Artificial Intelligence in Medicine 60 (2014) 165–177

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index

Antonio Oseas de Carvalho Filho[a], Wener Borges de Sampaio[a], Aristófanes Corrêa Silva[a,*], Anselmo Cardoso de Paiva[a], Rodolfo Acatauassú Nunes[b], Marcelo Gattass[c]

[a] Federal University of Maranhão, Av. dos Portugueses, SN, Campus do Bacanga, Bacanga, 65085-580 São Luís, MA, Brazil
[b] State University of Rio de Janeiro, São Francisco de Xavier, 524, Maracanã, 20550-900 Rio de Janeiro, RJ, Brazil
[c] Pontifical Catholic University of Rio de Janeiro, R. São Vicente, 225, Gávea, 22453-900 Rio de Janeiro, RJ, Brazil

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form 10 October 2013
Accepted 8 November 2013

Keywords:
Quality threshold
Genetic algorithm
Support vector machine
Lung cancer
Nodule detection
Computer-aided detection
Medical image

ABSTRACT

Objective: The present work has the objective of developing an automatic methodology for the detection of lung nodules.
Methodology: The proposed methodology is based on image processing and pattern recognition techniques and can be summarized in three stages. In the first stage, the extraction and reconstruction of the pulmonary parenchyma is carried out and then enhanced to highlight its structures. In the second stage, nodule candidates are segmented. Finally, in the third stage, shape and texture features are extracted, selected and then classified using a support vector machine.
Results: In the testing stage, with 140 new exams from the Lung Image Database Consortium image collection, 80% of which are for training and 20% are for testing, good results were achieved, as indicated by a sensitivity of 85.91%, a specificity of 97.70% and an accuracy of 97.55%, with a false positive rate of 1.82 per exam and 0.008 per slice and an area under the free response operating characteristic of 0.8062.
Conclusion: Lung cancer presents the highest mortality rate in addition to one of the smallest survival rates after diagnosis. An early diagnosis considerably increases the survival chance of patients. The methodology proposed herein contributes to this diagnosis by being a useful tool for specialists who are attempting to detect nodules.

1. Introduction

Lung nodules are potential manifestations of lung cancer, and their early detection is essential for diagnosis. The most crucial factor related to the occurrence of this type of cancer is the past use of tobacco. In most populations, lung cancer cases related to tobacco use represent 80% or more of the total number of cases [1]. In comparison to non-smokers, smokers have an approximately 20–30 times greater risk of developing cancer. In general, the incidence rates of lung cancer in a particular country directly reflect the consumption of tobacco [2].

The detection of lung nodules is still a challenging task [3]. Lung nodules are difficult to detect using computed tomography (CT), as their densities may be equal to those of other structures or other issues; for example, they may have low density or a small size in a complex anomaly area (connected to vessels and lung edges). Another factor that hinders detection is that specialists have a large number of CT images to analyze. This process is repetitive and exhausting, and the specialists' attention might falter, resulting in an analysis mistake, especially when there are other anomalies in the image. Therefore, this type of analysis is often subject to errors [4].

To minimize such errors, a large amount of research has been conducted to improve computer-aided detection (CAD) and computer-aided diagnosis (CADx). These systems provide a second opinion, helping the radiologist in the interpretation of exams and indicating the diagnosis of lung nodules [5]. Table 6 list some works that propose methodologies for the detection of lung nodules.

The present work introduces a method for automatic detection of solitary lung nodules. The proposed methodology comprises the following stages: (1) acquisition of images from the Lung Image Database Consortium image collection (LIDC-IDRI) and division into training and test sets for validating the methodology; (2) and (3) preprocessing, where the pulmonary parenchyma is segmented and filters are applied to make the nodules more visible; (4) clustering of structures inside the pulmonary parenchyma, which are very similar to lung nodules, using the quality threshold (QT) clustering

* Corresponding author: Tel.: +55 9832728243; fax: +55 9832728841.
E-mail addresses: antoniooseas@gmail.com (A.O. de Carvalho Filho), wenersampaio@gmail.com (W.B. de Sampaio), ari@dee.ufma.br (A.C. Silva), paiva@deinf.ufma.br (A.C. de Paiva), rodolfoacatauassu@yahoo.com.br (R.A. Nunes), mgattass@tecgraf.puc-rio.br (M. Gattass).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.11.002

algorithm; (5) extraction and selection of the most suitable features based on shape and texture; and (6) finally, the segmented structures are classified as nodules and non-nodules. In this stage, we use a micro-genetic algorithm, a variant of the genetic algorithm, to find the best training model and support the vector machine for the final classification.

We believe we bring contributions to this area in the following aspects: (a) the use of the QT algorithm for segmentation of structures that resemble a lung nodule, (b) the use of the genetic algorithm to automatically determine the best training model to be used in the classification stage of the methodology (this process is usually determined manually, and much time is spent determining the best model), and (c) the use of texture measurements based on the diversity index to characterize the nodule.

This paper is divided as follows: In Section 3, we present all the steps of the automatic segmentation of lung nodules, explaining in detail the techniques used for segmentation of the pulmonary pleura and contrast enhancement, the quality threshold algorithm, and the extraction of shape and texture features from the mass candidates, which compose the proposed methodology and evolutionary methodology. In Section 4, we analyze all the tests obtained with the application of the methodology. Finally, in Section 5, we present our final remarks about this work.

2. Related work

The literature available offers acknowledged studies dealing with the problem of automatic segmentation of pulmonary nodules. The following is a summary of some works with the same objective as ours.

Dehmeshki et al. [6] present a method called genetic algorithm template matching for automatic detection of lung nodules. The computation of the fitness function is based on the geometric shape of the voxel, and then combined with the global distribution of the nodule's intensity. The authors report a rate of 14 false positives per exam.

Opfer and Wiemker [7] show that the result of the proposed method's sensitivity is directly related to the size of the nodule, presenting sensitivity of 89% for nodules with a diameter equal to or larger than 4 mm and a sensitivity of 60% for cases smaller than 4 mm.

The methodology presented by Tong et al. [8] follows these steps: (1) segmentation of the pulmonary parenchyma, where detection of suspect nodule candidates is performed using a smoothed Gaussian function for noise reduction, and then a selective filter technique with a Hessian matrix is applied; (2) feature extraction; and (3) classification. The authors obtained a sensitivity of 95%.

The clustering performed by the algorithm proposed by Nie et al. [9] makes use of mean-shift clustering and convergence index resources, forming a CAD system that yields promising results in the detection of lung nodules, with a sensitivity of 89%.

The work by yan Jiang and yu Cheng [10] presents a CAD system with a new approach to overcome lung nodule segmentation problems in two main stages. (1) The transition region is determined by wavelet coefficient statistical resources, and (2) the precise boundaries of the nodule are segmented based on an improved version of the two joint level methods. To validate the methodology, 47 slices were used, and there was a sensitivity of 80%.

A robust and automatic algorithm for lung nodule segmentation is proposed by Sun et al. [11]. A mean-shift estimation method is applied to lung nodule segmentation. This work presents a method for choosing a new bandwidth. The new width determination was used to analyze the K–L divergence rule. The method was evaluated using 95 slices containing 36 nodules, presenting a sensitivity of 100% and 3 false positives per slice.

The work by Dehmeshki et al. [12] describes an algorithm for segmenting different types of lung nodules. It is based on region growing and fuzzy connectivity. Region growing takes place inside a volume mask, which is created by a first application of a local adaptive segmentation algorithm. The initial results achieved a sensitivity of 84%.

The work presented by Miyake et al. [13] illustrates the use of a technique based on temporal subtraction. For the automatic detection of lung nodules, a new method based on artificial neural networks and a temporal subtraction image was developed. To validate the proposed methodology, the authors applied the method to 6 cases, achieving sensitivity of 80.5% for lung nodules smaller than 20 mm in size, with 7.5 false positives per exam.

The detection scheme proposed by Ye et al. [5] was developed to detect two types of nodules: solid and ground-glass opacity ones. The proposed method was trained and validated on a set of 108 exams, presenting a sensitivity of 90.2%.

The work developed by Netto et al. [14] uses a grouping technique based on neural networks, known as growing neural gas, in the extraction of pulmonary structures stage. To reduce false positives, a support vector machine (SVM) was used to classify suspect structures as nodules or non-nodules. Twenty-nine exams from the Lung Image Database Consortium (LIDC) were used. They were divided into training and testing images, a sensitivity of 85.93% and a false positive rate of 0.138 per exam were achieved.

The architecture of the CAD system described by Messay et al. [15] for lung nodule detection was evaluated using LIDC images. In total, 84 exams containing 143 nodules with sizes between 3 and 30 mm were used. This work presents a combination of intensity thresholding and morphological processing to detect nodule candidates. The stepwise technique is used to determine the subset with optimum features. Finally, the authors report a rate of 3 false positives per exam and an average sensitivity of 82.66% for the nodules, using 40 selected features.

The methodology developed by Xiaomin et al. [16] is divided into three stages. In the first stage, a 2D multi-scale filter is used. In stage 2, blob-shaped nodules and non-nodules are differentiated. In stage 3, the shape features of each region are extracted, and a classifier based on automated rules to reduce false positives is applied. The method was applied to 30 exams and presented a sensitivity of 100% and false positive rate of 8.4 per exam.

The CAD proposed by Tan et al. [17] includes innovations such as the use of a new classifier of selective features described as feature-deselective neuro-evolving augmenting of topologies. This classifier is based on neural networks and genetic algorithms. Two other classifiers are also used, namely fixed-topology artificial neural networks and SVM. The process executed by the CAD consists of four steps: preprocessing, detection of nodule candidates, feature selection, and classification. The model was validated by the sensitivity computation (87.5%).

The objective of the study conducted by Suárez-Cuenca et al. [18] was to investigate the usefulness of different methods to combine the classifier to improve the performance of a CAD system in the detection of lung nodules. With a sensitivity of 80%, the number of false positives per case for the six individual classifiers was 6.1 for linear discriminant analysis (LDA), 19.9 for quadratic discriminant analysis, 8.6 for artificial neural networks, 23.7 for SVM-dot, 17.0 for SVM-poly, and 23.35 for SVM-ANOVA. The number of false positives by case for the five combination methods was 3.4 for the majority rule, 6.2 for average, 5.7 for product, 9.7 for neural network, and 28.1 for the likelihood ratio method.

The combination of three CAD systems developed by Italian MAGIC-5 is presented in Camarlinghi et al. [19]. The first is based on a 3D highlight algorithm. The second is based on the normal