Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

PREDICTION OF CANCER USING HISTOPATHOLOGY BY APPLYING MACHINE LEARNING ALGORITHMS

E.NAGARAJAN [1], CH.UTHPALA [2], CH.DEEKSHITHA [3]

[1]Assistant Professor, Department of Computer Science and Engineering, Sathyabama University, Chennai-119
[2, 3] U.G Student, Department of Computer Science and Engineering, Sathyabama University, Chennai-119, India

ABSTRACT

Predicting cancer at its early stages has always been a challenging task. Different ways are already in practice to detect cancer, which of most are very costly methods. Our objective is to propose a cancer diagnosing application that uses the simple digital images of human tissues. An extreme impression of digital images on changing society is proving it as an important component in science and technology. Histopathology plays an important role in automated diagnosis of cancer using tissue images. Our work aims to develop an infrastructure to design an automated cancer diagnosing system by processing and extracting data and information from the biopsy tissue images. The tissue images are segmented using the component graph of watershed transform to extract the pixels of the image. The curvilinear structures and features of the tissue image are extracted and compared with the features of the trained database images. A robust matching algorithm SIFT, with repeated structures is used for image comparison. The images stored to database in an array are iteratively used for feature comparison. The algorithm results can be used to conclude whether the given image is cancerous or not. The early detection of cancer disease increases the probability of curing the disease. Our discussion was helpful in building this application at a very low cost whose usage is easy and effortless. The implementation of this application includes techniques of simple image processing, image transformation and image comparison. The machine learning algorithms are opted to make it easy for the application to train itself with the changes that might be made to application or data-base for future use.

KEYWORDS: ***Digital image, watershed transform, curvilinear structure, image processing, SIFT algorithm.***

CH.DEEKSHITHA
U.G Student, Department of Computer Science and Engineering, Sathyabama University, Chennai-119, India

Received on: 30-11-2016
Revised and Accepted on 22-02-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.2.b279-284

INTRODUCTION

The recent surveys show that cancer is one of the most serious health problems in the world. [1]In recent years the image processing techniques are finding place in medical areas for improving the early detection and treatment stages, since the time factor is very important to discover and treat the disease in patient as fast as possible, especially in different kinds of cancer. This problem set is designed to demonstrate the advancement in cancer detection research made possible by adopting the image processing techniques. Cancer diagnoses are usually based on visual recognition of abnormal cell growth and harvested tissues from biopsies. When a pathologist does a job of evaluating a tissue sample, he observers the stained tissue slides, received from the pathology department and tries to analyze it by comparing it using their knowledge and experience and may seek experts help when needed. Besides, the pathologists try to match the structural patterns with their memory of samples and mentally recall the diagnosis process so that they can make some diagnosis in the specific test case. There is a chance for existence of controversial areas and inter-observer variations even after this training.

Biopsy and cancer tissues

A sample tissue taken from any part of the body for further analysis and inspection is called as “biopsy”. Biopsies are most often done to look for cancer. Recent advances in quantitative histopathology have made possible a much efficient way by which tumors can be diagnosed. An improvement in diagnostic accuracy is essential for further proceedings of cancer diagnosis and treatment is made possible through the quality development and integrity. Histopathology plays an important role in automated diagnosis of cancer using tissue images. A huge amount of image data is managed effectively with an intension to provide cases that have similarity with the test cases being inspected. Malignant lesions are not easily and readily distinguishable from normal tissue samples. Using high magnification techniques, apparent changes of the cancer tissue nucleus can be detected. The untightened chromatin present in lesions starts appearing denser due to passage of less amount of light through the clustered granules during trans-illumination.

Effective image processing

The tissue images are used to extract different morphological features such as color and texture. The process of image segmentation helps in isolating the data of interest like lesion or nuclei from the tissue background. The analysis of digital nuclear images helpsin unlocking the data invisible to naked eye. Here we propose an automated computerized system that segments the input tissue image through machine learned algorithm and extract features to compare them with the database of trained features. Our proposal is to discover the fundamental truth of tissue images resemblance using different tissue properties rather than any single morphological feature/element.

Our manuscript describes the pre-processing of given image to remove noise and the image is enhanced to make it easy to define the information entropy. Then the image is segmented through the water-shed algorithm and the morphological features of image are extracted to compare them with image features in database. The features of tissue sample images are considered to find the resemblance with trained tissue images stored to database through the retrieval process by stabilizing and improving the automated cancer detection.

RELATED WORK

Subashini et all. proposed a system[1] in which the image is changed into black and white i.e binary image using threshold. After this Fourier transform is used which helps to perform low pass filter and high pass filter for binary image. Then a log transformation is used to compress the light pixel and expand the dark pixel so that it is visible clearly.

$s = k\log(1 + a)$ where k is a constant and a >=0.

After this process the binary image may be blur, may have noises, sounds etc. Again all the process should be done from beginning and this is the main disadvantage. This process may take lot of time for finding the cancer affected area in the tissue. In the “Tissue sensing adaptive radar” (TSAR) [2] proposed by Sill and Fear, a method of microwave breast imaging is used for tumor detection. A resistively loaded Wu–King monopole antenna is fabricated, and reflections from the breast model over the frequency range of 1–10 GHz are recorded. It involves skin subtraction, focusing and tumor detection using through antenna fabrication. In the novel image processing algorithm proposed by Paramkusham et al. through their paper ”Early stage detection of Breast cancer using Novel Image processing techniques, Mat lab and Lab view implementation”[3], the algorithm is implemented for a.)Masses b.)Superposition c.)Extraction d.)This is totally done on Mat lab i.e. comparing the images. Fuzzy C-Means clustering is used to obtain fuzzy information related to cancerous tissues. This was proposed by Feng et al. for their work to detect prostate cancer through segmentation from multipara- metric MRI based on fuzzy Bayesian model in the year 2014[4].

METHODOLOGY

Before the image is processed for cancer detection it should be pre-processed to refine and enhance the image data which reduces the development of un-desired distortion, and improve some important features of the images useful for further processing. Firstly the given image is enhanced to improve the interpretability and perception of image information and to provide a better image for further processing. Errors encountered during the process of image load is resulted in the form of pixels preventing the reflection of real image’s intensity is termed as “noise”. This noise is eliminated from the image in such a way that the real image is kept discernible and sharp. Our approach consists of three main phases. The first level consists of image segmentation of a gray-scale image obtained though image processing using water-shed method. It starts by restoring the image whose edges are detected using the topological gradient; a simple way of restoring the real image ‘x’ after removing its noise is made possible by solving the PDE problem formulated as below.

$$\begin{cases} -div(k\nabla x) + x = y & in \quad \Omega, \\ \partial_n x = 0 & on \quad \partial\Omega, \end{cases}$$

Where k is a +ve constant (called the conductivity), ∂n denotes the normal derivative and n is the outward unit normal to ∂Ω. And then the watershed transformation takes place to enhance the digital analysis of the segmented image.It was proposed in[5], in which the given images are primarily changed as a directed valued graphs of different neighboring points, called the ingredients of graph 'f', on which the algorithm of watershed transform explained below is applied to compute the transform values of the input image[6].

The steps are as follows:

Algorithm 1: Smallest Distance Watershed

Input: Digital grey scale image *G* = (*N,E,im)* with cost function *cst*.
Output: labeled image *lbl*on N.
#defined WSHED as 0
//uses image *dst.*on output, *dst [x] =im[x],* for all x∈ N
For all *x∈N do*
//Initialize
Assign lbl[x] = 0; *dst [x]* = ∞
End for
For all local minimum values (*mi)* do
For all x∈ mi do
Assign Lbl[x] = i; dst [x] = im[x]
End for
End for
While N ≠ empty set do
y= Minimum-distance (V)
// find *y*∈ N with shortest distance, *dst [y]*
N=N \ {y}
For all *x*∈ N with *(x,y)* ∈*E* do
If *dst [y] + cst[x, y] <dst[x]* then
dst [x]←dst[y]+cst(x, y)
lbl [x]←lbl[y]
Else if *lbl[x]* ≠WSHED and *dst [y] + cst[x, y] = dst[x]* and *lbl[x]* ≠*lbl[y]* then
Lbl[x] = WSHED
End if
End for
End while

The watershed based transform proves it's supremacy over other models by proving closed contours, which is found most useful while segmenting an image. The other is, this process takes very low computation times when compared to other segmentation methods. The segmentation produced by the Open CV implementation is driven completely by the user-supplied or inputted seeds to the segmentation algorithm. In fact, the number of image segments produced by the Open CV algorithm equals with the number of seeds inputted by the user --- even when two different seed are placed in the same homogeneous region of the image.

Recognising image feature

The easily recognizable image features like "Edges" are called as low-level image features and are seen with naked eyes without much trouble. They store the very significant features, making it easy to recognize the image data and components of the image from its edge-detected version. There is still a chance for existence of other low-level features that could be used in computer vision, one of which is curvature. The change in direction of edge may be called as "curvature". The rate in change is defined by the number of points in a curve; points with rapid direction change are called as corners, and points with a small change in direction of edge are termed to be straight lines. The shape description and key point matching are made easy through the use of these features which hold the important data of the processed image. As soon as the segmentation is done, the images are processed for extracting the essential features using the principal curvature-based region detector (PCBR) method. The structure based detectors are used for detecting regions. The curvature based detector is a structure-based affine-invariant detector. This process takes place in few steps that are explained below briefly.

Curvilinear structure detection

Instead of edges the "structure-based detector uses curved features termed as "ridges". A single response is given for all curvature based features like edges, curves and lines. The Hessian matrix to get the principle curvature is as given below

$$L(K) = \begin{bmatrix} L_{XX}(K) & L_{XY}(K) \\ L_{XY}(K) & L_{YY}(K) \end{bmatrix}$$

Finding attributes and fault-tolerant system
Determining image- regions using watershed transformations
Selecting stable regions.

Digital –analysis

After the features are extracted from the images using above mentioned methods and algorithms the images are analyzed digitally.

Figure (1)[10]
The digital image (pixels) of cancer tissue

The histogram or analytical description like color, etc.… of a digital image is constructed using the pixel data of the digitalized-image. The detailed optical values called pixels are described through the "digital image" format.

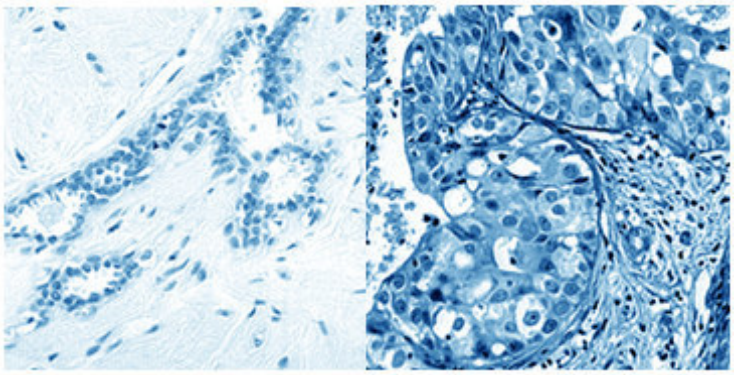

Figure (2)[10]
Normal breast tissue DCIS cancer of breast

In the digital analysis shown in fig(3), the left-histogram represents the normal tissue analysis whose major part of pixels are below 0.5 and most other pixel s are purely white(0.0 values recorded). The other histogram on right represents the digital analysis of a DCIS (Ductal Carcinoma in situ) sample, whose significant pixel values are above 0.5 and most others reaching 1.0 showing pure black pixels. The vigorous cancer activity can be found through this fault-less and definite test conducted.

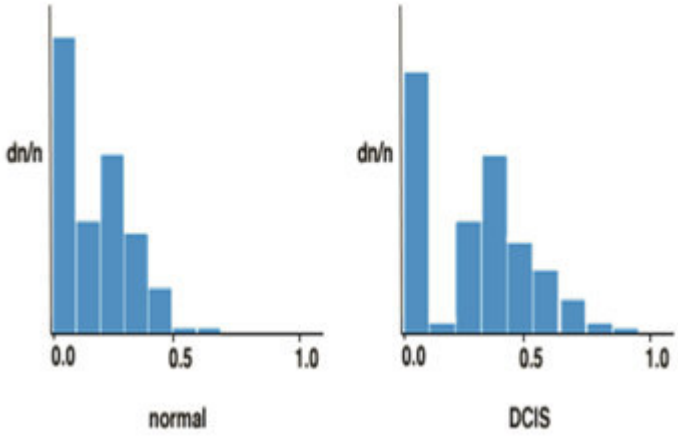

Figure (3)[10]
The Digital analysis of tissue image

It is necessary to conduct a high-order statistical analysis in order to characterize lesions and tumors.

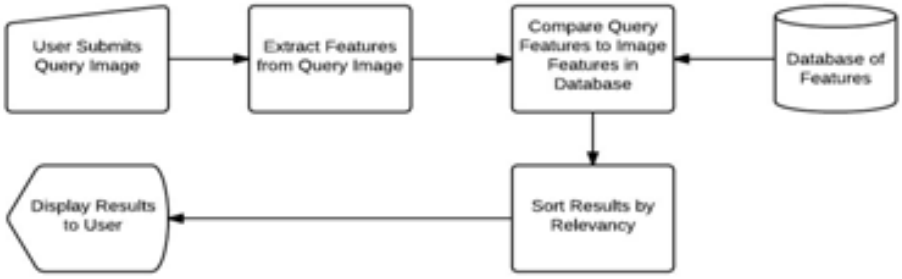

Figure (4)
The overall work flow diagram

Images are retrieved from data set by comparing the color and texture features. The color features are compared using color attributes namely, 1) Momentum of color 2) color histogram 3) Color Coherence Vector (CCV).

Detecting local features in images

To judge whether the input tissue image is cancerous or not, it is essential to compare the features extracted from the input image and the features of the images stored to the database of tissue images. The SIFT (Scale Invariant Feature Transform) method can be used at a vast extent to detect the local features of an image. A complete stage description and implementation of the SIFT algorithm can be briefly explained in following steps[7]. 1. The Gaussian scale-space is calculated, 2. The Difference of Gaussian (DOG) is calculated with the 'σ' as the standard deviation for Gaussian function,

$$g_\sigma(L) = Le^{-\frac{L^2}{2\sigma^2}}, -[4\sigma] \le L \le [4\sigma], \qquad L \epsilon Z$$

L is set so that $\Sigma g\sigma\ (L)$ =1. Let G denote the digital Gaussian convolution of parameter σ and v be a digital form of an image of size X*Y. The two-dimensional (2D) convolution (discrete) is used to calculate Gσv, "Gaussian smoothing constant" of a digital image.

$$\mathrm{G}_\sigma \mathrm{v}(l,k) := \sum_{l'=-[4\sigma]}^{[4\sigma]} g_\sigma(l) \sum_{k'=-[4\sigma]}^{[4\sigma]} g_\sigma(k)\, \bar{\mathrm{v}}(l-l', k-k')$$

For a range of values of 'σ' defined in the above formula (i.e. σ ≥ 0.7), the digital Gaussian smoothing parameter satisfying a relation with an error below 10−4 for pixel points varying between 0 and 1[8], 3. Image's key-points are calculated, 4. The location of image's key-points are refined, 5. Key-points without stability (having noise) are filtered, 6. The key-points without stability (laying on edges of the image) are filtered, 7.For each key-point of the digital image, an orientation attribute is assigned, 8. The given input image's key-point descriptor is designed. A key point descriptor is constructed and the matching key point algorithm is applied to it. More sophisticated techniques have been designed allowing matching of image key points[9]. The following is the algorithmic description for the method to be used.

Algorithm 2: Matching Key points

Input: *KA = {(ua, va, Øa, fa)*},
// key points relative to images xA.
KB = {(ub, vb, Øb, fb)}
// key points relative to image xB.
Output: *M = {(ua, va, Øa, fa), (ub, vb, Øb, fb)}*
// matching points.
Arguments: *L*relative, match relative threshold
For each *key point fa in KA do*
Find *fb, fb' //*nearest and 2nd nearest neighbors of *fa*
For each *key point f in KB do*
Compute *d (fa, f)*
End for
f*d (fa ,fb)* <*C*relative,match*d(fa ,fb')* then
Add pair *(fa,fb)* to *M.*
End if
End for

The matching algorithm results are used to judge whether the given input image is cancer affected or not. The accuracy of the results depend on the robustness of the algorithm applied for image comparison.

DISCUSSION

Cancer has been recognized a deadly disease as its diagnosis and cure have been very difficult. Our key idea is to help cancer specialists in better prediction of cancer using the tissue images[1, 2]. This could be made possible by using the relevant histo-pathological images [1]. The discussion led to our work on building an application that could automate the cancer prediction using the tissue images through the image comparison technique against a very large data-base of trained tissue images that leads to the betterment in cancer prediction[6, 8, 9]. So we worked on different algorithms of image processing and image comparison and drafted our work by these methods. The key-point matching process decides the feasibility of output[9]. The probability to the best cancer prediction depends on the size of the data-set. A very large data-set is to be trained using a Machine Learning Algorithm the trine the data-set to compare against the query image and output it's similarity context to any of the cancerous or non-cancerous image present in the data-set and classify accordingly. The application reduces a pathologist's effort in diagnosing the cancer existence in a given tissue image.

CONCLUSION

Multiple image processing steps like image pre-processing to remove noise and increase the resolution of the image to extract the complete image data have been proposed. Pre-processed image is subjected to segmentation to extract the pixels that can be used for further image analysis. The features of the processed image like texture and pathology are extracted through the proposed algorithm. These features are compared with the previously trained image features that are stored in database for knowledge of distinguishing cancerous and non-cancerous tissues. Thus the automated system to diagnose the cancer tissue images is built through the proposed algorithms.

CONFLICT OF INTEREST

Conflict of interest declare none.

REFERENCES

1. Subashini MM, Sahoo SK, Sagar S. Cancer cell diagnostics based on tissue morphology using image processing. InSustainable Energy and Intelligent Systems (SEISCON 2012), IET Chennai 3rd International on 2012 Dec 27 (pp. 1-5). IET.
2. Sill JM, Fear EC. Tissue sensing adaptive radar for breast cancer detection-experimental investigation of simple tumor models. IEEE Transactions on Microwave theory and Techniques. 2005 Nov;53(11):3312-9.
3. Paramkusham S, Rao KM, Rao BP. Early stage detection of breast cancer using novel image processing techniques, Matlab and Labview implementation. InAdvanced Computing Technologies (ICACT), 2013 15th International Conference on 2013 Sep 21 (pp. 1-5). IEEE.
4. Guo Y, Ruan S, Walker P, Feng Y. Prostate cancer segmentation from multiparametric MRI based on fuzzy Bayesian model. In2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI) 2014 Apr 29 (pp. 866-869). IEEE.
5. Meijster, A., and Roerdink, J. B. T. M. A proposal for the implementation of a parallel watershed algorithm. In Computer Analysis of Images and Patterns. Eds., vol. 970 1995, pp. 790-795.
6. Roerdink JB, Meijster A. The watershed transform: Definitions, algorithms and parallelization strategies. Fundamenta informaticae. 2000 Jan 1;41(1, 2):187-228.
7. Otero IR, Delbracio M. Anatomy of the SIFT Method. Image Processing On Line. 2014 Dec 22;4:370-96.
8. Otero IR, Delbracio M. Computing an exact Gaussian scale-space. Image Processing On Line. 2016 Feb 2;6:8-26.
9. Rabin J, Delon J, Gousseau Y. A statistical approach to the matching of local features. SIAM Journal on Imaging Sciences. 2009 Sep 18;2(3):931-58.
10. The University of Arisona, March 26,1998.
11. Tan PH, Cheng L, Srigley JR, Griffiths D, Humphrey PA, Van Der Kwast TH, Montironi R, Wheeler TM, Delahunt B, Egevad L, Epstein JI. International Society of Urological Pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 5: surgical margins. Modern Pathology. 2011 Jan 1;24(1):48-57.
12. Doyle S, Feldman M, Tomaszewski J, Madabhushi A. A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Transactions on Biomedical Engineering. 2012 May;59(5):1205-18.

Reviewers of this article

Dr. S. Prayla shyry, Ph.D

Assistant Professor, Department of CSE, Sathyabama University, Rajiv Gandhi Salai, Jeppiaar Nagar, Chennai, Tamil Nadu 600119, India

Asst.Prof.Dr. Sujata Bhattacharya

Assistant Professor, School of Biological and Environmental Sciences, Shoolini University, Solan (HP)-173212, India

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript