Development and Evaluation of a Computer-Aided Diagnostic Algorithm for Lung Nodule Characterization and Classification in Chest Radiographs using Multiscale Wavelet Transform

Amal M. Al Gindi [1,*], Essam A. Rashed [1] and Mostafa-Sami M. Mostafa [2]

[1] Department of Mathematics, Faculty of Science, Suez Canal University, Ismailia 41522, Egypt
[2] Department of Computer Science, Faculty of Computers and Informatics, Helwan University, Cairo, Egypt
algindi_a@yahoo.com

Abstract: Lung cancer is of one of the diseases that lead to high mortality rate globally. It is a leading cause of cancer death for both men and women. So, the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. This work presents a multiscale wavelet transform for comparing the use of four different wavelet families, that are Daubechies, Haar, Biorthogonal and Reverse Biorthogonal Spline wavelets with a total of twenty three wavelets applied separately, for the process of characterization of malignant and benign lung nodules. This is done using three different Regions of Interest (ROI's) sizes to separate the nodules to avoid the difficulties and traditionality of segmentation methods used by researchers to separate very small objects from background. A set of real labeled database images is used to evaluate the proposed system. Results showed that the system is certainly competitive to those described by other known conventional CAD nodule detection systems especially in the area of classification into benign and malignant tumors till present.
[Amal M. Al Gindi, Essam A. Rashed and Mostafa-Sami M. Mostafa. **Development and Evaluation of a Computer-Aided Diagnostic Algorithm for Lung Nodule Characterization and Classification in Chest Radiographs using Multiscale Wavelet Transform.** *J Am Sci* 2014;10(2):83-92]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 14

Keywords: x-ray images, Biorthogonal and Reverse biorthogonal spline wavelets, feature extraction, Lung cancer diagnosis

I. Introduction

Lung cancer is one of the most common and-lethal type of cancer. The five-year survival ratio of lung cancer patients can be highly improved from 14% up to 49% if the lesion detected in the initial phases, when the tumor, presented in the term of lung nodule, is solitary and localize[1]. However, chest radiographs (CXRs) are used far more commonly for chest diseases because they are the most cost-effective, the most routinely available, the most dose-effective diagnostic tool, and they are able to reveal some unsuspected pathologic alterations [2]. Because CXRs are so widely used, improvements in the detection of lung nodules in CXRs could have a significant impact on early detection of lung cancer. Although CXRs are the most widely used modality for lung diseases, it has been well demonstrated that detection of lung cancer at an early stage in CXRs is a very difficult task for radiologists. The difficulties in detecting lung nodules in CXRs are threefold: (1) There is a wide range of nodule sizes, (2) nodules exhibit a large variation in density in CXR, and (3) nodules can be obscured by other anatomic structures. The reasons for misdetection may be due to difference in decision techniques, lack of clinical data, and structured noise in CXRs [3].

Computer aided detection of solitary pulmonary nodules is faced with many difficulties due to the existence of complicated anatomical structures in chest radiographs. Automatic classification of the regions of interest (ROI) as nodules needs to extract a class of powerful region-based features. The performance of region-based feature extraction hinges on a successful nodule segmentation algorithm of suspicious regions [4]. Image features extraction is an important step in the preprocessing of signal processing techniques. The features of digital image could be extracted directly from the spatial data or from a different space. Using a different space by special data transform such as Fourier or wavelets transforms could be helpful to separate a special data that contain specific characteristics from the background. Detecting the features of image texture is a challenging process because these features (or this texture) are mostly variable and scale-dependent [5].

A content based medical image retrieval CAD system for early detection of lung cancer nodules from the chest Computer Tomography (CT) images is presented in [6]. There are different phases involved which are extraction of lung region from chest computed tomography images, segmentation of lung region, feature extraction from the segmented region, and classification of occurrence and non occurrence of cancer in the lung. They used 2D discrete wavelet transform, then using the characteristics to do image compression. Finally, the classification process is achieved using SVM classifier which performs a nonlinear mapping using kernel functions this maps data into a higher dimensional feature space. The

results shows that there are few mis-detections but overall efficiency of vision based efficiency measure is more than 80%.

Various CAD schemes, which utilize both digital image processing techniques and artificial neural networks (ANN's), have been proposed to assist radiologists in detecting lung nodules. A CAD scheme for reduction of false positives in lung nodule detection using two-level neural classification was developed in [7]. The system employed a neural-digital CAD system based on a parameterized two-level convolution neural network (CNN) architecture and on a special multilabel output encoding procedure through employing a receiver operating characteristics (ROC) method with the area under the ROC curve A_z as the performance index to evaluate all the simulation results. The two-level CNN showed superior performance ($A_z = 0.93$) to the single level CNN ($A_z = 0.85$). An efficient lung nodule detection scheme is presented in [8] by performing nodule segmentation through multiscale wavelet based edge detection and morphological operations; classification by using a machine learning technique called Support Vector Machine (SVM). This methodology uses three different types of kernels like linear, Radial Basis Function (RBF) and polynomial, among which the RBF kernel gives better class performance with a sensitivity of 92.86% and error rate of 0.0714.

I. 1 Lung Nodules

A lung nodule is defined as a spot on the lung that is 3 cm in diameter or less. If an abnormality is seen on an x-ray of the lungs that is larger than 3 cm, it is considered a lung mass and not a nodule, and is more likely to be cancerous. Lung nodules must be at least 1 cm in size before they can be seen on a chest x-ray, they are also quite common, and are found on 1 in 500 chest x-rays, and 1 in 100 CT scans of the chest. Approximately 150,000 lung nodules are detected in people in the United States each year. Roughly half of smokers over the age of 50 will have nodules on a CT scan of their chest [9].Most of smokers who undergo thin-section CT have been found to have small lung nodules, most of which are smaller than 7 mm in diameter. Fewer than 1% of very small (<5-mm) nodules in patients without a history of cancer indicate malignant behavior. The importance for serial follow-up for all small nodules is that some of them will turn out to be cancers and that early detection will provide an opportunity for cure [10].

I. 2 Lung Cancer Types and Classification

Lung cancer is a disease characterized by uncontrolled cell growth in tissues of the lung. This growth can spread beyond the lung in a process called metastasis into tissues in the lung and finally into other parts of the body. Most cancers that start in lung, known as primary lung cancers, are carcinomas. The main types of lung cancer are small-cell lung carcinoma (SCLC), and non-small-cell lung carcinoma (NSCLC). Lung cancers are classified according to histological type. This classification is important for clinical treatment of the disease. Most lung cancers are carcinomas and malignancies that arise from epithelial cells. Lung carcinomas are categorized by the size and appearance of the malignant cells seen by a histopathologist under a microscope. The two broad classes are non-small-cell and small-cell lung carcinoma.

This paper organized as follows. Section I is an introduction to lung cancer and its techniques. In section II the materials and methods are proposed, which includes the database used in training and testing the system, and an evaluation of the CADe scheme. Section III presents the proposed scheme which includes image enhancement techniques through applying High Frequency Emphasis filtering to all the images used in the database, also includes the feature generation and extraction approach using four different wavelet families and also an introduction to wavelet transformation and the used wavelets. A description of the classification methodology used is presented in section IV. The achieved numerical results comparing the four mother wavelets through three different ROI sizes are presented and discussed in section V. Conclusion and future scopes are drawn in section VI.

II. Material and Methods

II. 1 X-Ray Image's Database

In the present study a set of 247 chest x-ray images from Standard Public Database by Japanese Society of Radiological Technology (JSRT) database [11], this database which is publicly available have been used in applying the new scheme. This database is selected according to the variety cases included and widely usage in similar research work..The posteroanterior chest films (34.6 cm × 34.6 cm [14 × 14 inches]) for this database were collected from 14 medical institutions by using screen-film systems over a period of 3 years. All nodules were confirmed by CT, and the locations of the nodules were confirmed by three chest radiologists who were in complete agreement. The images were digitized using an LD-4500 or an LD-5500 laser film digitizer (Konica, Tokyo, Japan), with a resolution of 2048x2048 pixels, 0.175-mm pixel size, and 12-bit gray levels corresponding to a 0.0–3.5 optical density range. The database consists of 154 images with confirmed lung nodules and 93 images without a nodule. One hundred nodules were malignant and 54 were benign. The nodules were grouped into five categories, based on the degree of subtlety for detection, which may be

influenced by the nodule size, occlusion by other structures, and nodule density as follows,

Level 1, extremely subtle in which detection is extremely difficult because of low contrast, small size, or overlap with a normal structure;

Level 2, very subtle in which detection is very difficult;

Level 3, subtle in which detection is difficult;

Level 4, relatively obvious in which detection is relatively easy;

and level 5, obvious in which detection is easy. Twenty-five cases were categorized as level 1, 29 as level 2, 50 as level 3, 38 as level 4, and 12 as level 5. This can be categorized in table 1.

Table 1.The Distribution of JSRT images

Category	Az	no. of images
Obvious	0.990	12
Relatively obvious	0.960	38
Subtle	0.876	50
Very subtle	0.753	29
Extremely subtle	0.753	25

The subtlety categories were characterized by the average area A_z under the Receiver-Operating-Characteristic (ROC) curves, which used in evaluating the performance.

II. 2. Evaluation of CAD Schemes

Evaluations of CAD schemes in the past studies were performed without nodules in the opaque portions of x-ray images because that correspond to the retrocardiac and sub-diaphramatic regions of the lung since the purpose of using these CAD scheme was to detect nodules in the lung fields only. But in our scheme we use the whole database (including images which have nodules in the hidden regions) because some lung nodules may be found there. However, designing a complete computerized classification system for chest nodule detection and classification is still a challenging problem, a survey is presented in [12]. A review of some of the most recent solutions is discussed below.

The importance of combining features calculated in different dimensions was developed in [13], they performed CAD experiments on 125pulmonary nodules imaged using multi-detector row CT (MDCT), which computed 192 2D, 2.5D, and 3Dimage features of the lesions. Leave-one-out experiments were performed using five different combinations of features from different dimensions: 2D, 3D, 2.5D, 2D+3D, and 2D+3D+2.5D. The experiments were performed ten times for each group. Wilcoxon signed-rank tests were applied to compare the classification results from these five different combinations of features. The results showed that 3D image features generate the best result compared with other combinations of features, i. e 3D features might be sufficient for the diagnosis of lung nodules regardless of the slice thickness of the CT scans. In [14] another way to improve accuracy of diagnosis is to find images with similar features from real labeled images, and the content-based image retrieval (CBIR) technology has been used for this purpose. Also, it presented a method to find and select texture features of solitary pulmonary nodules (SPNs) detected by computed tomography(CT) and evaluate the performance of support vector machine (SVM)-based classifiers in differentiating benign from malignant SPNs. Seventy-seven biopsy confirmed CT cases of SPNs were included in this study. A set of 25 features were finally selected from a total of 67 after 300 genetic generations. They constructed the SVM-based classifier with the selected features and evaluated the performance of the classifier by comparing the classification results of the SVM-based classifier with six radiologists' observations and ANN-based classifier. The evaluation results showed that the SVM-based classifier had good performance in differentiating the benign from malignant SPNs compared to an average performance of radiologists in this research and was more accurate than the ANN-based classifier in distinguishing benign from malignant SPNs. In [5] a multiresolution analysis system for interpreting digital mammograms is proposed and tested. This system is based on using fractional amount of biggest wavelets coefficients in multilevel decomposition to be the feature vector of the corresponding mammogram, the Daubechies-4, Daubechies-8, and Daubechies-16 wavelets were used in the decomposition process. A set of real labeled database is used in evaluating the proposed system. The results showed that using Dabechies-8 produces the most significantly correct classification rates. It was clear that Daubechies-8 is still the most successful choice of wavelets for their purpose.

In [3], table 2 is given which summarizes the performance comparisons among different CADe schemes for some recent literatures.

II. 3 The Proposed Scheme

Figure 1. shows the main steps of our CAD scheme for lung nodule classification in x-ray images which consists of (1) A two-stage image enhancement technique (2) Feature generation using multiscale wavelet analysis using four different mother wavelets with 23 different wavelets (3) feature extraction of the candidate ROI's using three different ROI sizes 32 x 32 pixels, 64 x 64 pixels and 128 x 128 pixels through three separate experiments, (4) dimensionality reduction of the selected features which done on three stages(5) classification of the nodule candidates into malignant or benign using the new approach.

Table 2. Performance comparisons of CADe schemes used the JSRT database for evaluation in [3].

	Sensitivity	FPs/image	Database
Wei et al. (2002)	80% (123/154)	5.4 (1333/247)	All nodule and normal cases in JSRT (247)
Coppini et al. (2003)	60% (93/154)	4.3 (662/154)	All nodule cases in JSRT (154)
Schilham et al. (2006)	51% (79/154)	2.0 (308/154)	All nodule cases in JSRT (154)
	67% (103/154)	4.0 (616/154)	
Hardie et al. (2009)	80% (112/140)	5.0 (700/140)	Nodule cases in JSRT (140)
	63% (88/140)	2.0 (280/140)	

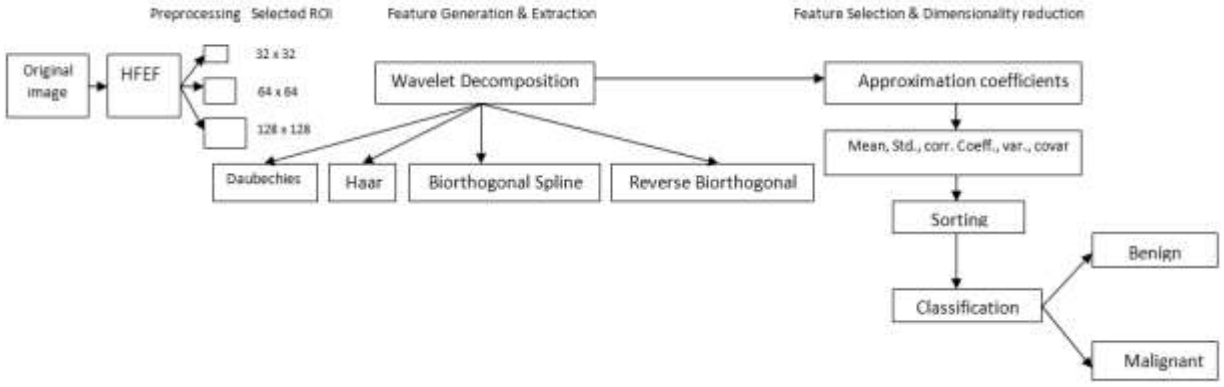

Fig.1 The main diagram for the CAD scheme

II.3.1 Two-stage image enhancement technique

We developed a two-stage enhancement technique in which the first stage is converting the image format from16-bit (BIG endian) to 32-bit (Little endian) format for all the images in the database . Big-endian and little-endian are terms that describe the order in which a sequence of bytes are stored in computer memory. The second stage is applying High-Frequency Emphasis Filtering and image histogram to the preprocessed images which is -just as lowpass filtering blurs an image - the opposite process, high pass filtering, sharpens the image by attenuating the low frequencies and leaving the high frequencies of the Fourier transform relatively unchanged.

II.3.1.a Frequency Domain Processing

Filtering in the frequency domain is quite simple conceptually, the general idea is that the image (*f(x,y)* of size *M* x *N*) will be represented in the frequency domain (*F(u,v)*). The equation for the 2-D, discrete Fourier transform (DFT) is given by

$$\boldsymbol{F(u,v)} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \boldsymbol{f(x,y)e^{-j2\pi(\frac{ux}{M}+\frac{vy}{N})}}$$

The frequency domain is simply the coordinate system spanned by *F(u, v)* with *u* and *v* are the frequency variables. This is analogous to the spatial domain which is the coordinate system spanned by *f(x, y)*, with x and y as spatial variables.

The concept behind the Fourier transform is that any waveform that can be constructed using a sum of sine and cosine waves of different frequencies. The exponential in the above formula can be expanded into sines and cosines with the variables *u* and *v* determining these frequencies, and the idea in frequency domain filtering is to select a filter transfer function that modifies F(u, v) in a specified manner. The foundation for linear filtering in both the spatial and frequency domain is the convolution theorem, which shows the relationship between the spatial domain and frequency domain which written as

$f(x,y) * h(x,y) \Leftrightarrow H(u,v)F(u,v)$

and, conversely,

$f(x,y)h(x,y) \Leftrightarrow H(u,v) * G(u,v)$

The symbol "*" indicates convolution of the two functions. The important thing to extract out of this is that the multiplication of two Fourier transforms corresponds to the convolution of the associated functions in the spatial domain.

II.3.1.b Lowpass and Highpass Frequency Domain Filters

Based on the property that multiplying the FFT of two functions from the spatial domain produces the convolution of those functions, fourier transforms can be used as a fast convolution on large images. Filters also can be created directly in the frequency domain. There are two commonly discussed filters in the frequency domain

Lowpass filters, also known as smoothing filters -

Highpass filters, also known as sharpening filters-

Given the transfer function $H_{lp}(u, v)$ of a lowpass filter, we obtain the transfer function of the corresponding highpass filter by using the simple relation

$H_{hp}(u, v) = 1 - H_{lp}(u, v)$

Examples of this kind of filters are ideal, Butterworth, and Gaussian highpass filters. However, because the average value of an image is given by F(0, 0) , and these highpass filters far zero-out the origin of the Fourier transform, the image has lost most of the background tonality present in the original.

II.3.1.c High-Frequency Emphasis (HFE) filtering

In which an approach to compensate for this problem of lost most of the background tonality of the original image is to add an offset to a highpass filter. When an offset is combined with multiplying the filter by a constant greater than 1, the approach is called *High-Frequency Emphasis* filtering because the constant multiplier highlights the high frequencies. The multiplier increases the amplitude of the low frequencies also, but the low-frequency effects on enhancement are less than those due to high frequencies .as long as the offset is small compared to the multiplier. High-frequency emphasis has the transfer function

$H_{hfe}(u, v) = a + bH_{hp}(u, v)$

Where a is the offset, b is the multiplier, and $H_{hp}(u, v)$ is the transfer function of a highpass filter [15]. A Butterworth highpass filter of order 2 and standard deviation value equal to 5% of the vertical dimension of the padded image width were used, then using high-emphasis filtering with a = 0.5 and b = 2.0 we got an advantage for the resulted image in which the gray-level tonality due to the low-frequency components was retained, and as known an image characterized by gray levels in a narrow range of the gray scale is an ideal candidate for histogram equalization, so, histogram equalization was an appropriate method to be used for further enhancement.

Figure 2 show the difference between the image before and after applying the high-frequency emphasis filter and histogram equalization.

The clarity of the bone structure and other details in the enhanced image is obvious although it appears a little noisy, but this is typical of X-ray images when their gray scale is expanded. The result obtained using a combination of high frequency emphasis and histogram equalization is superior to the result that would be obtained by using either method alone.

Fig. 2. Enhanced lung field using high frequency emphasis filtering

II.3.2 Nodule Selection and Extraction for selected ROI's

From the selected images in the JSRT database, the following steps were carried out

(i) Nodule Selection, a regions of interest (ROI's) were first selected depending on that the locations of the nodules were confirmed by three chest radiologists who were in complete agreement as we mentioned above .

(ii) Nodule Extraction, ROI's were extracted with different sizes128 x 128 pixels, 64×64 pixels and 32 x 32 pixels regardless of the nodule size, these selected nodule sizes is because the average size of all nodules included in the database was 17.3 mm. and the average pixel size is 0.175- mm, i.e. 32 x 32 pixels nearly equal 5.6 mm, 64 x 64 pixels nearly equal 11.2 mm, and 128 x 128 pixels nearly equal 22.4 mm, which permits to analyze nodules with some surrounding areas . The reason for selecting nodules with some surrounding areas is first, wavelets have the ability to differentiate and preserve details at various resolutions, which make it easy for them to separate small objects (nodules) from large objects (backgrounds) regardless of the size of the neighbourhood, second radiologists observations on the benign and malignant nodules and their background suggests that the surrounding areas may carry information which might differentiate them. [16]

III.3 Wavelet transform and the new approach

We used discrete wavelet transform for the process of ROI's feature extraction for two reasons first, since wavelet transform is an efficient tool for multi resolution and texture analysis, second, to avoid the problems of nodule segmentation and thresholding techniques. The idea of wavelets was briefly discussed in [17], the basic idea behind wavelet transform is to analyze different frequencies of a signal using different scales or resolutions. High frequencies are analyzed using low scales while low frequencies are analyzed in high scales, which is more flexible approach than the Fourier transform since it enabling analysis of both local and global features. In wavelet transform, all of the basis functions, which are called wavelets, are derived from scaling and translation of a single function, called mother wavelet. Depending on their properties, the wavelet transform can be divided into two categories, i.e., the redundant wavelet transform and the orthogonal wavelet transform. Orthogonal wavelet transform allows an input image to be decomposed into a set of independent coefficients, corresponding to each orthogonal basis i.e. orthogonality implies that there is no redundancy in the information represented by the wavelet coefficients, which results in efficient representation and other desirable properties.

In wavelets a set of functions can be generated by translations and dilations of the mother function [5]. Wavelets decomposition is based on applying 2D wavelets transform to the image and a set of four different coefficients are produced in each level of decomposition. The produced coefficients are

– Low frequency coefficients (A).
– Vertical high frequency coefficients (V).
– Horizontal high frequency coefficients (H).
– High frequency coefficients in both directions (D).

An image decomposed with 3-level wavelet transform is illustrated in Fig. 3

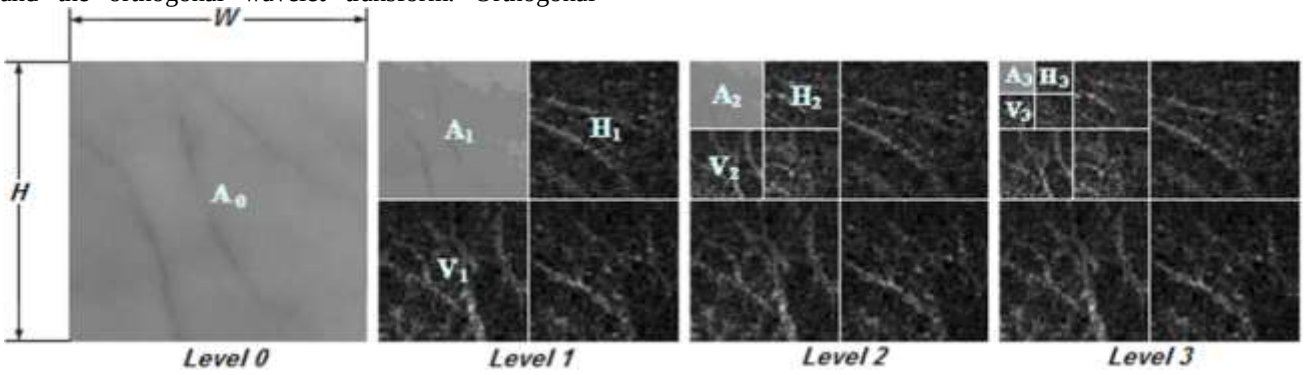

Figure 3 Three-level wavelet decomposition

In the proposed technique, we calculated the discrete wavelet transform (DWT) feature vector - for the enhanced images - which is a multi-resolution representation of the nodule signal. The features used for classification were selected using four different levels of decompositions based on four different mother wavelet families that are Daubechies (10 wavelets), Haar (1 wavelet), Biorthogonal Spline(8 wavelets, 4 for decomposition and 4 for reconstruction) and Reverse-Biorthogonal (4 wavelets) i. e a total of 23 wavelets. In our trial for reducing the dimensionality of the resulting feature coefficients we did the following 1) in each level of decomposition, a percentage of the low frequency coefficients i.e. approximation coefficients are used to represent the corresponding feature vector by covering all possible rates (10–90%), 2) the mean , standard deviation, variance, covariance and correlation coefficients were calculated for each level of decomposition separately, each in a separate experiment , 3) sorting the feature vector decendingly. The idea of using a percentage of the low frequency coefficients was previously discussed in [5] and extended here by using four different mother wavelets and a total of 23 wavelets, also, computing the parameters mean, standard deviation variance, covariance and correlation coefficients for the approximation coefficients in each of the four levels of decomposition to compare their performance and select the best one in classification problem. A set of 65% and 32% nodule images from JSRT respectively, extracted randomly through the 5 different subtlety levels, were used to train and test the proposed method including some of those used to produce the class core vectors and also nodule images in the hidden regions or opaque portions of the x-ray images that correspond to the retrocardiac and subdiaphramatic regions of the lung i.e. that obscured by mediastinum, the diaphragm and other anatomic structures, at the time the evaluations of CAD schemes in most of the past studies were performed with excluding nodules in these portions.

III.4Biorthogonal and Reverse-Biorthogonal Spline Wavelets

Polynomial splines are a common source for wavelet construction. Two approaches governed the construction of wavelet schemes that use splines. One is based on Orthogonal and semi-orthogonal wavelet in spline spaces which produces compactly supported spline wavelets. The other one employs splines in wavelet analysis [18]. Haar wavelet is the only orthogonal wavelet with linear phase, Biorthogonal wavelets feature a pair of scaling functions and associated scaling filters, one for analysis and one for synthesis. While, the family of Reverse Biorthogonal Wavelet Pairs is obtained from the biorthogonal wavelet pairs previously described. [Wavelet toolbox™ User's Guide R2013b]

IV. Classification

IV. 1 Euclidean Distance Classifier

Using the same classifier proposed in [5], the Euclidean distance is used to design the classifier that is based on calculating the distance between the feature vectors for the tested ROI's and the class core vectors which previously calculated in the training phase . The system automatically classifies the feature vector to a diagnosis class C_{diag} by finding the nearest class to this vector. This is done by testing the distance between this feature vector and all class core vectors, the following equations were used to the classification process C_m

$$\mathrm{Dist}(A, C_{diag}) = \mathrm{MinDist}, \quad (1)$$

$$\mathrm{MinDist} = \min_{1 \ll m \ll M} \left(\mathrm{Dist}(A, C_m)\right), \quad (2)$$

$$Dist(A, C_m) = \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{L^j} \sqrt{(A^j(i) - C_m^j(i)^2}, \quad 1 \ll m \ll M \quad (3)$$

While A_j is the coefficient vector of the jth decomposition level for diagnosis image, C_m^j is the class core vector for class m at decomposition level j, L^j is the length of coefficient vector at decomposition level j, M is the number of classifications classes, and J is the number of decomposition levels used. The class core vector for each experiment were calculated using the following equation,

$$C_m^j = \frac{1}{N^j} \sum_{n=1}^{N^j} \sum_{i=1}^{L^j} A_m^j(i), \quad 1 \ll m \ll M, \quad (4)$$

Where N_j is the number of selected ROI's to produce the class core vector at decomposition level j, and A_m^j is the coefficient vector for ROI's for the class m at decomposition level j.

V. Experimental Results

In this work an evaluation and performance comparisons to the problem of correct classification of benign and malignant lung nodules, for this purpose we applied the Discrete Wavelet Transform (DWT)-based feature extraction for 65% randomly selected ROI's all over the database through the 5 levels of subtlety. In each experiment we used the same methodology, the same approach, the same dataset and classifier. The aim of this evaluation is to select one of the wavelets for better classification purpose. Thus, we have used a series of experiments with the use of different wavelets and Euclidean distance classifier and also three different ROI sizes to select the one with the best classification accuracy. We computed the DWT-based feature extraction using a program code written in Matlab® version 7.9.0.529 (R2009b) with the use of Matlab® Image processing toolbox and Wavelet toolbox in applying the wavelet transform process.

The Daubechies, Haar, Biorthogonal and Reverse-Biorthogonal Spline were used in this work in the decomposition process, In each experiment, four levels of decompositions are applied, and then in each level of decomposition the mean, standard deviation, variance, covariance and correlation coefficients were calculated each in a separate experiment for each group of coefficients in each level to get the feature vector of the training set. Table 3 shows the total correct classification rate obtained when we used Daubechies-4, as an example, to decide which calculations will be better used to reduce the dimensionality of the resulting feature vector for the rest of the work.

In practical work we consider two main problems, the first one is to choose the best wavelets among 23 used wavelet candidates with the best calculations to classify lung nodules according to the level of risk into two classes benign and malignant, the second problem is to test their performance through three different ROI sizes which permits to analyze nodules with some surrounding areas.

It is clear from Table 3 that using mean and standard deviation we got better classification rate. For this reason we did a comparison through tables 4 and 5, between the 23 different wavelets of the 4-mother wavelets using mean and standard deviation to choose 3 different wavelets to be the candidate wavelets to complete the work.

In the process of applying the four wavelet families, it is clear from table 4 and 5 that Biorthogonal and Reverse Biorthogonal Spline wavelets results in better classification results than that of Daubechies and Haar, using mean (average) than that using standard deviation, and it is clear also that using Db-2 and Db-3 a 84% and 82% of correct classification rate is achieved respectively when computing the average (mean) of the approximation coefficients in each level, also using bior1.5 (used for decomposition) 95% of correct classification rate is obtained. For bior3.3 (used for reconstruction) a 92% is achieved, and for bior1.3 (used for decomposition) and bior2.4 (for reconstruction) a 89% of correct classification rate is

achieved. Also, using bior5.5 we got 87% and using bior3.1 84% of correct classification rate is achieved.

Table 3:Correct classification rate, in percentage, obtained using Daubechies-4 with different statistics

Mother Wavelet Name	Number of images (%)	Total Correct classification rate (%)				
Daubechies-4	32%	Mean	Std	Var	Cov	Corrcoef
		74	71	50	60	53

Table 4 Correct classification rate, in percentage, with 10 Daubechies wavelets for 32% of JSRT images

Daubechies Wavelet Family	Total Correct classification rate (%)	
	Mean	Std
Db-1	68	68
Db-2	**84**	76
Db-3	**82**	74
Db-4	74	71
Db-5	74	55
Db-6	74	58
Db-7	74	53
Db-8	74	53
Db-16	63	47
Db-32	58	66

Table 5 Correct classification rate, in percentage, with Haar, Biorthogonal and Reverse Orthogonal Spline wavelet families for 32% of JSRT images.

Wavelet Family	Total Correct classification rate (%)	
	Mean	Std
Haar	68	68
bior1.1	68	68
bior1.3	89	79
bior1.5	**95**	76
bior2.2	89	76
bior2.4	89	76
bior3.1	84	71
bior3.3	**92**	76
bior5.5	87	68
bior6.8	76	66
rbio1.3	**92**	76
rbio2.4	**92**	74
rbio3.3	**92**	76
rbio4.4	89	76

Table 6 Correct classification rate, in percentage, with Biorthogonal Spline wavlets (bior1.5 & bior3.3) and Reverse Orthogonal Spline wavelet rbio2.4 for ROI size of 128 x128 pixel

Wavelet coefficients (%)	bior3.3		bior1.5		rbio2.4	
	Benign	Malignant	Benign	Malignant	Benign	Malignant
10%	50	52	31	48	77	60
20%	31	56	31	48	23	48
30%	76	64	62	72	69	72
40%	23	76	38	72	38	60
50%	46	72	62	72	46	72
60%	69	80	77	84	69	80
70%	77	88	77	88	77	84
80%	77	88	77	**92**	77	84
90%	85	**96**	85	**92**	85	**92**
100%	85	**96**	85	**100**	85	**96**

Table 7 Total correct classification for the specified wavelets for three ROI sizes for 100% of Wavelet coefficients

ROI Size	bior3.3	bior1.5	rbio2.4
128 x 128	**92**%	**95**%	**92**%
64 x 64	74%	74%	74%
32 x 32	63%	58%	58%

Also it is clear that using Reverse Biorthogonal Spline we got results which are similar to that of Biorthogonal Spline, as using rbio1.3 and rbio2.4 and rbio3.3we got 92% of correct classification rate is achieved. So, we decide to use each of bior3.3, bior1.5 and rbio2.4 to complete the work and decide which wavelet is the best in the classification of benign and malignant lung nodules. The same procedure of applying four levels of decompositions is used, and then a fractional percent of the coefficients is used to be the feature vector of the corresponding nodule image. Table 6 shows the successful classification rate during using the three specified wavelets in all possible fractions of coefficients (10 – 90%).

From Table 6, it is clear that the highest rates is obtained using a set of 100% of the resulting feature vector (coefficients), as using bior3.3 a 96% correct classification rate for malignant and 85% for benign is reached and the same results is obtained using a set of 90% of wavelet coefficients. Also using bior1.5 a 100% of correct classification for malignant nodules and 85% for benign is obtained while using rbio2.4 a 96% for malignant and 85% for benign is obtained. Also, it is clear that using smaller fractions of coefficients is not suitable for our case for this reason we decide to complete our experiment with the other two ROI sizes in case of 100% of the coefficients i.e. all the coefficients in the feature vector. Table 7 shows the classification results in percentage for each of the three selected wavelets bior3.3, bior1.5 and rbio2.4 in three cases of ROI sizes. From which we can notice that classification accuracy is sensitive to ROI size and it is clear that for the three selected wavelets we got the highest classification accuracy with the largest ROI size since it we obtain 95% correct classification using bior1.5 with ROI size 128 x 128 pixel, and it is decreased as ROI size decrease which agrees with the idea discussed previously in section III. 2 that wavelets have the ability to differentiate and preserve details at various resolutions and radiologists suggestions that the surrounding areas around the nodule may carry information which might differentiate them. This generally means that the CAD scheme proposed in this article appears to perform very well in comparison to that in Table 2.

VI. Conclusion and Future Work

Early detection of cancereous pulmonary nodules may be a way to improve a patient's chances for survival. This paper presented a detailed comparison of the performance of four different wavelet families, when applied separately for the problem of lung cancer diagnosis. Experiments were applied on real labeled data even those in the hidden regions which is a valuable area of future work of most of the recent researches. In this study, a new approach is presented using an idea has been proved earlier which is using multiresolution analysis or wavelet decomposition analysis in lung nodule characterization and feature extraction. The results is obtained using only the JSRT dataset this suggested that the results is enhanced compared with other performance comparisons and that the proposed scheme appears to be promising compared to most of the previously published researches.

For future work, other feature analysis, feature extraction and dimensionality reduction tools might be used. The classification performance might be improved using other classifiers such as support vector machines instead of Euclidean distance which may affect the results achieved here positively or negatively.

References

1. Murphy G. P., W. Lawrence Jr., and R. E. Lenhard Jr., *Clinical Oncology*.Washington, DC: Amer. Cancer Soc., 1995
2. Murphy G. P. *et al.*, *American Cancer Society Textbook of Clinical Oncology*,2nd edition, The Society, Atlanta, GA, (1995).
3. Chen S., K. Suzuki, and H. MacMahon, "Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification," Med. Phys. 38 (4), 1844-1858 (2011).
4. Coppini G., S. Diciotti, M. Falchini, N. Villari, G. Valli, "Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms," *IEEE Transactions on Information Technology in Biomedicine*, Vol.7, 344-357 (2003).
5. Rashed E. A., I. A. Ismail, S. I. Zaki," Multiresolution mammogram analysis in multilevel decomposition," Pattern Recognition Letters 28, 286-292 (2007).
6. Yadav N. G., "Detection of Lung Nodule Using Content Based Medical Image Retrieval,"

International Conference on advanced Engineering & Technology, (2012).
7. Lin J. S., S. C. B. Lo, A. Hasegawa, M. T. Freedman and S. K. Mun, "Reduction of False Positives in Lung Nodule Detection Using a Two-Level Neural Classification," IEEE TRANSACTIONS ON MEDICAL IMAGING. Vol.15, No. 2, 206-217 (1996).
8. Aarthy K. P., U. S. Ragupathy, "Detection of lung nodule using multiscale wavelets and support vector machine," International Journal of soft Computing and Engineering (IJSCE) 2(3), 32-36 (2012).
9. L. E. MD, "Lung Nodules - Symptoms, Causes, and Diagnosis," About.com Guide *Medical Review Board*, (2013).
10. MacMahon H., Austin J., Gamsu j. e*t al.*, "Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society" Radiology. 237, 395-400 (2005).
11. Shiraishi J., Katsuragawa S., Ikezoe J. *et al.*, Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," AJR, Am. J. Roentgen. **174**(1), 71–74 (2000).
12. El-Baz A., Beache G, Gimel'farb G. *et al.*, "Computer-Aided Diagnosis System for Lung Cancer: Challenges and Methodologies: Review Article," International Journal of Biomedical Imaging, 1-46 (2013).
13. XU Y. Lee M., Boroczky L. *et al.*, "Comparison of Image Features Calculated in Different Dimensions for Computer-Aided Diagnosis of Lung Nodules ," Medical Imaging Proc. of SPIE (2009).
14. Zhu Y., Tan Y., Hua Y. *et al.*, "Feature Selection and Performance Evaluation of Support Vector Machine (SVM)-Based Classifier for Differentiating Benign and Malignant Pulmonary Nodules by Computed Tomography,"Journal of Digital Imaging, 23(1), 51-65 (2010).
15. Gonzalez R. C., R. E. Woods, S. L. Eddins, "Digital Image Processing Using Matlab," Pearson Education, Inc.(2004).
16. Teresa O., "Wavelet-based Pulmonary Nodules Features Characterization on Computed Tomography (CT) Scans," ProQuest Information and Learning Company, (2008).
17. Sakka E., A. Prentza, I. E. Lamprinos and D. Koutsouris, "Microcalcification Detection using Multiresolution Analysis based on Wavelet Transform," Biomedical Engineering Laboratory, National Technical University of Athens.
18. Averbuch A. Z., V. A. Zheludev,"A New Family of Spline-Based Biorthogonal Wavelet Transforms and Their Application to Image Compression, "IEEE TRANSACTIONS ON Image Processing, VOL. 13, NO.7, (2004).

2/13/2014