Classification of Cervix Tumor using Bag of Visual Word Classifier

Hema Rajini N

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamilnadu, India.

Abstract— *A tumor segmentation and classification system has been designed and developed on computed tomography images. Image processing is used in the medical field for detection of tumor. Image segmentation is an important part of image processing. Segmentation is the process of subdividing an image into distinct regions. The algorithm has the steps of preprocessing, cervix extraction, cervix boundary correction, image segmentation, feature extraction and image classification. The image is preprocessed using adaptive median filtering and fuzzy thresholding. The cervix is extracted using canny edge detection and border tracing algorithm. The cervix boundary correction is performed using adaptive concave hull algorithm. Segmentation is performed using region growing based technique. Then for the segmented tumor region, the features are extracted using the gray level co-occurrence matrix. After the features are extracted, the image is classified as the benign or malignant cervix by using the bag of visual word classifier.*

Keywords— Segmentation, Region growing, Feature extraction, Gray level co-occurrence matrix, Bag of visual word.

I. INTRODUCTION

Segmentation is a vital process in digital image processing that has found extensive applications in several areas. Image segmentation plays an important role in the field of biomedical applications. Cervix cancer is the fourth most common cancer among women globally and the second most common cancer among Indian women. India alone bears 23% of the global cervical cancer burden. The cervix cancers have the ability to spread to other parts of the body that leads to death. The segmentation technique is widely used by the radiologists to segment the input medical image into meaningful regions. The specific application of this technique is to detect the tumor region by segmenting the abnormal Computed Tomography (CT) cervix tumor image.

The medical images can be segmented manually or automatically. But the accuracy of image segmentation using the segmentation algorithm is more when compared with the manual segmentation. Large amount of time was spent by radiologist and doctors for identification of tumor and segmenting it from other cervix tissues. However, exact labeling cervix tumors is a time-consuming task, and considerable variation is observed between doctors. Subsequently, over the last decade, from various research results it is being observed that it is very time-consuming method but it will be performed faster using image processing techniques. Primary cervix tumors do not spread to other body parts and can be malignant or benign and secondary cervix tumors are always malignant. Malignant tumor is more dangerous and life threatening than benign tumor. The benign tumor is easier to identify than the malignant tumor. Also, the first stage tumor may be malignant or benign but after first stage it will change to dangerous malignant tumor which is life threatening.

A tumor is an abnormal growth of cells that serves no purpose. A benign tumor is not a malignant tumor. It does not invade nearby tissue or spread to other parts of the body the way cancer can. In most cases, the outlook with benign tumors is very good. But benign tumors can be serious if they press on vital structures such as blood vessels or nerves. Therefore, sometimes they require treatment and sometimes they do not require treatment.

II. LITERATURE SURVEY

Armato et al. has proposed an automated lung segmentation technique for thoracic CT image [1]. Here the information is reliable. This method uses gray-level thresholding to segment the lungs within each computed tomography section. But the segmentation is inaccurate 5% - 17% lung nodules are missed. Jiantao Pu et al. proposed an adaptive border marching algorithm for automatic lung segmentation on chest CT images [2]. This method is more efficient and straight forward to implement. But the juxtaplueral nodule is missed while minimizing over segmentation of adjacent areas. Gomathi et al. proposed an image segmentation approach using Modified Fuzzy C-Means (FCM) algorithm and Fuzzy Possibilistic c-means algorithm (FPCM) [3]. This approach is a generalized version of standard FCM Clustering algorithm. The limitation of the conventional FCM technique is eliminated in modifying the standard technique. The Modified FCM algorithm is formulated by modifying the distance measurement of the standard FCM algorithm to permit the

labeling of a pixel to be influenced by other pixels and to restrain the noise effect during segmentation. The FCM is more robust in the presence of noise and the performance is good. But fail to segment the images corrupted by noise. Standard FCM, modified FCM, FPCM are compared to explore the accuracy of this proposed approach.

Hua et al. proposed an automatic algorithm for pathological lung CT image segmentation that uses a graph search algorithm [4]. This method is more efficient and mostly used in 3D images. But the time complexity is high. Ying Wei et al. developed the fully automatic lung segmentation and repairing based on improved chain code and Bresenham algorithm [5]. The segmentation accuracy and implementation speed have been improved greatly than the rolling ball method. It requires Low computational cost. The performance is good. It reduces the missing of juxtaplueral nodules. But the complexity is high. Shiying Hu et al. developed fully automatic method for identifying the lungs in 3D pulmonary X-ray images [6]. It is a simple method which does not require the user interaction and time complexity is reduced. But the optimal size selection is difficult because the algorithm depends on the size and shape of the selected morphological structuring elements.

In this proposed method the image is preprocessed using adaptive median filtering and fuzzy thresholding. The cervix is extracted using canny edge detection and border tracing algorithm. The cervix boundary correction is performed using adaptive concave hull algorithm. Segmentation is performed using region growing based technique. Then for the segmented tumor region, the features are extracted using the Gray Level Co-occurrence Matrix (GLCM). After the features extracted, the image is classified as the benign or malignant cervix by using the Bag of Visual Word (BOVW) classifier.

The rest of this paper is organized as follows. The introduction part is given in Section 1 and the studies of several research papers are portrayed in the Section 2. Section 3 describes the proposed methodology following Section 4 signifies the experimental results and the work are concluded in Section 5.

III. METHODOLOGY

The main purpose of this work is to identify the region of tumor and to do the detailed diagnosis of that tumor; this will be helpful for treating the cancer patient. The proposed methodology is shown in Fig.1. The cervix segmentation algorithm is introduced to segment the cervix and correct the segmentation defects caused by different types of nodules.

1.1 Pre-processing

Before cervix segmentation, some pre- processing techniques are applied on the CT images to produce appropriate images for the next steps. First, image enhancement and noise removal need to be performed by adaptive median filtering of the CT images. The adaptive median filter detects noisy pixels by comparing them to their neighbourhood [7]. Both the size of the neighborhood and the threshold for the comparison are adjustable. A pixel which exceeds this threshold and is also structurally misaligned with similar pixels in its neighbourhood is classified as noise. In that case, the noisy pixel is replaced by the median pixel value of the non-noise pixels in its neighbourhood. After noise removal, a fuzzy thresholding method is used for CT image binarization this is a robust binarization method and helps to remove all the unrelated parts in the cervix areas the goal being keeping only the important areas that are helpful for cervix segmentation [8].

1.2 Cervix extraction

Following the pre-processing step, the canny edge detector is used to specify the borders of the connected components [9]. Then, the well-known border tracing algorithm is used to compute the cervix border sequential points; forming a collection of directed closed contours. The extra contours can be removed based on the number of pixels in each contour. In the next step, a novel adaptive concave hull algorithm is applied to obtain accurate initial cervix masks for an active contour model. Angles and curvature are probably the most widely used features for concavity calculation but both of them are vulnerable to noise. In this method, an adaptive concavity degree computation of each boundary pixel is applied for each point on the cervix boundary, its corresponding boundary points by distance h are considered as the two endpoints of its line segment. The concavity degree is measured as the ratio of the outside points of the line segment to the total number of the line segment points, as shown in equation (1):

$$C_i^h = \frac{\sum Li \cap R'}{(2h-1)} \quad (1)$$

Where, L_i is the region of the line segment, C_i is the concavity degree corresponding to the I^{th} boundary pixel, R is the region of the cervix image and R' is its complementary region. Furthermore operator $\sum$ shows the number of non-zero values in the corresponding operand. Thus, the total number of the line segment points is equal to (2h – 1) because two end points are not considered as the line segment points during computation. In addition, the other innovation in the adaptive concavity degree computation is that instead of choosing a single value for parameter "h" multiple values are selected for this parameter.

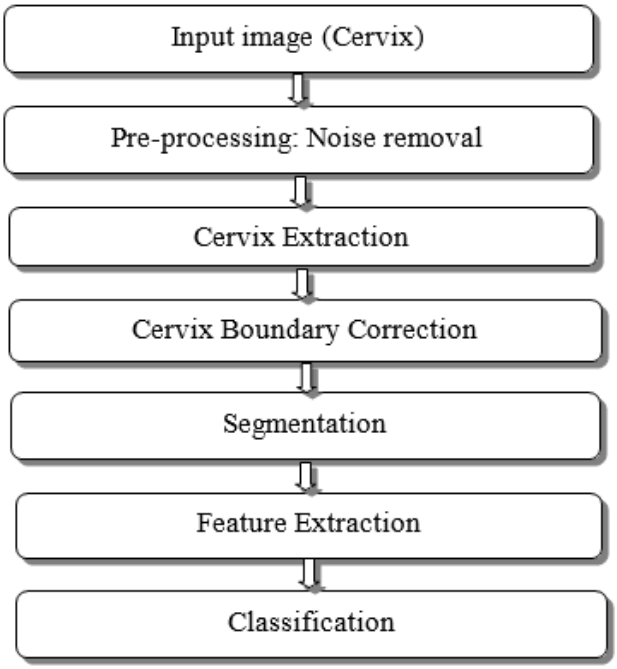

Fig. 1: Methodology of the Proposed Technique

Tumor nodules can be classified into three groups based on their diameter as small (diameter ≤ 5), medium (5 ≤ diameter ≤ 8) and large (8 ≥ diameter). Thus, three values are considered for 'h' corresponding to these three categories that help us to have a robust algorithm that can include different kind of tumor or nodules in this cervix segmentation method. In the concave image there is lower intensity on the border points that are in non-convex parts while for the other parts of the contours have high intensity.

1.3, Cervix boundary correction

In this step the boundary points of the concave image are traced and a novel method is used to smooth the cervix borders by closing the holes on the borders that are caused by tumor or nodules in order to include them in the segmented cervix. Each of these cervix contours in the concave image can be characterized with a histogram of the measured concavity degree of the sequence of points that is referred to as concavity histogram. Since this histogram is noisy a Gaussian filter is used for denoising to remove false local maxima and produce a robust segmentation method. Finally find the pair of strong local maximums in the concavity histogram that have strong deep valley between them. These points indicate holes on the cervix border because small perturbations in shape might cause a rapid change of curvature value in the histogram. If their Euclidean distance between these points is lower than a pre-defined threshold these two points indicate the two side of the tumor or nodule otherwise this curvature is a natural border along the cervix like the hole on the cervix boundary that is due to the space between cervixes. Thus, the nodule curvature parts should be closed but the main borders should be kept.

1.4, Segmentation

In the last step the result of the previous step is applied as an initial mask of a region growing. This iterative process will deform the model by finding the cervix. The region growing is useful for segmenting the borders of cervix accurately and to exclude some tiny white parts that exist in the cervix areas following the result of previous step and decreasing the effects of over segmentation [10]. Thus, this approach produces robust cervix segmentation.

E. Feature extraction

Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problem stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Feature extraction is done by using GLCM. The GLCM represent spatial distribution and gray levels dependency within a local area. There are many texture features available, but here only four features are used. They are Energy, Contrast, Correlation, and Homogeneity. The features which are extracted in this work are given in eqs. (2)-(5).

1. Energy

$$Energy = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j)^2 \qquad (2)$$

The energy of a texture describes the uniformity of the texture. Energy is 1 for a constant image.

2. Contrast

$$Contrast = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G} \sum_{j=1}^{G} P(i,j) \right\}, |i-j| = n \qquad (3)$$

Contrast is a measure of the local variations present in an image. This measure of contrast favors contributions from *P (i, j)* away from the diagonal, i.e. *i= j*. If there is a large amount of variations in an image, the *P[i,j]*'s will be concentrated away from the main diagonal and the contrast will have a high value.

3. Correlation

$$Correlation = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \quad (4)$$

Correlation is a measure of gray level linear dependence between the pixels at the specified positions relative to each other. The correlation will be higher if an image contains a considerable amount of linear structure.

4. Homogeneity

$$Homogeneity = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{P(i,j)}{1+|i-j|} \quad (5)$$

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Homogeneity is 1 for a diagonal GLCM. A homogeneous image will result in a co-occurrence matrix with a combination of high and low *P [i,j]*'s. A heterogeneous image will result in an even spread of *P [i,j]*'s.

F. Classification

The Bag of Visual Word technique is also called as bag of words. Visual image categorization is a process of assigning a category label to an image under test. Categories of image for this project will be benign or malignant images. This method contains following steps.

Step 1: Load Image Sets: Cancer image dataset contains two categories: benign and malignant, which can be loaded for classification.

Step 2: Prepare training and validation image Sets: to equalize the images in each and every category.

Step 3: Create a visual vocabulary and train an image category classifier:

- Extracts SURF (Speed up Robust Features) features from all images in all image categories.
- Constructs the visual vocabulary by reducing the number of features through quantization of feature space using K-means Clustering.

Step 4: Evaluate Classifier Performance: First test it with the training set, which should produce near perfect confusion matrix.

Step 5: Try the newly trained classifier on test images: then apply the newly trained classifier to categorize new images.

Step 6: Accuracy and confusion matrix will be displayed.

IV. RESULTS AND DISCUSSION

In the experimental setup input images are extracted. During testing normal and abnormal cervix images are taken for analysis. The images are taken from that of CT scan images. The test image can be of any size. The input images are preprocessed. Then the enhanced images are segmented. The dense tissue regions within the tumors are segmented. Finally, the image is classified as normal or an abnormal cervix.

The CT image is collected from International Cancer Center (ICC), Neyyoor. The image consists of both benign and malignant cervix image. The benign cervix images are shown in Fig. 2. The malignant cervix images are shown in Fig. 3.

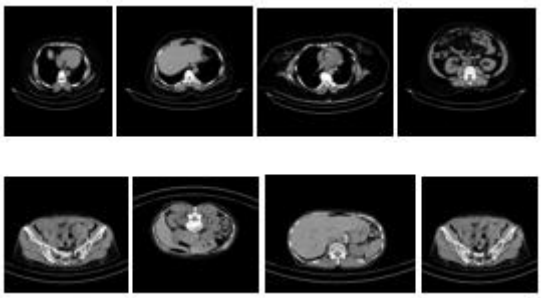

Fig. 2: Benign cervix images

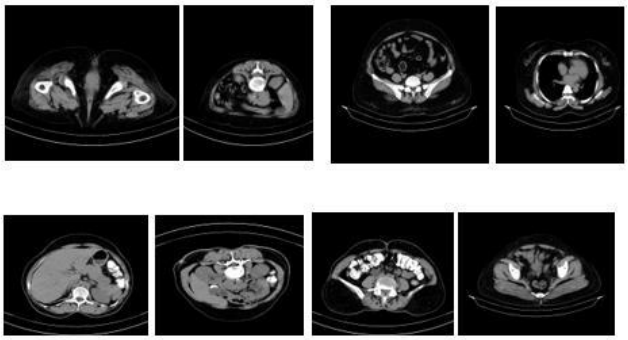

Fig. 3: Malignant cervix images

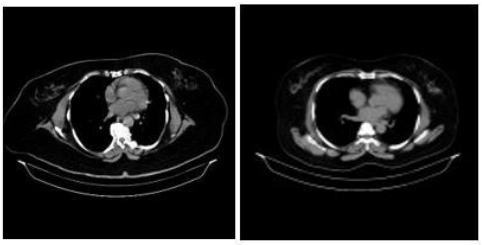

Fig. 4: Input Images

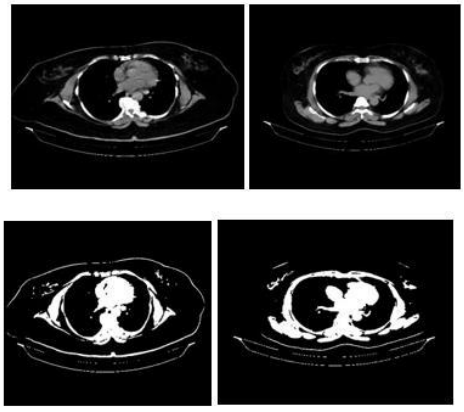

Fig. 5: Preprocessing result

The input images are filtered to remove noise and

filtering is done using the adaptive median filter and fuzzy thresholding. Preprocessing result for the corresponding input image is shown in Fig. 5.

Cervix extraction and cervix border correction also done for identifying the region of interest. Its results are shown in Fig. 6.

Fig. 6: Cervix extraction and cervix border correction result

The preprocessed output is segmented to detect the tumor region. Segmentation is done using the region growing algorithm. The output of segmentation is shown in Fig. 7.

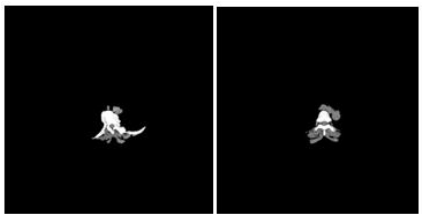

Fig. 7: Segmentation Result

From the segmented region the features are extracted. Features are extracted using GLCM.

Table 1: Feature Extraction Result

GLCM Features	Benign			Malignant		
	1	2	3	1	2	3
Contrast	0.040	0.036	0.025	0.023	0.041	0.031
Correlation	0.961	0.970	0.971	0.966	0.967	0.972
Energy	0.964	0.9409	0.9637	0.9642	0.9476	0.9547
Homogeneity	0.996	0.9944	0.9962	0.9960	0.9944	0.9959

It represents spatial distribution and gray levels dependency within a local area. There are many texture features available, but here only four features are used. They are Energy, Contrast, Correlation, and Homogeneity. The extracted feature values are shown in table 1.

Features are extracted for the segmented region and based on this feature values, the input image is classified as benign or malignant cervix by using the BOVW classifier.

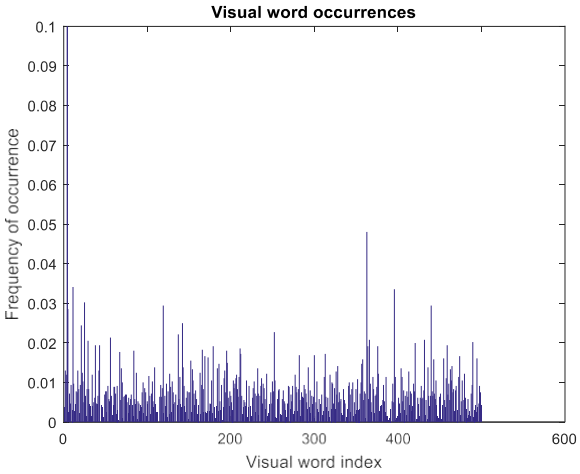

Fig. 8: Visual vocabulary BOVW classifier

This is the extension of NLP algorithm. BOVW is the supervised learning model. Use the computer vision system toolbox functions for image category classification by creating a bag of visual words. The process generates a histogram of visual word occurrences that represent an image. This histogram is used to train an image category classifier.

Fig. 8 and 9 shows that the bag of visual words classification results in which the accuracy is 92.8%. Before classification the image of features is extracted and then confusion matrix is created for this method.

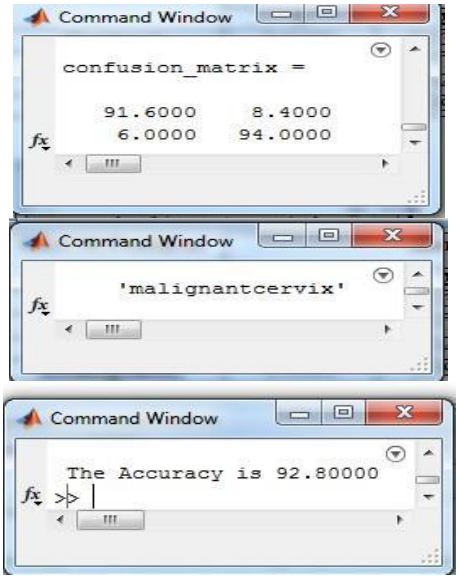

Fig. 9: Classifier result

V. CONCLUSION

The proposed system is used to detect and identify the cervix tumor by segmenting cervix CT images. This method has the stages of preprocessing, cervix extraction, cervix boundary correction, image segmentation, feature extraction and image classification. The image is preprocessed using adaptive median filtering and fuzzy thresholding. After this process fatty tissues and other unwanted details get smoothened. The cervix is extracted using canny edge detection and border tracing algorithm. The cervix boundary correction is performed using adaptive concave hull algorithm. Segmentation is performed using region growing based technique. Then from the segmented tumor region, the features are extracted using the GLCM. After the features are extracted, the image is classified as the benign or malignant cervix by using the bag of visual word classifier. The area of tumor and its type of tumor is found using this method. A classification with an accuracy of 92% has been obtained by bag of visual word classifier.

REFERENCES

[1] Armato, S.G., & Sensakovic, W.F., (2004). Automated lung segmentation for thoracic CT impact on computer-aided diagnosis. Academic Radiology 11(9), 1011-1021.

[2] Jiantao Pu, Justus Roos, Chin A Yi, Sandy Napel, Geoffrey D. Rubin & David S. Paik, (2008). Adaptive border marching algorithm: automatic lung segmentation on chest CT images. Computerized Medical Imaging and Graphics 32(6), 452-462.

[3] Gomathi, M. & Thangaraj, P., (2010). A New Approach to Lung Image Segmentation using Fuzzy Possibilistic C-Means Algorithm. International Journal of Computer Science and Information Security 7(3), 222- 228.

[4] Hua, P., Song, Q., Sonka, M., Hoffman, E. A. & Reinhardt, J. M., (2011). Segmentation of pathological and diseased lung tissue in CT images using a graph-search algorithm. IEEE International Symposium on Biomedical Imaging: From Nano to Macro 2072-2075.

[5] Ying Wei, Guo Shen & Juan-Juan Li, (2013). A fully automatic method for lung parenchyma segmentation and repairing. Journal of Digital Imaging 26(3), 483-495.

[6] Shiying Hu, Eric A. Hoffman & Joseph M. Reinhardt, (2001). Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Transactions on Medical Imaging 20(6), 490-498.

[7] Hwang, H. & Haddad, R.A., (1995). Adaptive median filters: new algorithms and results. IEEE Transactions on Image Processing 4(4), 499- 502.

[8] Jiahui Wang, Feng Li & Qiang Li, (2009). Automated segmentation of lungs with severe interstitial lung disease in CT. Medical Physics 36(10), 4592-4599.

[9] Xu T, Mandal M, Long R, Cheng I & Basu A, (2012). An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. Computerized Medical Imaging and Graphics 36(6), 452-463.

[10] Withey, D.J. & Koles, J.H., (2007). Three Generations of Medical Image Segmentation: Methods and available Software. International Journal of Bioelectromagnetism 9(2), 67-68.