Improved Method for Watershed Segmentation of Lung CT Scan Images for Early Diagnosis of Cancer

Shubhpreet Kaur and Gagandeep Jindal

Abstract—The purpose of this research work is to segment human lung CT scan images for early detection of cancer. As a part of this work ‘Watershed Technique’ is implemented as the ‘Segmentation’ method. This method is based on the filters available in the ‘Insight Segmentation and Registration Toolkit’ (ITK) [9]. Image segmentation plays a vital role in several medical imaging programs by automating or assisting the delineation of physiological structures along with other parts of interest.

NIH/NCI Lung Image Database Consortium (LIDC) served as the repository of lung images in the DICOM format. In this short article, different methods that can be used for efficient visualization as well as automatically extracting the organ regions from abdominal CT (Computerized tomography) data especially from lung that can be further used in various medical diagnosis applications like CBMIR (Content-based medical image retrieval) have been suggested. The achieved results support the use of the proposed algorithm for extraction of lung nodules from the image.

Keywords—CAD, DICOM, CBMIR, ITK, MATITK, Computerized Tomography, LIDC.

I. INTRODUCTION

In image processing, segmentation falls in to the category of extracting different image attributes of an original image. Segmentation subdivides an image into constituent regions or objects. The level to which that subdivision carried out is a problem specific.

The purpose of the segmentation of the lung region in the CT image is to achieve a better orientation in the image. For separating the regions of interest or objects of interest from other parts of the image a region growing approach is used to distinguish between the specific nodules of lungs and other suspicious region. Region growing [2] is a technique for extracting a region of the image that is connected based on some predefined criteria. This criterion can be based on intensity information and/or edges in the image. In its simplest form, region growing requires a seed point that is manually selected by an operator, and extracts all pixels [6] connected to the initial seed with the same intensity value.

The watershed transform is a popular segmentation method coming from the field of mathematical morphology [4].

II. WATERSHED SEGMENTATION

The description of this transform[1] is quite simple: if we consider the image as a topographic relief, where the height of each point is directly related to its gray level, and consider rain gradually falling on the terrain, then the watersheds are the lines that separate the “lakes” (actually called catchments basins) that form. Generally, the watershed transform is computed on the gradient of the original image, so that the catchments basin boundaries are located at high gradient points [10].

The watershed transform has been widely used in many fields of image processing, including medical image segmentation due to the number of advantages that it possesses: it is a simple, intuitive method, it is fast and can be parallelized and it produces a complete division of the image in separated regions even if the contrast is poor, thus avoiding the need for any kind of contour joining [5]. Furthermore, several researchers have proposed techniques to embed the watershed transform in a multiscale framework.

Advantages of watershed transform:

- The watershed lines always correspond to the most significant edges between the markers. So this technique is not affected by lower-contrast edges, due to noise, that could produce local minima, and, thus, erroneous results, in energy minimization methods.
- Even if there are no strong edges between the markers, the watershed transform always detects a contour in the area. This contour will be located on the pixels with higher contrast.

Fig.1: Basic steps involved in the research
An important task was to design, develop and implement an automated system for cancer detection [8]. Next step was to identify what features must be taken into consideration of a Dicom image for successfully detecting [11] the lung cancer. It is of great significance to find casual association, as well as, correlation between and among the cancer indicating features extracted from image for developing highly successful diagnostic system. Therefore, nature of data collected from various sources (various patients and subjects images). Therefore, the nature of the data was analyzed using regression technique. The main objective for doing analysis of this dataset was to explain variation of one variable (called dependent variable) defined in medical fraternity as state of cancer having either malignant or benign state represented as 0 and 1, respectively in our thesis, based on variations of one or more other variables called independent variables, which includes:

1. Image number
2. Subtlety (difficult to detect the nodule)
3. Number of lesions
4. Sphericity (roundness of a 3-D object)
5. Malignancy
6. Solidity (texture)
7. Lobulation (boundary marked or not)
8. State (benign or malignant)

There are basically two approaches to regression:
• A hit-and-trial approach.
• A pre-conceived approach.
The general regression model (linear) is of type:

\[ Y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n \]

Where, \( y \) is the dependent variable and \( x_1, x_2, \ldots, x_3 \) are the independent variables expected to be related to \( y \) and explain or predict \( y \). \( b_1, b_2, \ldots, b_n \) are the coefficients of the respective independent variables.

A. Flow of steps involved in watershed segmentation pipeline:

1. Input a Medical Image of lungs
2. Pre-process the Image (to remove noise, smooth and enhance its
3. Extract regional properties of connected objects within the image as well as meta information from DICOM images
4. Based on threshold, generate an image having particular sized cells (Cancer)
5. Segment using Watershed to identify the various regional maximas and minimas
6. Segment the image based on the distance

B. Algorithmic steps for Watershed Segmentation [3]:

1. Accept feature map as input and build an image boundary around the borders of the image.
2. Threshold small fluctuations on the low end of the input feature map.
3. Locate and label all single pixel regional minima (pixels those are lower in value than all their 4-neighbors). Pixels are labelled with unique integer value, beginning with 1.
4. Identify all flat regions, give them a distinct label and locate the lowest pixel surrounding the boundary of the region.
5. Trace all pixels that are not members of flat region.
6. Trace all remaining unlabeled flat regions to their respective regional minima (starting with lowest pixel that has already been labelled. By this step all the pixels in the image are labelled with a distinct region label) [7].
7. Mark the flat regions that are regional maxima and mark boundary pixels between region.
8. Depth of each region i.e. Watershed is computed as a difference between the largest and smallest valued pixel in that watershed.
9. Watersheds are thresholded according to their distances.
10. Sequentially labelling all individual regions bounded by the marked edges in the binary image to create a segmented image.

III. RESULTS

Figure 3 illustrates the original image and steps of the segmentation process applied on Lung CT scan training data set.

129
Fig.3: Final results of feature extraction displayed on Lung CT scan images (dataset1): (a) Original images, (b) image after ROI, (c) Results of image superimposition, (d) effects of watershed segmented, (e) extracted cancerous nodule from the image

Input: The dataset consisting of 30 observations is given in table. The dataset is referred to as REGDATA.sav.

Regression Output: The results of the regression model are in the Table in tabular form. Column of the table, titled ‘B’ lists all the coefficients for the model. According to this, 

a (intercept) = -.247
b₁ = .005
b₂ = .138
b₃ = 1.090
b₄ = -.084
b₅ = -.243
b₆ = .266
b₇ = .014

Regression Output:

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>6.772</td>
<td>7</td>
<td>.967</td>
<td>43.279</td>
<td>.000003</td>
</tr>
<tr>
<td>Residual</td>
<td>.469</td>
<td>21</td>
<td>.022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7.241</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Predictors: (Constant), Lobulation, Subtlety, Solidity, Number of lesions, Sphericity, Image number, Malignancy

b) Dependent Variable: State

IV. DISCUSSION

From Table II (analysis of variance table), the last column indicates the p-level to be 0.000004. This indicates that the model is statistically significant at a confident level of (1-0.000003)*100, or 99.997. The p-level indicates the significance of the F-value [10].

The R² value is 0.935, from the Table V; we also note the t-tests for significance of individual independent variables indicate that the significance level (p-level < 0.10) of 0.10, only Subtlety, Number of lesions, Sphericity, Solidity, and Malignancy are statistically significant in the model. The other 2 independent variables are individually not significant.

Thus, from feature selection criteria it is clear that our dataset have cancerous (malignant) lesions that had been identified by using important textual characteristics such as: Subtlety, Number of lesions, Sphericity, Solidity, and Malignancy.
V. CONCLUSION

Medical image segmentation definitely has a large potential in the medical domain. Watershed Segmentation method can be used on a large variety of images and in a wide area of applications. This article shows an improved segmentation algorithm for lung CT images. Further obtained segments can be used for content based medical image retrieval as a diagnostic aid.

We have applied segmentation tools on several pulmonary CT images of lung which are obtained from NIH/NCI Lung Image Database Consortium (LIDC) dataset that offers the opportunity to perform the proposed research. Experiment results show that the proposed method can improve the speed, robustness and accuracy of diagnosis as physician can judge a particular case in right time and with full information which are obtained from Dicom files as Meta information. Hence, this technique is very useful for research for medical fraternity and students.

REFERENCES