AUTOMATIC DETECTION AND EXTRACTION OF LUNGS CANCER NODULES USING CONNECTED COMPONENTS LABELING AND DISTANCE MEASURE BASED CLASSIFICATION

Mamdouh Monif¹, Kinan Mansour², Waad Ammar² and Maan Ammar¹

¹AL Andalus University for Medical Sciences, Faculty of Biomed. Eng., Al Qudmos, Syria
²Al Andalus University Hospital, Al Qudmos, Syria

ABSTRACT

We introduce in this paper a method for reliable automatic extraction of lung area from CT chest images with a wide variety of lungs image shapes by using Connected Components Labeling (CCL) technique with some morphological operations. The paper introduces also a method using the CCL technique with distance measure based classification for the efficient detection of lungs nodules from extracted lung area. We further tested our complete detection and extraction approach using a performance consistency check by applying it to lungs CT images of healthy persons (contain no nodules). The experimental results have shown that the performance of the method in all stages is high.

KEYWORDS

lungs cancer, lungs area extraction, nodules detection, distance measure, performance consistency check.

1. INTRODUCTION

According to World Health Organization (WHO) statistics, lung cancer cases exceeded 13% of all cancer cases appeared in the world, surpassing breast cancer which came second with 11.9%[1]. These facts made lung cancer a major concern for both related specialists and scientists seeking efficient computer aided diagnosis.

Early diagnosis can improve the effectiveness of treatment and increase the patient's chance of survival [2]. The previous facts motivated researchers to pay a great attention to researches that work on automated diagnosis of lung cancer in a wide field known as Computer Aided Diagnosis Systems for Lung Cancer. A reliable computer diagnosis of the disease will help screening a large number of images created every day enabling specialized doctors to work with only little amount of candidate images and raising their efficiency [3]. Hundreds of published scientific papers appeared during the past 3 decades [2]. These published research works were reviewed in several review papers [2,3,4,5,6] to evaluate the overall situation of research on this subject, identify the challenges, and propose specified points to improve the performance of the CAD approaches for lung cancer detection and diagnosis. The review papers discussed in general the four main steps of processing of lungs images: segmentation of the lung fields (regions),
detection of nodules inside the lung fields, segmentation of the detected nodules, and diagnosis of the nodules as benign or malignant.

In general, all review papers called for further improvements in the performance of the available systems, especially, in segmentation of lungs (lungs from chest image) and detection of lung nodules (segmenting and detecting nodules from lung images) and considered this improvement as challenges for further investigation in this field.

In this paper, we introduce an automatic method for detection and extraction of Lungs Cancer Nodules from chest CT images. The method provides good improvements in segmenting lungs accurately from chest CT image with a considerable variety of shapes so that it can be considered as flexible and effective. The method uses 2D Connected Components Labeling (2D-CCL) technique to extract the lungs area from chest image, and after extraction of some suitable features, the method makes use of a weighted distance measure based classification technique to detect nodules with high accuracy.

2. RELATED WORKS

Different techniques were used by researchers to extract nodules from 2D and 3D, CT images. Using two dimensional CT images, Kaur R., et al. [7] used PCA (Principal Component analysis) to extract nodules from lung cancer CT images, and Miwa T., et al.[8] used morphological N-Quoit Filter to automatically extract nodules based on shape and gray level information. Homma N., et al. [9] used Gabor filter and the difference of pixel values along the object axis to detect nodules, and Gomathi M., et al. [10] used FPCM and extreme learning machine for the same purpose. Those were some sample references from the period (2002-2013). Recently, S. Makaju et al. [11], used in 2018 watershed technique for segmentation and some shape and density features to detect nodules., S. Wang et. al. [12] used in (2020) Residual Neural Networks and N. Khehrrah et. al [13] used in (2020) the histogram and some morphological operators to extract the lung, and a threshold based technique to select candidate nodules. The works mentioned above on detecting lung nodules from 2D chest CT images for Computer Aided diagnosis are naturally not exhaustive but give a good idea about the diversity of techniques and methods used.

Ammar M. et al. [14], used the CCL technique to extract liver area from the complicated 2D abdominal CT image to be used for diagnosis of liver cancer. Based on this experience, the authors explored the possibility of using the same technique for the detection and extraction of lung cancer nodules from 2D chest CT images, and presented the encouraging results they obtained in this paper.

3. USED DATA

Images from CT scans of lungs of 11 persons were provided by Alsham Imaging Center and 102 others by Tishreen Hospital. The images in each scan are about 80, and the thickness of each slice is 2 mm. The specialist selected one image from each scan to be used in this study. We divided the 113 images into 2 groups: (1) lungs of 98 cancer patients containing nodules, and (2) lungs of the remaining 15 persons with no nodules (from healthy persons). Both groups were used for lung area extraction from the CT image. For nodule detection, we used the first group for developing and testing the algorithm, and used the second group to check the consistency of the algorithm performance, since the algorithm that detects nodules in the lungs of cancer patients must detect “no nodules” in the images of the lungs of healthy persons. We transformed all the images from DICOM format to JPG format for processing in MATLAB environment. Fig. 1
shows an example from the CT slices of healthy persons, and another one from CT slices of a cancer patient lungs.

![A lung CT image of a healthy person](image1.png) ![A lung CT image of a cancer patient](image2.png)

Fig. 1. A CT image of a healthy person (left), and another one of a lung cancer patient (right). The nodules diagnosed by doctor, are surrounded by a small circles.

4. **NODULES AUTOMATIC EXTRACTION METHOD**

As can be seen in Fig. 1, the original lungs CT image is a complicated content one because it contains, as well as the lungs area, the name of the medical center, the name of the patient, the date, and several other types of information and shapes to help the doctor in diagnosis and archiving. Besides, the lung nodes and the nodules are rather similar in shape, and their gray levels are similar to those of the surrounding region. This situation makes direct extraction of the lungs with their pictorial content from the original image (by thresholding, for example) impossible. Therefore, the general method we use to extract the nodules consists of two main stages, as shown in Fig. 2. Of course, each stage consists of several steps. In the first stage, the lungs area is extracted from the original CT image, and in the second one, the nodules in the lungs area are detected and extracted. If no nodules found in the lungs image, it is marked as "healthy lungs". The output image contains the extracted nodules that can be used later in diagnosis research.

4.1. **Lungs Area Extraction Stage**

The automatic extraction of lungs area from the abdominal CT image is a complicated process. It is rich of details that we should consider in order to extract the lungs accurately from different CT shapes. Therefore, we explain the steps of this stage rather briefly here, then explain later the nodules detection and extraction process. In this stage, we extract the lungs area from the original chest image through two essential processes. In the first one, the known "lungs mask" is extracted through several steps. Then the mask is multiplied by the original image to extract the lungs area from the chest image with the original gray levels preserved. We explain the complete steps in the following:

1– Thresholding the input original image automatically using Outsu's method [15] to select the threshold, then all pixels values in the image below this threshold are set to zero "0", and the rest ones are set to "1", resulting in a binary image. Fig. 3(b) shows the result of thresholding the original image shown in Fig. 3(a).
2- Labeling the connected components in the thresholded image to give each distinct object (connected component) a unique label that enables us, in principle, to compute all possible kinds of features and use them appropriately to extract lungs area or to distinguish the nodules.

3- Finding the largest component which is the closed region surrounding the lungs, by selecting the component that has the maximum area measured by number of pixels in the thresholded image. Fig. 3(c) shows the largest component obtained from Fig. 3(b).

4- Applying a "closing" process to the complement of Fig. 3(c) to remove any remaining printed characters or tiny objects attached to the largest component and to close small holes. The result of this process is shown in Fig. 3 (d).

5- Applying a hole-filling process to the complement of Fig. 3 (d) to get the complete inner area of the largest component, shown in Fig. 3 (e).

6- Removing any organs may remain near the lung: This is done by applying an "opening" process followed by a "closing" one using a circular "structuring element" with a diameter (D=10). This case does not appear in some CT scans. Fig.3 (g) shows an organ removed from Fig. 3(f) by this step. Note that we use here a different image to show this case, which does not appear in all CT images.

7- Multiplying Fig. 3(d) by Fig. 3(e) to get what we called Lung Mask shown in Fig. 3 (h).
Finally, multiplying the Lung Mask by the original image in Fig. 3(a) to extract the lungs area with original gray levels preserved, as shown in Fig. 3 (i).

Fig. 3. Sample results of the eight steps lungs extraction process.

The method explained above extracted successfully lungs areas from CT images of different general shapes and complications, as shown in Fig. 4. These results give an idea of the flexibility and effectiveness of the proposed method.

4.2. Extraction of nodules from lungs image

After having the lungs area extracted, we extract nodules through three main stages: (1) Preparing the lungs area image for feature extraction, (2) Extracting features used for classification, and (3) Classifying the regions remaining in the prepared image into "nodules", and "not nodules". Regions classified as "nodules" remain in the resultant image and the others are removed, as shown in Fig. 2. We explain the three stages in the following in necessary details.
<table>
<thead>
<tr>
<th>Original CT image</th>
<th>Extracted lungs area</th>
<th>Original CT image</th>
<th>Extracted lungs area</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Original CT image" /></td>
<td><img src="image2" alt="Extracted lungs area" /></td>
<td><img src="image3" alt="Original CT image" /></td>
<td><img src="image4" alt="Extracted lungs area" /></td>
</tr>
<tr>
<td><img src="image5" alt="Original CT image" /></td>
<td><img src="image6" alt="Extracted lungs area" /></td>
<td><img src="image7" alt="Original CT image" /></td>
<td><img src="image8" alt="Extracted lungs area" /></td>
</tr>
<tr>
<td><img src="image9" alt="Original CT image" /></td>
<td><img src="image10" alt="Extracted lungs area" /></td>
<td><img src="image11" alt="Original CT image" /></td>
<td><img src="image12" alt="Extracted lungs area" /></td>
</tr>
</tbody>
</table>

Fig. 4. Sample results of extracting lungs areas from images of different general shapes and complication, with very good accuracy.

### 4.2.1. Preparing the lung area for feature extraction

Preparing the image of lungs area for feature extraction is done in three steps, as shown in Fig. 5. These steps are:

1. Binarizing the image using Outs's method to select the threshold in the same way used in step 1 of lungs area extraction process explained in section 4-1.
2. Labeling connected components in the lungs area binary image.
3. Removing small components with areas less than 15 pixels, since as our investigation of all images showed that the regions of such areas are not nodules. Removing these small components will save some of the computation time needed for feature extraction and classification. Fig. 6 shows a sample result with the remaining CCs.
4.2.2. Feature extraction

Specialized doctors diagnosed the nodules used in this study. We found by visual interactive investigation of our data using MATLAB programming facilities that, in general, the nodules have a rather circular shape with almost specific distribution of gray levels, Fig. 7 shows two examples of the interactive examination we made. Therefore, we extracted two groups of features: shape features group and density features group. We will explain the two groups in the following two subsections.
4.2.2.1. Shape Features used in the Study

After labeling the remaining components in the lung image, we extracted the following shape features for every component:

\[ F_{s1}: \text{form factor} = \frac{4\pi \times \text{Area}}{\text{Perimeter}} \]
\[ F_{s2}: \text{roundness} = \frac{4\pi \times \text{Area}}{\pi \times \text{MaxDiameter}^2} \]
\[ F_{s3}: \text{solidity} = \frac{\text{Area}}{\text{ConvexArea}} \]
\[ F_{s4}: \text{extent} = \frac{\text{TotalArea}}{\text{Area Bounding Rectangle}} \]
\[ F_{s5}: \text{compactness} = \sqrt{\frac{4\pi \times \text{Area}}{\pi \times \text{MaxDiameter}}} \]
\[ F_{s6}: \text{aspect ratio} = \frac{\text{MaxDiameter}}{\text{MinDiameter}} \]

Where \( F_s \) stands for: shape feature.

4.2.2.2. Density features

Density features are those related to gray levels of image components. Several density features can be extracted from a gray image [16]. We extracted the following ones:

\[ F_{d1}: 3^{\text{rd}} \text{ normalized moment:} \quad \mu_2 = 255 \times \sum_{i=0}^{255} (z_i - m)^2 p(z_i) \]
\[ F_{d2}: 4^{\text{th}} \text{ normalized moment:} \quad \mu_3 = 255 \times \sum_{i=0}^{255} (z_i - m)^3 p(z_i) \]
\[ F_{d3}: \text{standard deviation:} \quad \sigma = \sqrt{\mu_2} \]
\[ F_{d4}: \text{Smoothness measure:} \quad R = 1 - \frac{1}{\sigma^2} \]
\[ F_{d5}: \text{similarity measure:} \quad U = \sum_{i=0}^{255} p_i^2(z_i) \]
\[ F_{d6}: \text{entropy:} \quad e = -\sum_{i=0}^{255} p(z_i) \times \log_2[p(z_i)] \]

Where: \( m = \sum_{i=0}^{255} z_i \times p(z_i) \); \( F_d \): density feature, \( m \): average of gray levels; \( z_i \): value of the pixel at gray level \( i \); \( p(z_i) \): value of the histogram at gray level \( i \).

4.2.3. Detection of nodules from extracted lungs with remaining CCs

The image of the extracted lungs area with the remaining connected components CCs, like that in Fig. 6, contains nodules and none nodules CCs. Therefore, detecting nodules in this situation
belongs to the Pattern Recognition standard two-classes problem, in which a test sample must be classified whether to belong to the first class (nodule, here), or to the second class (not-nodule). This situation is exactly similar to signature verification problem in which the test signature must be judged whether to be genuine or a forgery. Ammar M. developed a reliable system for signature verification that gives its decision based on a threshold on a Weighted Euclidian Distance Measure computed from suitable features [17]. This system is US patented [18], and still working commercially in hundreds of US banks since about two decades examining about 3 million checks every day. Based on that, this approach is expected to work well in nodule detection. We detected nodules-components by computing the Weighted Euclidean Distance (WED) of each CC in the image using shape and density features explained in the previous two subsections, then using a global threshold obtained from the training group, we classified each CC with a WED less than this global threshold as a "nodule", because the low WED means that the CC resembles the nodules more than the arbitrary CCs. CCs that have a WED equal or larger than the threshold is considered as not-nodule.

5. TRAINING

The training to obtain the global threshold is done on 30 images among the 98 ones containing nodules diagnosed by the specialized doctor. Fig. 8, for example, shows the feature values and the WED of the 4 CCs in a simple case image containing one nodule diagnosed by the doctor, and denoted by "T" in the last column. It is obvious that this nodule is easily separable from the other CCs due to the considerable difference in the WED value. This was an example, but the complete training was done using all shape and density features in 30 images from the 98 images containing nodules used in this study. It is worth noting that the number of remaining CCs in an image reaches 41 in some images.

<table>
<thead>
<tr>
<th>CC #</th>
<th>form factor</th>
<th>roundness</th>
<th>Solidity</th>
<th>Extent</th>
<th>Compactness</th>
<th>Aspect ratio</th>
<th>Distance</th>
<th>Doc. Diag.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.0391719</td>
<td>1001.60565</td>
<td>0.833333</td>
<td>0.535714</td>
<td>31.64815402</td>
<td>2.188309697</td>
<td>1.539006</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>16.5141083</td>
<td>432.901445</td>
<td>0.93</td>
<td>0.9375</td>
<td>20.80628379</td>
<td>1.136570857</td>
<td>1.453795</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>31.7389553</td>
<td>23175.2454</td>
<td>0.885697</td>
<td>0.653846</td>
<td>152.2341793</td>
<td>1.281090021</td>
<td>0.201863</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>17.431828</td>
<td>7495.76987</td>
<td>0.82</td>
<td>0.414141</td>
<td>86.57811425</td>
<td>2.48330078</td>
<td>1.306872</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 8. Shape feature values and Weighted Euclidean Distance Measure of the remaining CCs of a simple test image.

6. RESULTS AND DISCUSSION

After training on the 30 images using shape features once, and both shape and density features once again to obtain the global threshold that separates the nodules from the other CCs in each case, we obtained the results shown in Table 1, where: TP is "True Positive", "FP" is "False Positive", "FN" is False Negative, and the “Sensitivity” is computed by this equation: Sensitivity = TP/(TP+FN).
Table 1. Results of the nodules detection using shape and density features on test cases of 68 images.

<table>
<thead>
<tr>
<th>Feature kind</th>
<th>Sensitivity%</th>
<th>FPs (per case)</th>
<th>No. CCs</th>
<th>No. of Nodules Diagnosed by doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape features</td>
<td>95.1</td>
<td>2.1</td>
<td>1836</td>
<td>173</td>
</tr>
<tr>
<td>Shape + Density features</td>
<td>97.2</td>
<td>1.98</td>
<td>1836</td>
<td>173</td>
</tr>
</tbody>
</table>

We can conclude from Table 1 that: the average diagnosed nodules is about 2.54 per image, and the average number of remaining CCs in an image before detection and extraction is 27. It is worth noting that FPs is a false alarm, and it is not dangerous.

Using the density features with shape features improves the sensitivity by about 2%, and the improvement in FPs reaches about 6%.

6.1. Performance consistency check

As a consistency check of the performance of the nodule detection and extraction algorithm, we tested the lung images of the 15 healthy persons as a consistency-test-group. The result was excellent where 9 false positives (FPs) appeared in 8 images (4% of CCs, as an average). Fig. 9 shows the result for a test image containing seven nodules where all detected with one FP, and the second is for a healthy person where no nodules are detected, and consequently will be classified as "healthy lung".

![Fig. 9](image)

Fig. 9. The first column is the input image, the second column is the extracted lungs image, and the third column is the result of detection and extraction of nodules. In the first row: TPs=7, and FPs=1. In the 2nd row: TPs=0, FPs=0, healthy person (perfect result).
6.2. Comparison with some other works

As can be seen in Table 2 below, the introduced method gave a general performance compared with the best results reported, with rather higher sensitivity and it approached the minimum (FPs) false positives.

Table 2. Performance comparison between other works and the introduced method.

<table>
<thead>
<tr>
<th>Authors</th>
<th>year</th>
<th>Sensitivity%</th>
<th>FPs per case</th>
<th>No. nodules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu [19]</td>
<td>2010</td>
<td>97</td>
<td>4.3</td>
<td>32</td>
</tr>
<tr>
<td>Orozco [21]</td>
<td>2012</td>
<td>96.15</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>Teramoto [22]</td>
<td>2013</td>
<td>80</td>
<td>4.2</td>
<td>103</td>
</tr>
<tr>
<td>Shao [23]</td>
<td>2012</td>
<td>89.47</td>
<td>11.9</td>
<td>44</td>
</tr>
<tr>
<td>Bergtholdt [24]</td>
<td>2016</td>
<td>85.90</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Wu [25]</td>
<td>2017</td>
<td>79.23</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Saien [26]</td>
<td>2018</td>
<td>83.98</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Khehrah [12]</td>
<td>2020</td>
<td>93.75</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Monif (our work)</td>
<td>2021</td>
<td>97.2</td>
<td>2.1</td>
<td>173</td>
</tr>
</tbody>
</table>

7. Future Work

We plan to work on improving the performance of the classification by using the feature selection technique developed by Ammar M. [17].

8. Conclusions

We have introduced a method for the automatic detection and extraction of lungs cancer nodules from CT images using connected component labeling (CCL) and weighted distance measure based classification. The obtained results have shown clearly the high performance of the method in both extraction of lungs area from the CT image and in the correct detection of the cancer nodules. This high performance was also supported by the consistency check we made which reveals the ability of the method to detect the healthy image at the same time.

The experiments conducted and their results analysis presented in this paper enable us to conclude that using CCL technique with appropriate sequence and parameters of some morphological operations may lead to a high performance approach for extracting lungs areas from complex chest CT image with a wide variety in shapes. We can conclude also that using CCL technique with the high flexibility it offers in manipulating CCs in the extracted lungs areas, with the usage of WEDM and a threshold based classification, may provide an efficient method for detecting and extracting nodules to be used later for diagnosis. We found also that using density features with shape ones improves to some extent the performance. We hope that we have introduced a positive effort in the general direction of the research for building an actual automatic lung cancer detection and diagnosis systems.

Acknowledgments

The authors wish to thank Alsham Medical Imaging Center and Tishreen Hospital for providing the lung images data. We appreciate also the support provided by Al Andalus University for Medical Sciences and its hospital.
REFERENCES


AUTHORS

Mamdouh Monif PhD in Biomedical Engineering from Technical University of Graz, Austria, 2000, Associate Professor and Dean of the Faculty of Biomedical Engineering - Al-Andalus University for Medical Sciences, 2020 to present. US patent: SYSTEMS AND METHODS TO MEASURE FLUID IN A BODY SEGMENT- US 9895069 B2; Feb. 2018. Published several papers in biomedical engineering field.

Kenan Mansour MD Obstetrics and Gynaecology Specialist, Tishreen Hospital, 2011. Medical Manager of Al Andalus University Hospital, 2019 to present. Medical Education Master student, Syrian Virtual University, 2020. Interested in medical image analysis and diagnosis.

Waad Ammar MD General Surgery Specialist, Tishreen Hospital, 2020. At present, working at Al Andalus University Hospital. Medical Education Master student, Syrian Virtual University, 2020. Interested in medical image analysis and diagnosis.


© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.