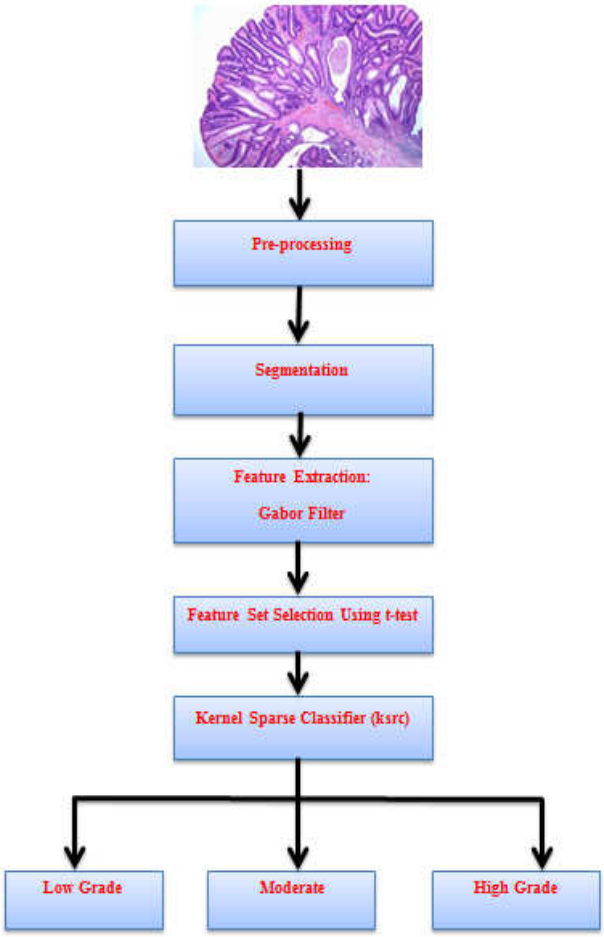


Figure.2: Block diagram for tissue image classification

A. Pre-processing

Digitalized colon tissue images also consist of some noise. Before applying the processing techniques on the images to extract the low level features, the images need to be pre-processed to remove unwanted information and to get enhanced images with the most relevant information. To remove noise and get an enhanced image, the Wiener filter is applied on the grey scale image [10]. The Wiener filter minimizes the mean square error between the estimated random process and the desired process. The Wiener filter is based on statistical approach. It removes additive noise and inverts the blurring simultaneously. Next step is to adjust the contrast of filtered image for get better enhanced image for further processing. We use adaptive histogram equalization (AHE) to improve the contrast in the image. The adaptive method computes several histogram, each corresponding to a distinct section of the image and uses them to redistribute the lightness values of the image. For an example image in figure 2.1, that shows the tissue image after and before pre-processing.

(a)Sample Input (b)After Pre-proceesing

B. Segmentation

Image segmentation is an important and challenging problem and a necessary primary step in image analysis. Traditionally, most segmentation has on unsupervised segmentation, grouping elements of the image according to a criterion such as homogeneity. Recently, supervised image segmentation methods have gained popularly since these methods give the user the ability to affect the segmentation as necessary for a particular application. In this work we using graph cut segmentation for segmenting the tissue image of colon [8].

B.1 Graph Cut

Let an undirected graph be denoted as G = < V, E > where V is a series of vertices and E is the graph edge with connect every two neighbour vertices. The vertex V is composed of two different kinds of nodes (vertices). The first kind of vertices is neighbourhood nodes which correspond to the pixels and the other kind of vertices are called terminal nodes which consist of s (source) and t (sink). This type of graph is also called s-t graph where, in the image s node usually represents the object while t node denote the background and there is links which connect the neighbouring pixels within the image [9]. And the second type of edge is called t-links which connect the terminal nodes with the neighbourhood nodes. Also each edge is assigned with a non-negative weight denoted as w_e which is also named as cost. A cut is a subset of edges E which can be denoted as C and expressed as C⊂E. The cost of the cut |C| is the sum of the weights on edges, C which is expressed as follows

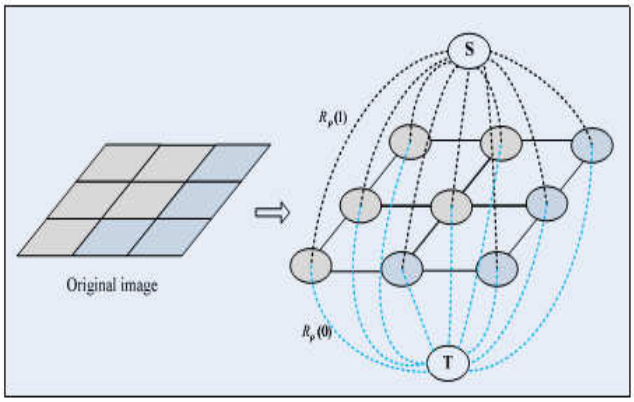


Figure.2.2.1 Illustration of s-t graph

B.2 Graph Cut Segmentation

Image segmentation can be regarded as pixel labelling problems. The label of the object (s-node) is set to be 1 while that of the background (t-node) is given to be 0 and this process can be achieved by minimizing the energy-function through minimum graph cut. In order to make the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

segmentation reasonable, the cut should be occurred at the boundary between object and the background. Namely, at the object boundary, the energy (cut) should be minimized. Let L= {1,2,3,..i,..p,.. } where p is the number of the pixels in the image and {0,1} in ∈ . Thus, the set L is divided into two parts and the pixels labelled with 1 belong to object while others are grouped into background. The energy function is defined as following equation which can be minimized by the min-cut in the s-t graph.

where, *R(L)* is called regional term which incorporates the regional information into the segmentation and *B(L)* is called boundary term which incorporates the boundary constraint into segmentation, α is the relative importance factor between regional and boundary term. When α is set to be 0, it means that the regional information is ignored and only considering the boundary information. Figure 2.2 shows that after graph cut segmentation.

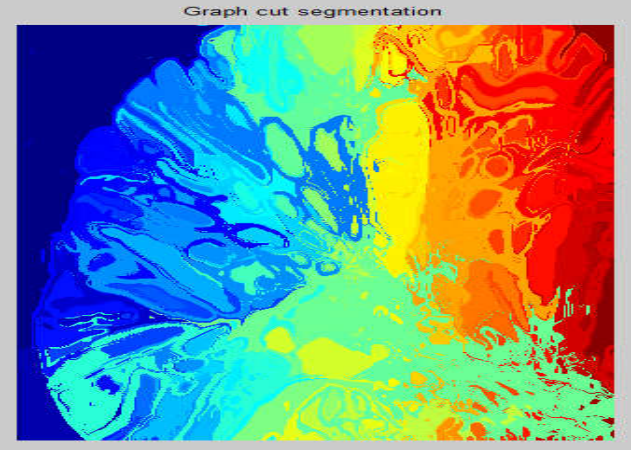


Figure 2.2 : After Graph Cut Segmentation

C. Feature Extraction

A key stage of mass detection and classification by CAD (computer-aided diagnosis) schemes is feature analysis and extraction. The feature space is very large and complex due to the wide diversity of normal tissues and variety of abnormalities. Hundreds of features might be derived from an image. But not all of the features are suitable for mass classification. Too many irrelevant features not only make the classifier complicated, but also will reduce the accuracy of the classification. The most important issue is that to select features that are able to represent the characteristics of masses in the digitalized tissue images, and based on these features; the malignant mass can be significantly discriminated from the benign masses by the classifier. Texture features are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other in the image. The Gray Level Co-occurrence Matrix (GLCM) method is a way of extracting second order statistical texture features. Within the large number of texture features available, some of the features are strongly correlated with each other. From those, different features are selected and tested and finally, subsets of features are accepted as optimum for this work. Co-occurrence matrix of the tissue image is computed first. From these the selected features are extracted. They are mentioned below: Contrast: It measures the local intensity variation. Contrast returns a measure of the intensity contrast between a pixel and its neighbor over the whole image.

$$\Sigma_{i,j}|i-j|^2 p(i,j)$$

where *P(i, j)* indicates co-occurrence matrix.
Homogeneity: It is also called inverse difference movement. It returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

$$\Sigma_{i,j}\frac{p(i,j)}{1+|i-j|}$$

Correlation: It returns a measures of how correlated a pixel is to its neighbor over the whole image.

$$\Sigma_{i,j}\frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j}$$

Energy: It provides the sum of squared elements in the GLCM.

$$\Sigma_{i,j} p(i,j)^2.$$

Mean: Average or mean of matrix elements.

$$\mu = \frac{1}{N}\Sigma_{i=0}^{N-1} x_i$$

Standard Deviation: This feature puts relatively high weights on the elements that differ from the average value of *P (i, j)*

$$\sigma^2 = \frac{1}{N-1}\Sigma_{i=0}^{N-1}(x_i-\mu)^2$$

Skewness: Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

$$s = \frac{E(x-\mu)^3}{\sigma^3}.$$

Kurtosis: It measures the variation in probabilistic distribution.

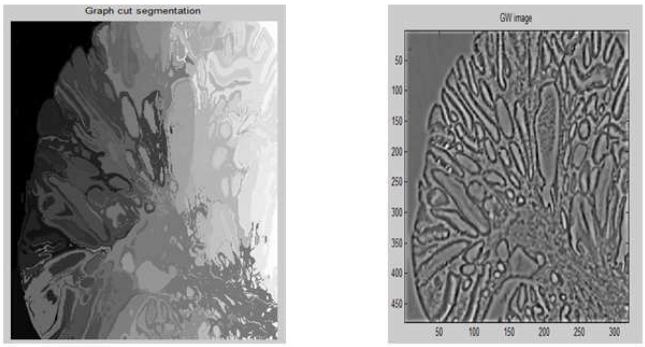
$$k = \frac{E(x-\mu)^4}{\sigma^4}.$$

We propose a Gabor filter algorithm for feature extraction. A set of Gabor filters with different frequencies and orientations may be helpful for extracting useful features from an image. Gabor filters have been widely used in pattern analysis applications. Gabor filters are directly related to Gabor wavelets, since they can be designed for a number of dilations and rotations. In general, expansion is not applied for Gabor wavelets, since this requires computation of bi-orthogonal wavelets, which may be very time-consuming. Usually, a filter bank consisting of Gabor filters with various scales and rotations is created. The filters are convolved with the signal, resulting in a so-called Gabor space. Relations between activations for a specific spatial location are very distinctive between objects in an image. Important activations can be extracted from the Gabor space in order to create a sparse object representation. A circular 2-D Gabor filter in the spatial domain has the following general form,

In spatial domain (above equation) the gabor filter is a complex plane wave (2D Fourier basis function) multiplyed by an origin-centred Gaussian[6]. Gabor features, referred to

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

as Gabor jet, Gabor bank or multi-resolution Gabor feature, are constructed from responses of Gabor filters in by using multiple filters on several frequencies fm and orientations θn.Such Gabor filters have been widely used in various applications. In addition to accurate time-frequency location, they also provide robustness against varying brightness and contrast of images. Furthermore, the filters can model the receptive fields of a simple cell in the primary visual cortex. Based on these properties, in this paper, we try to apply a Gabor filter to tissue images feature extraction. In fact, the imaginary part of the Gabor filter automatically has zero DC because of odd symmetry. This adjusted Gabor filter will convolute with a sub-image. Using this technique, extract the features, then it saved in feature set. After this, generate a Gabor Wavelet Image that is shown in below (Figure 2.3).



(a)Input for feature extraction (b)Output of Gabor Wavelet

Figure 2.3 : Feature extraction using Gabor filter algorithm

D. Feature Selection

For best classification performance, we must find a way to select the most informative subset of features. Reducing the number of features is important. For many data sets with larger number of features and limited number of observations may leads to classification. Reducing features can also save storage and computation time and increase comprehensibility. Feature selection algorithms select a subset of features from the original feature set. Our goal is to reduce the dimension of data by finding a small set of important features which can give good classification performance. Feature selection algorithms can be roughly grouped into two types, filter methods and wrapper methods. Filter are usually used as pre-processing step, because they simple and fast. t-test is used in this proposed system [7]. The most common type of t-test is used to assess whether the means of two classes are statistical different from each other.

F. Classification

The classification of tissue image is achieved in two main steps. In the first step, that is the training stage, selected set of features from the colon tissue images are classified and stored. In the second step, features from the input images are classified using sparse coding based classifier. All samples in training dataset are treated uniformly in the same class during the learning process of KSRC. Sparse classifier is a non – linear classifier. It has been recently demonstrated effective for robust multi – class classification. Kernel methods have the ability to capture the nonlinear relationship between features in high dimensional recognition accuracy[4]. In this work, have to determine whether the tissue images are low, moderate or high grade. In short, a sample in the training dataset may not completely belong to one class. The combination of the CAD scheme and experts knowledge would greatly improve the detection accuracy of the abnormalities.

III. RESULT AND DISCUSSION

In this model structural and Gabor filter features represented by key features. Mainly developed model uses Gabor Filter algorithm for frequency and orientation, t-test for feature reducing the feature set and kernel sparse classifier is used to classify the tissue image as low, moderate and high grade of malignancy, and these are based on features extracted from the tissue images. The features extracted from the tissue images are contrast, energy, homogeneity, correlation, mean, standard deviation, skewness, kurtosis and gabor wavelet features. Then the tissue image classification is mainly performed with the help of kernel sparse classifier. From the classification result identify the grade of colon cancer.

The accuracy of image also calculated based on True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) values. The above values are defined based on confusion matrix. The table 3.1 shows that the feature value comparison. This work also compare with other model, which is not used the gabor filter algorithm and t-test and also using SVM classifier. Figure 3.2 shows that the comparison (ROC) of existing and proposed models.

Table 3.1 : Feature Value Comparison

	LOW GRADE	MODERATE	HIGH GRADE
Contrast	6.3285	8.3261	9.4458
Correlation	0.1794	0.1402	0.1123
Energy	0.0236	0.0179	0.0102
Homogeneity	0.4811	0.4372	0.4283
Mean	0.1444	1.6284	9.1108
Standard Deviation	3.4815	2.1957	0.1721
Skewness	26.3184	55.4021	222.483
Kurtosis	731.85	134.2901	5.44

The proposed system is tested using a database containing 50 high grades, 50 moderate grade and 10 low grade images. 100 cancerous images are taken out of which 50 are high grade cases and 50 are moderate cases. Within each case different types of images are chosen with different mass size, shape, orientation etc. Image of type .jpg format is accepted.

Table 3.2 Experimental Results of previous method

No. of Colon Images	Successfully Classified Images	Failed to Classify	Efficiency in %
5	5	0	100
15	14	1	98
25	23	2	96

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

50	46	4	94
100	91	8	92

In table 3.2 shows the experimental result of previous work with SVM classifier and table 3.3 shows that experimental result of proposed method. Figure 3.1 show the analysis of efficiency of these two methods. Classification is based on KSRC classifier. The satisfying performance of 94% classification rate demonstrates that this system is valuable to improve diagnosis and grading.

Table 3.3 Experimental Results of proposed method

No. of Colon Images	Successfully Classified Images	Failed to Classify	Efficiency in %
5	5	0	100
15	15	0	100
25	24	1	98
50	48	2	96
100	96	6	96

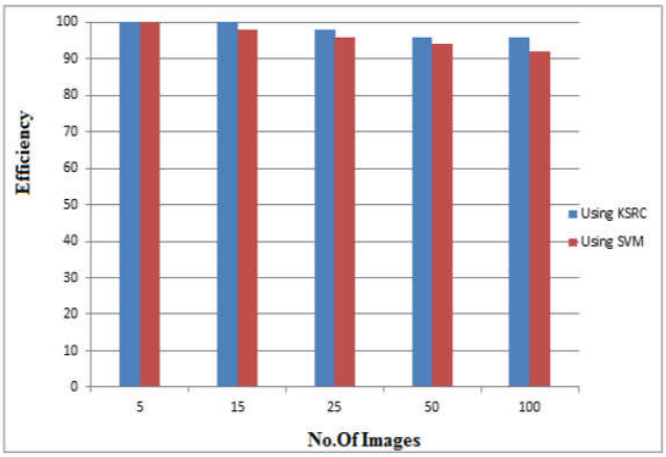


Fig 3.1: Colon Cancer detected using SVM and proposed method

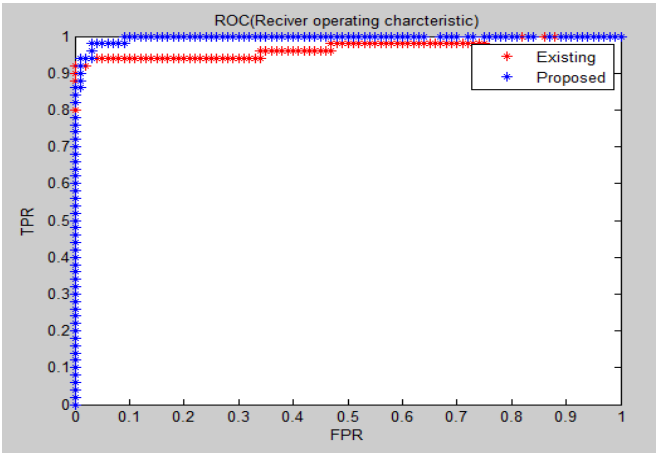


Fig 3.2 : ROC of proposed model

IV. CONCLUSION

Colon cancer disease is among the most serious cancer types in the world but curable ones if it can be diagnosed early. The developed novel model that makes use of gabor filter algorithm for feature extraction, t-test for reducing the feature set and kernel sparse classifier for non – linear classification, these techniques are used for tissue image classification. The detection of tumors in colon tissue image is divided into three main stages. Higher accuracy of the classifier can be achieved by the selection of optimum feature set. Features like the time (spatial) and frequency information are extracted by using t-test algorithm and the tunable kernel size allows it to perform multi-resolution analysis. The advantages of this model are high classification accuracy and higher efficiency. The satisfying performance of 94% classification rate demonstrates that this study is valuable to improve early diagnosis. This work can be improved by revising the segmentation function as well as by using different set of features.

REFERENCES

1. Saima Rathore, Mutawarra Hussain, Ahmad Ali, and Asifullah Khan, (2013), "A Recent Survey on Colon Cancer Detection Techniques" *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 10, No.3.*
2. Ashiya,(2013) "Notes on the Structure and Functions of Large Intestine of Human Body," http://www.preservearticles.com/201105216897/notes-on-the-structure-and-functions-of-large-intestine-of-human-body.html/
3. PubMed Health, "Colon Cancer," http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001308/, Feb. 2013.
4. Ju Han, Hang Chang, Leandro Loss, Kai Zhang, Fredrick L. Baehner, Joe W. Gray, Paul Spellman, and Bahram Parvin, (2011) "Comparison of Sparse Coding and Kernel Methods for Histopathological Classification of Gliobastoma Multiforeme ", Proc IEEE Int Symp Biomed Imaging.
5. Sufan Y Ababneh, Jeff W Prescott, Metin N,(2011)," Automatic Graph Cut Based Segmentation of bones from knee magnetic resonance image for osteoarthritics research", Elseviar/Medical Image Analysis 15,438-448
6. John G Daugman, (2009), " Uncertainly relation for resolution in space, spatial frequency and Orientation optimised by Two Dimensional visual cortical filters", J.Opt.Soc.Am A/Vol 2.No.7/July.
7. Deqip Wang, Hui Zhang, Rui Liu, Weifcng Lv, Dutao Wang, (2014), " t-Test feature selection approach based on term frequency for text categorization", Elseviar/Pattern Recognition Letters,45,1-10
8. A. D. Belsare and M. M. Mushrif , (2012), "Histopathological Image analysis using Image Processing Techniques : An Overview", Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.4
9. Cigdem Gunduz- Demir, Melih Kandemir, Akif Burak Tosun, Cenk Sokmensuer, (2010), "automatic segmentation of colon glands using object- graphs," Elsevier Medical Image Analysis,vol. 14pp.1-12.
10. Erdem Ozdemir, Cigdem Gunduz-Demir, (2013), "A hybrid classification model for digital pathology using structural and statistical pattern recognition," IEEE Trans. Knowledge Medical Imaging., vol. 32, no. 2, pp. 474-483.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.