Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.143 – 153

RESEARCH ARTICLE

DETECTION OF SKIN CANCER USING HYBRID OF SVM-ID3 ALGORITHM

Greeshma Rajan[1], G Shivaraj[2]

[1]PG Scholar, Dept. of ECE, SNS College of Engineering, Coimbatore, India

[2]Assistant Professor, Dept. of ECE, SNS College of Engineering, Coimbatore, India

ABSTRACT: This project presents a study of identifying or detecting early symptoms of skin cancer. Realization of automatic cell segmentation technique is presented. A program is developed in MATLAB and effectiveness of the developed program is tested with biomedical sample images. The study reveals that automatic cell detection technique aid cell segmentation with convincing results and increases the speed of analysis. It also provides consistent accuracy in objective cell segmentation. NC ratio plays a vital role in identifying or detecting early disease symptoms such as skin cancers during medical imaging analysis.

Keywords - Cell segmentation, Third Harmonic Generated Microscopy, SVM, ID3, Hybrid of ID3-SVM Algorithm, Nuclear-to-Cytoplasm (NC) ratio

1. INTRODUCTION

Detection of early skin cancer is done by a medical procedure named biopsy [2], which involves the removal of tissues from a living object. The removed tissue is processed by an extensive preparation procedure which includes fixing, embedding, sectioning, staining and placed under a microscope for pathologist examination. Errors may occur during tissue processing and it leads to in accurate diagnosis. It is painful, side effects may also occur like infection and spreading of cancer cells. Optical virtual biopsy techniques for cells and tissues imaging provides capable microscopic details about the benign and malignant lesions without tissue removal. This non-invasive *in vivo* virtual biopsy avoids or minimizes the above mentioned disadvantages involved in virtual biopsy procedure. It also reduces the cost and time consumption in traditional biopsy procedures.

Various non-invasive imaging techniques such as con-focal microscopy [2], two-photon fluorescence (2PF) microscopy, and second harmonic generation (SHG) microscopy have been developed and applied for in vivo human skin diagnosis. Skin disease changes may occur in the deep dermis layer of the skin i.e., several hundred microns below the skin surface. Above mentioned techniques are limited by photo damage, lower resolution, lower penetrability or low contrast. Higher harmonic generation microscopy (HHG) [1], which combines the second and third harmonic generation modalities based on 1230-1250 nm, and can provide high penetration, high resolution and

rich contrast .Second harmonic generation (SHG) light nearly disappeared beyond the depth of 200μm, and the image produced by the SHG also becomes out of focus and lose sharpness at a depth of 250 μm, but still visible at a range of 350μm. THG gives a clear boundary definition between the cell nuclei and cytoplasm [1]. Nuclear-to-Cytoplasm [1] ratio plays a vital role in identifying early symptoms of diseases like skin cancer, whose ratio is generally larger in skin cancer than in normal cells and also, provides information about the type and stage of the disease. Several cell segmentation algorithms have been established for describing the exact position of the round objects and gives information about the size, shape and area to obtain useful properties. Image thresholding [4]-[6] is one of the cell segmentation methods used to segment the objects out of background; it lacks adaptability for global thresholding. This method is computationally expensive and does not consider clustered cells i.e., it cannot separate the touching nuclei. Watershed based- segmentation [4], is a popular morphological image segmentation tool, and often produces over segmentation due to false markers. To reduce over-segmentation fragment merging [6] and marker-controlled watershed transform are used. Fragment merging [6], combines the compactness score and probability density function (PDF) score to obtain more correctly segmented nuclei but, it is sensitive to the size of nuclei. Marker-controlled watershed [7]-[8], replaces the region minimum with predefined markers, each of which represents the object but, the difficulty in performing this method is marker extraction. Convergence index filter [9]-[10], degree of convergence is based on the distribution of the gradient vector not on their magnitude. It is based on the maximization of the convergence index at each point of the spatial co-ordinates. Some of the convergence index filters are COIN filter, Adaptive Ring filter, IRIS filter, Sliding Band filter (SBF). Support region, is the major difference among these filters.

2. IMAGE BACKGROUND

Almost all the parts of our body is completely covered by skin, to precisely diagnose the skin diseases, biopsy is the most common method used today. But, it is invasive, painful to the patient and may also risk patient's life by causing infection or spreading of cancer cells. And, it consumes time for fixing, embedding and staining or pathological analysis. For early diagnosis of skin diseases invasive physical biopsy procedures are removed by non-invasive in vivo virtual biopsy. Confocal microscopy [2], Two photon fluorescence (2PF) microscopy, Second Harmonic Generation (SHG) microscopy, Higher Harmonic Generation (HHG) microscopy, which combines the second and third harmonic generation (THG) microscopy are some of the non-invasive in vivo virtual biopsy imaging techniques.

Gwo Giun et al. (2013) proposed an automatic cell segmentation [1] approach for analysing nuclear-to-cytoplasm (NC) ratio for third-harmonic generated virtual biopsy images. Third-Harmonic Generation (THG) [1] is a nonlinear process, related to the interaction of light with matter, which generates light waves with three times the frequency of the source. It obeys the laws of conservation of energy and involves virtual electron transitions alone. Hence, there is a possibility of less photo damage and photo bleaching and no energy deposition on the interacted matters. THG [3] signals measured on the chicken skin, muscle and fat and stronger THG signals have been found in the skin. Based on that, THG microscopy applied in wide areas, such as the microscopic imaging of plant leaf cells, cultured cells, live zebra-fish embryos and lives mouse skin. THG arises from cell membrane, cytoplasm organelles, hemoglobin, elastic fiber and lipid bodies and it is mainly subscribe by the cytoplasm of the keratinocytes, boundaries of collagen fibers and red blood cells where SHG provides the collagen fibers in the dermis. The resolution of each image is 512 x 512 and is stored in gray-level 12-bit TIFF files, where pixel values are proportional to their third harmonic responses. Fig. 3(a) shows an image from our dataset in which equal amounts of third harmonic responses have been provided in red and blue channels for visualization purposes. The magenta region represents cytoplasm, and the dark part surrounded by magenta cytoplasm represents nuclei. The proposed algorithm for cell segmentation is not limited to these images but can be utilized for analysis of other multivariate images.

3. PROPOSED METHOD

The block diagram of cell segmentation and NC ratio (figure.1) analysis is divided into two parts: nuclei and cytoplasm segmentation. Here the skin cancer can be detected by using an automatic cell segmentation technique, is called Nuclear and Cytoplasm Contours (NCCs) method. NCCs detector is automatically detect the cytoplasm and nucleus contours of a cell in a tissue image. Here an Adaptable Threshold Decision (ATD) method is used to separate the cell from the tissue image and then a maximal gray level gradient difference method is used to extract the nucleus from the cell. From this segmentation result, the NC ratio is evaluated which is important in identifying or detecting the skin cancer. Based on this NC ratio value, here classify the cancerous cells and normal cells in the tissue image.

To improve the accuracy of the cell segmentation, some trained datas are needed, so the classifier algorithm can be used here, which differentiate between the cancerous and normal cells in the tissue image. Here three types of classifier algorithms are used namely Support Vector Machine (SVM), Induction Decision Tree 3 (ID3) and Hybrid of SVM – ID3.Most commonly SVM is used for classification. The main disadvantage of SVM is low speed and it takes more processing time, but it gives the better accuracy. In order to improve the processing time, another classification algorithm ID3 is used. But the accuracy of the ID3 algorithm is low as compared the SVM algorithm. To avoid the disadvantage of both these algorithm, here introduce a new algorithm that is hybrid of SVM and ID3, through which it can be achieved that the higher accuracy in the lesser time And also compared the time and accuracy of these three types of classifier algorithms.

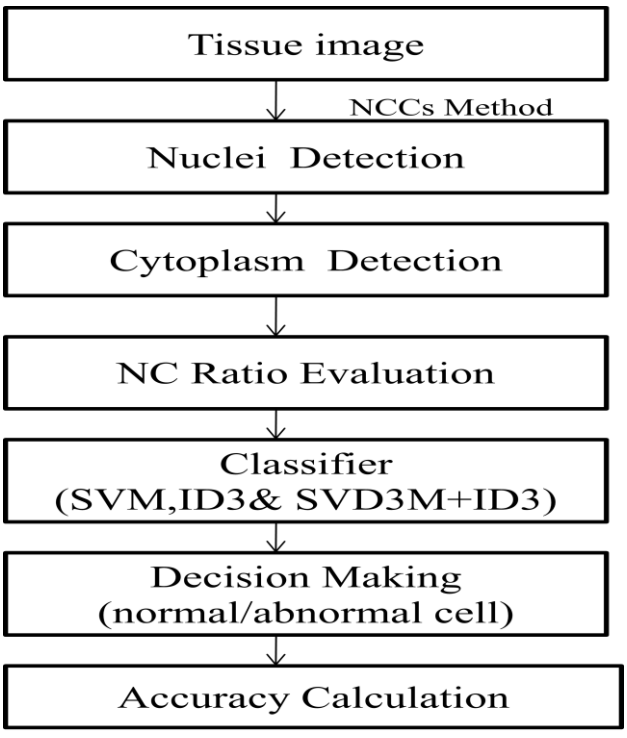

Fig3.1 Block diagram of the cell segmentation and classification

3.1 AUTOMATIC NCC DETECTION

The NCC detector incorporates two phases, cytoplast contour detection phase and nucleus contour detection phase. In the former, the ATD method is used to segment the cytoplast from a tissue image. In general, two different adjacent objects have dissimilar gray-level distributions and the boundary of an object usually has a high gradient. Hence, the maximal gray-level-gradient-difference (MGLGD) method is proposed to draw the nucleus contour based on the gray-level difference between the nucleus and cytoplast in a cell as well as the gradient of the nucleus contour. This section will address both stages in detail.

3.2 NUCLEUS CONTOUR DETECTION PHASE

The nucleus of a cell in a cervical smear image is often much smaller than the cytoplast of the cell and the background of the image, and the gray-level histograms of nuclei are often variable. Often, the ATD method cannot give a good threshold to precisely discriminate the nucleus from the cytoplast of a cell. In this paper, the maximal gray-level-gradient-difference (MGLGD) method is proposed to sever the nucleus from the cytoplast. The nucleus contour detection phase is composed of three stages—gradient calculation, MGLGD, and contour connection.

3.2.1 GRADIENT CALCULATION

Sobel operator is one of the most simple and effective gradient computation methods. The NCC detector uses sobel operator to calculate the gradients of all the cell pixels in I. Let Ig(x,y) be the gradient of the pixel I(x,y) computed by sobel operator. In the sobel operator, where the gray levels of all the cell pixels in gray scale images are stretched from 0 to 255. Sobel operator is sensitive to noise, hence, the NCC detector employs the Mean Vector Difference enhancer to highlight the gradient of object contour and to suppress the gradient of noise contour.

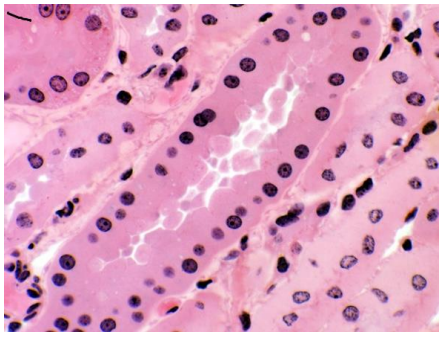

Figure .1 Original tissue image

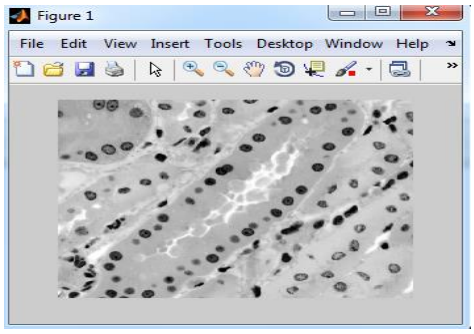

Figure .2 Gray scale image

3.2.2 MAXIMAL GRAY-LEVEL-GRADIENT-DIFFERENCE (MGLGD) METHOD

In general, two different adjacent objects have dissimilar gray level distributions. Given an image region R that contains only two objects A and B, such as considering the image, an initial contour C0 (marked by red) partitions R into two sub-regions. Each of the pixels on C0 will be repeatedly moved. When C0 is moved to the contour of B, the difference of the average gray-levels of A and B is maximal. Moreover, the pixels located on the object contour have great gradients for the most part. Based on these two properties, we propose the maximal gray-level-gradient difference (MGLGD) method to draw the contour of the nucleus.

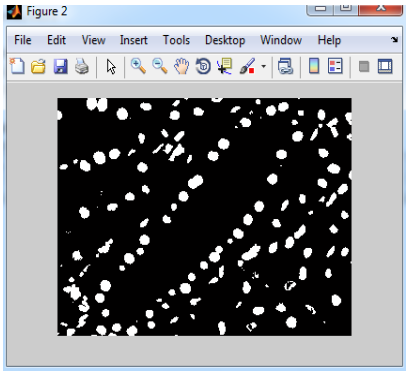

Figure .3 Detection of nuclei

i. CONTOUR CONNECTION

However, in each iteration, the new Cj may not be a closed curve since the solutions of (x0 ,y0), are discrete values. each iteration, the MGLGD method must mend the breaking C0j.Let Wr be a 5 x 5 related window of I(x0 ,y0), where I(x0 ,y0) is located at the center of Wr.

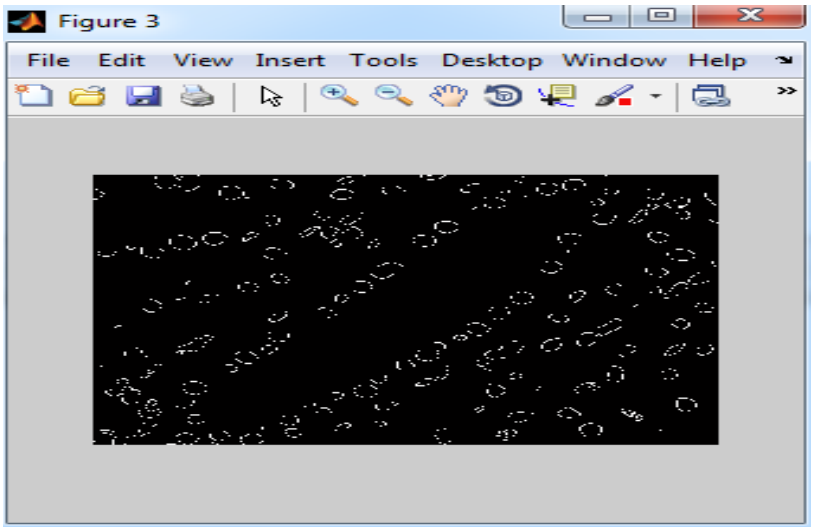

Figure .4 Nuclear contours

We number the pixels in Wr,. An endpoint is one of the two ends of a line. Let I(x0 ,y0) be an endpoint on C0 j and I(x0 ,y0) be the pixel numbered 1 in Wr. Let the other endpoint be the pixel numbered k in Wr. If k = 3, 5, 7, or 9, then the (k x 1)th pixel is added to C0j; otherwise, the (4 x (1 + (k _ 9)div4) + 9)th pixel is assigned to an element of Cj, where div is the integer division.

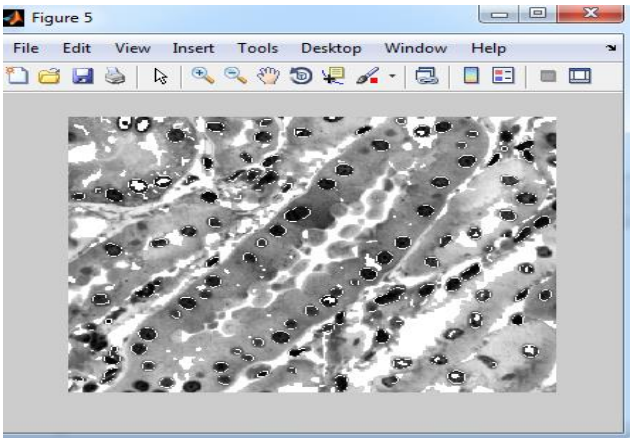

Figure .5 Nuclei contour detection

3.3 CYTOPLAST CONTOUR DETECTION PHASE

The thresholding technique is commonly used to separate objects from an image via some threshold values. Otsu's method (Otsu, 1979) is a well-known thresholding method since it is simple and effective. Unfortunately, when the standard deviations or the numbers of data between classes are quite different, Otsu's method cannot successfully give a proper threshold (Hou, Hu, &Nowinski, 2006) . To solve the problem of Otsu's method, Tsai et al. (2010) proposed the ATD method, which uses the standard deviations of group, quantity of data, and group interval as the factors of determining the optimal thresholds. In this paper, the ATD method is used to draw the contour of the cytoplast. The NCC detector uses the ATD method to divide the cytoplast and background of the image. Let I be a cervical smear image and I(x,y) the pixel located at the coordinates (x,y) on I. Assume T⁄ is the threshold obtained by the ATD method from the gray-level histogram of I. Then, the NCC detector generates Ib by Formula,

$$I_b(x, y) = \begin{cases} 0, & if\ I(x,y) >= T^* \\ 1, & otherwise, \end{cases} \qquad 1$$

where I_b is a binary image and a 1-bit (resp. 0-bit) denotes a black pixel (resp. a white pixel). Figs(a) and (b) are the I_b and the cytoplast contour obtained from the original image in Fig. 1(a) by the NCC detector. We call I(x,y) a cell pixel only if $I_b(x,y) = 1$, where all the cell pixels comprise a cell region.

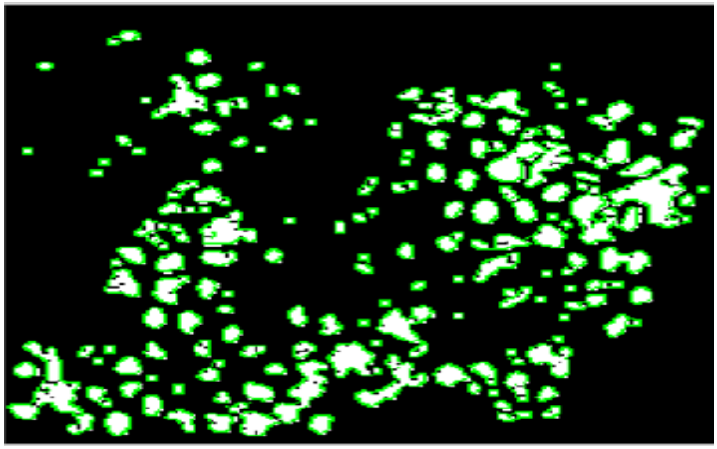

Figure .6 Cytoplasm contour detection

3.4 CELL SIZE AND NC RATIO EVALUATION

Cellular size and nuclear size are indicators not only of the developing status of some diseases but of skin and other quantifiable physical factors. For example, cellular and nuclear size in the layers of basale cells in forearm skin has been found to increase with age. The NC ratio, which has been defined as the volume ratio of nucleus to cytoplasm, is commonly used in diagnosis. A protocol has been developed to obtain accurate NC ratios. Although the NC ratio is defined as a volume ratio, it can be approximated by an area ratio of nucleus to cytoplasm.

3.5 SUPPORT VECTOR MACHINES (SVMS)

SVMs are relatively new types of classification algorithms. An SVM expects a training data set with positive and negative classes as an input (i.e. a binary labeled training data set). It then creates a decision boundary (the maximal-margin separating boundary) between the two classes and selects the most relevant examples involved in the decision process (the so-called support vectors). The construction of the linear boundary is always possible as long as the data is linearly separable. If this is not the case, SVMs can use kernels, which provide a nonlinear mapping to a higher dimensional feature space. The dot product has the following formula:

$$k(x,y) = k(x \cdot y + 1)^d \quad (2)$$

where *x* and *y* are the vectors of the gene expression data. The parameter *d* is an integer which decides the rough shape of a separator. In the case where *d* is equals to 1, a linear classification algorithm is generated, and in the case where *d* is more than 1, a nonlinear classification algorithm is generated. In this paper, when *d* is equals to 1, it is called the SVM dot product, when *d* is equals to 2, it is called the SVM quadratic dot product and when *d* is equals to 3, it is called the SVM cubic dot product. The radial basis kernel is as follows,

$$k(x,y) = \exp[-(x-y)^2/2\sigma^2] \quad (3)$$

where σ is the median of the Euclidean distances between the members and non-members of the class. The main advantages of SVMs are that they are robust to outliers, converge quickly, and find the optimal decision boundary if the data is separable [7]. Another advantage is that the input space can be mapped into an arbitrary high dimensional working space where the linear decision boundary can be drawn. This mapping allows for higher order interactions between the examples and can also find correlations between examples. SVMs are also very flexible as they allow for a big variety of kernel functions. Sequential minimal optimization (SMO) [20] is used in this paper to train an SVM. SVMs have been shown to work well for high dimensional microarray data sets [10]. However, due

to the high computational cost it is not very practical to use the wrapper method to select genes for SVMs, as will be shown in our experimental results section.

3.5 ITERATIVE DICHOTOMISER 3

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan[1] used to generate a decision tree from a dataset. The ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy or information gain of that attribute. Then selects the attribute which has the smallest entropy (or largest information gain) value. The set is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before. Recursion on a subset may stop in one of these cases: Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labeled with the class of the examples ,there are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labeled with the most common class of the examples in the subset, there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected . the data was split, and terminal nodes representing the class label of the final subset of this branch. Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which

3.6 PROPOSED ALGORITHM

The proposed method is a hybrid approach to embedding SVM in ID3 (SVM-ID3) for pre-pruning the tree while carrying out the classification. This resulting hybrid system is categorized as embedded hybrid system where the technologies participating are integrated in such a manner that they appear to be inter-twined. The proposed model is similar to the classical recursive partitioning schemes, except that the leaf nodes created are Support Vector Machine categorizers instead of nodes predicting a single class. The SVM classifier has been used for pre-pruning the DT resulting in a smaller DT than a complete on application of C4.5. The proposed model uses the C4.5 algorithm for constructing a decision tree. Root node of the decision tree is selected based on a chosen threshold value of the continuous attribute. For this the standard entropy minimization technique is used. In the next step the Significance of Node is computed by using 10x10 cross-validation accuracy estimates for SVM at the node. Computation of Significance of Node is followed by the computation of Significance of Split. The Significance of Split is computed by taking the weighted sum of the significance of the nodes. Here, the weight given to a node is proportional to the number of instances that go down to that node. Significance of Node and Significance of Split are computed and compared and the results attempt to approximate whether the generalization accuracy for SVM classifier at each leaf is higher than a single SVM classifier at the current node. A split is defined to significant if the relative (not absolute) reduction in error is greater than 5% and there are at least 20 instances in the node. If there are n training samples, and m attributes, then the computational complexity of the algorithm for the proposed model has been worked out to be O(m.n2).

4. EXPERIMENTAL RESULT

Automatic cell segmentation and NC ratio evaluation were performed using the proposed algorithm on about 600 THG virtual biopsy images of the stratum basale layer of human forearm skin from 31 healthy volunteers. The evaluated NC ratios and cell sizes were discussed and interpreted by a dermatologist and a radiologist. From the Figure 5.4(a), it can be observed that one of the experimental results of cell segmentation is given. Its profile, including the evaluated NC ratios, cell sizes, and position is given in Table 5.1. Most of the cells were segmented accurately. From the table 5.1 it can be observed that the experimental results of cell segmentation including the evaluated NC ratios, cell sizes and also found that the range of NC ratio lies between 0.29-0.51 with average 0.326.

SI NO	Cell Area (pixels)	Nuclear Area (pixels)	Cytoplasmic Area(pixels)	NC Ratio
1	924	229	695	0.329496403
2	649	160	489	0.327198364
3	901	228	673	0.338781575
4	1076	316	760	0.415789474
5	836	201	635	0.316535433
….	….	….	…..	….
149	862	222	640	0.326875
150	972	319	653	0.488514548
151	943	296	647	0.4574970297
152	722	217	505	0.32970297
153	1226	303	923	0.328277356
154	1221	408	813	0.501845018
155	1357	400	957	0.417972832
156	1341	452	889	0.508436445
157	1323	195	1128	0.28672234
158	1066	265	801	0.330836454
Total	56250	15732	40518	-
Average	969.8276 (pixels)	271.2414 (pixels)	698.5862 (pixels)	0.325971879

Table .1 calculation of NC Ratio

From the table 1 it can be observed that the experimental results of cell segmentation including the evaluated NC ratios, cell sizes and also found that the range of NC ratio lies between 0.29-0.51 with average 0.326. .Based on this NC ratio value, classified the normal and abnormal cells in the tissue image. To improve the accuracy of the cell segmentation here classifier algorithms are used and comparison can be listed in the table 2.

SI.NO	Name of the algorithm	Accuracy (%)	Time(second)
1	ID3	95	2.36
2	SVM	97	5.62
3	Hybrid of ID3-SVM	98	1.08

Table 2 Comparison of the 3 algorithms

From the table 2,it can be observed that the id3 have low accuracy and time consumption as compared to the SVM algorithm. Then the SVM requires high accuracy and it takes more the processing time as compared to the ID3. But the hybrid of the ID3 and SVM provides better speed and accuracy within the less time. Based on the results, the estimated NC ratios can be used to assist medical doctors to noninvasively identified the symptoms of diseases with abnormal NC ratios in addition to revealing the type and stage of the developing disease.

5. CONCLUSION

This method presented a computer-aided design for automatic cell segmentation and NC ratio analysis. Realization of automatic cell segmentation has been studied using a watershed algorithm. Here the nuclei and cytoplasm are segmented from tissue image and then found the NC ratio from it. A MATLAB programming has been developed for the implementation of this method .NC ratio value is achieved by this proposed method is 0.29-0.51 and mean and standard deviation are 0.33, 0.01 respectively. The proposed method saves much time and provides convincing segmentation result. Processing time can be reduced by using this technique.

Acknowledgements

The authors would like to thank a great support of Anna University and SNS Group of Institutions to complete this project successfully.

REFERENCES

[1] Gwo Giun (Chris) Lee, Senior Member, IEEE, Huan-Hsiang Lin, Ming-Rung Tsai, Sin-Yo Chou, Wen-Jeng Lee, Yi-Hua Liao, Chi-Kuang Sun, Fellow, IEEE, and Chun-Fu Chen, ' Automatic Cell Segmentation and Nuclear-to-Cytoplasmic Ratio Analysis for Third Harmonic Generated Microscopy Medical Images', IEEE Transactions On Biomedical Circuits And Systems, Vol. 7,pp. 158-168, 2013.

[2] S.-Y. Chen, H.-Y. Wu and C.-K. Sun, 'In vivo harmonic generation biopsy of human skin', J. Biomed. Opt., vol. 14, no. 6, p. 060505, 2009.

[3] S.-Y. Chen, S.-U. Chen, H.-Y. Wu, W.-J. Lee, Y.-H. Liao and C.-K. Sun, 'In vivo virtual biopsy of human skin by using non invasive higher harmonic generation microscopy', IEEE J. Sel. Topics Quantum Electron., vol. 16, no. 3, pp. 478–492, 2010.

[4] K. Z. Mao, Peng Zhao, and Puay-Hoon Tan, 'Supervised Learning-Based Cell Image Segmentation for P53 Immuno histo chemistry', IEEE Transactions On Biomedical Engineering, Vol. 53, pp. 1153-1163, 2006.

[5] Bin Fang, Wynne Hsu, and Mong Li Lee, 'On the Accurate Counting of Tumor Cells', IEEE Transactions On Nanobioscience, Vol. 2, pp.94-103,2003.

[6] Xiaobo Zhou, Fuhai Li, Jun Yan, and Stephen T. C. Wong, 'A Novel Cell Segmentation Method and Cell Phase Identification Using Markov Model', IEEE Transactions On Information Technology In Biomedicine, Vol. 13, pp.152-157, 2009.

[7] Chanho Jung and Changick Kim, 'Segmenting Clustered Nuclei Using *H*-minima Transform-Based Marker Extraction and Contour Parameterization', IEEE Transactions On Biomedical Engineering, Vol. 57, pp.2600-2604, 2010.

[8] Xiaodong Yang, Houqiang Li, and XiaoboZhou, 'Nuclei Segmentation Using Marker-Controlled Watershed, Tracking Using Mean-Shift, and Kalman Filter in Time-Lapse Microscopy', IEEE Transactions On Circuits And Systems Vol. 53, pp. 2405-2414, 2006.

[9] Hidefumi Kobatake, and Shigeru Hashimoto, 'Convergence Index Filter for Vector Fields', IEEE Transactions On Image Processing, Vol. 8, pp.1029-1038, 1999.

[10] Pedro Quelhas, Monica Marcuzzo, Ana Maria Mendonça and Aurélio Campilho, 'Cell Nuclei and Cytoplasm Joint Segmentation Using the Sliding Band Filter', IEEE Transactions On Medical Imaging, Vol. 29, pp. 1463-1473, 2010.

[11] Hyejun Ra, Wibool Piyawattanametha, Yoshihiro Taguchi, Daesung Lee, Michael J. Mandella, 'Two-Dimensional MEMS Scanner for Dual-Axes confocal Microscopy', Journal of microelectromechanical systems, Vol. 16, No.4, 2007.

[12] Kung-Bin Sung, Chen Liang, Michael Descour, Tom Collier, Michele Follen, and Rebecca Richards-Kortum, 'Fiber-Optic Confocal Reflectance Microscope With Miniature Objective for *In Vivo* Imaging of human skin tissues', IEEE Transaction on Biomedical Engineering, Vol. 49, pp. 1168- 1172, 2002.

[13] Terry B. Huff,Yunzhou Shi, Yan Fu, Haifeng Wang and Ji-Xin Cheng, 'Multi-modal Nonlinear Optical Microscopy and Applications to Central Nervous System Imaging', IEEE journal on Quantum Electronics, Vol. 14, pp.4-9, 2008.

[14] Paul J. Campagnola, Heather A.Clark, William A. Mohler, Aaron Lewis, Leslie M. Loew, 'Second-harmonic imaging microscopy of living cells', Journal of Biomedical Optics, 2001.

