35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

On the Combination of Wavelet and Curvelet for Feature Extraction to Classify Lung Cancer on Chest Radiographs

Hamada R. H. Al-Absi, Brahim Belhaouari Samir, Taha Alhersh and Suziah Sulaiman

Abstract— **This paper investigates the combination of multiresolution methods for feature extraction for lung cancer. The focus is on the impact of combining wavelet and curvelet on the accuracy of the disease diagnosis. The paper investigates feature extraction with two different levels of wavelet, two different wavelet functions and the combination of wavelet and curvelet to obtain a high classification rate. The findings suggest the potential of combining different multiresolution methods in achieving high accuracy rates.**

I. Introduction

Cancer is a widely used term nowadays that explains the situation in which cells of a specific organ in the human body start to grow in an uncontrolled way. According to the National Cancer Institute [1], there are more than 100 types of cancer affecting different organs in the human body; One of these types is lung cancer.

Lung cancer is a disease that causes death more than any other type of cancer [2]; in fact, nearly 13% of the total new cancer cases diagnosed in 2008 were lung cancer [3]. Primarily, the main cause of this disease is cigarette smoking [1]. In order to contribute to the early detection of this disease and others as well, which could prevent death; many computer aided diagnosis systems have been developed.

Computer Aided Diagnosis (CAD) is a field that analyzes medical images to look for any signs of abnormality. Research on CAD has been taking place in many research institutions around the world to develop systems with high capabilities in disease diagnosis and many of this research focuses on lung cancer diagnosis. For instance, A system for lung cancer detection introduced by Sousa et al. [4]. The system consists of a thorax extraction step to remove all external objects. After that, a lung extraction step to identify the lung region and then another step to reconstruct the lung region to prevent any removal of the lung. A structure extraction step is then executed to select the dense structures from the lung; after that, a tubular elimination step to remove the many of the pulmonary trees and finally, a false positive step to select regions that have a high probability of having a nodule. The performance of the system was 84.84% sensitivity and 96.15% specificity. Another CAD system for lung nodule detection was introduced by Lee et al. [5]. The system introduced a classifier based on random forest that is aided by a clustering method to detect lung nodules. 32 scans of patients' lungs were used for the assessment of the proposed system. The performance of the system was 98.33% sensitivity and 97.11% specificity. A system for lung nodule classification was introduced in Zhang et al. [6]. The system uses rule-based and SVM classifier. First, each region is subjected to feature extraction, where a combination of seven shape features, two gray features, and four texture features were calculated. All extracted features are used as an input to the classifier, which is a combination of rule-based and SVM; the system achieved an overall accuracy of 84.39%. Orozco et al [7] presented a system to classify lung nodule in frequency domain using SVM. The system begins with a manual selection of the regions of interest, which are subjected to the calculation of the 2D Discrete Cosine Transform and 2D Fourier Transform. After that, statistical texture feature is extracted. A total of 24 texture features are extracted and then reduced to 2 after a selection process. The selected features were classified with SVM classifier that is based on the Radial Basic Function (RBF). The system achieved a sensitivity of 96.15% and specificity of 52.17% with 82.66% preciseness. Kumar et al [8] introduced a CAD system for the diagnosis of lung nodules. The system begins with pre-processing the input images to enhance the sharpness using bi-orthogonal wavelet and enhance the contrast using bi-histogram equalization. The second step in the system is segmentation of regions of interest using region growing method. The final step of the system is the decision making (diagnosis); this step is done using the fuzzy inference system. The system achieved an accuracy of 90% with sensitivity and specificity of 86% and 84% respectively. Another system was proposed by Chen et al [9] for the classification of pulmonary nodules based on the neural network ensemble. In the system, the neural network ensemble was constructed with a multilayer neural network with the back-propagation algorithm, radial basis probabilistic neural network and learning vector quantization neural network. The outputs of the three networks were generated in the Bayesian probability form, to be put in the weighted sum by the Bayesian criterion and classify the images. Results achieved in this system concluded that the neural network ensemble produces higher accuracy than the individual networks with 78.7%.

This paper presents a CAD system that is comprised of feature extraction, feature selection and classification. Particularly, it aims on investigating the effect of the combination of wavelet and curvelet for feature extraction for lung cancer diagnosis. In addition, the paper examines the combination of multiple wavelet functions as well as the

Hamada R. H. Al-Absi is a PhD student at the Department of Computer & Information Sciences, Universiti Teknologi PETROAS, Malaysia (phone: +60175989915; e-mail: Hamada.it@gmail.com).

Brahim Belhaouari Samir is an Assistant Professor of Mathematics at the College of Science, Alfaisal University, Riyadh, Kingdom of Saudi Arabia. (e-mail: sbelhaouari@alfaisal.edu).

Taha Alhersh is an independent researcher, Jordan (e-mail: taha.trh@gmail.com).

Suziah Sulaiman is a Senior Lecturer at the Department of Computer & Information Sciences, Univeristi Teknologi PETRONAS, Malaysia (e-mail: suziah@petronas.com.my)

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

combination of different levels in the same wavelet for feature extraction. This paper is part of an on-going research on the use of wavelet transform and curvelet transform for lung cancer diagnosis.

II. Materials and Methods

A. Dataset

In this study, the Japanese Society of Radiological Technology (JSRT) standard chest radiographs dataset is utilized [10]. The dataset contains both nodules and non-nodules images with a total of 247 images. The distribution of the images in this dataset is further shown in Table I.

TABLE I. IMAGE DISTRIBUTION IN THE JSRT DATASET

Class	Benign	Malignant	Total
Abnormal	54	100	154
Normal	-	-	93
Total			247

The images in the dataset are x-ray images with 2048x2048 pixels in size. The coordinates of the abnormalities in the 154 images that contain nodules are provided. These coordinates are used to extract the regions of interest with a 128x128 pixels for each sub-image. Fig. 1 shows the first image in the dataset (JPCLN001.IMG) with the extracted regions that contain the abnormality.

(a) (b)

Figure 1. An Example of the JSRT Dataset (JPCLN001.IMG) (a) Original Chest Radiograph (b) Extracted sub-image [10]

As for the normal images, random regions from the 93 images are extracted. These randomly selected regions represent different areas in the lung.

B. Methods

As shown in fig. 2, our CAD system consists of a feature extraction phase, a feature selection phase and a classification phase. The following sub-sections explain each phase.

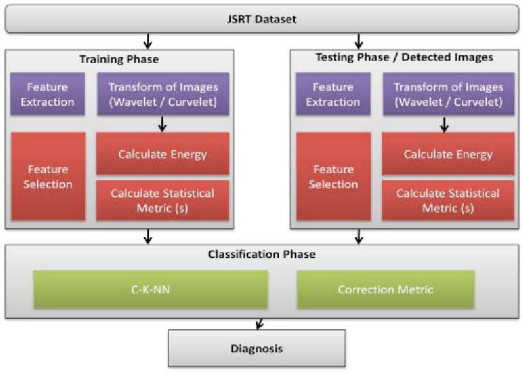

Figure 2. System Overview

1) Feature extraction

Feature extraction is a significant step in CAD system. In this paper we utilize wavelet and curvelet separately and combined in order to investigate their performance and their contribution to the diagnosis of lung cancer. Both wavelet and curvelet are well known techniques and further explanation of each can be found in [11] and [12].

2) Feature selection

Feature selection is another important step in CAD system. The aim of this step is to remove the extracted features that are redundant and have no contribution to the classification process. After applying wavelet, curvelet or the fusion of both of them, the generated coefficients are subjected to a two-step feature selection, namely, statistical energy and statistical metric.

a) Statistical Energy Calculation

The statistical energy is calculated as follows:

$$\text{Energy\_metric}(k) = \frac{\Sigma_i \Sigma_j \left|n_j^i(k)\right|}{\Sigma_i n_i(k)}, \qquad (1)$$

where $n_j^i(k)$ is feature k of image j in class i . A certain threshold is fixed to remove coefficients with small energy. This threshold is identified after sorting the calculated values in ascending order and plotting them. The remaining coefficients are subjected to the next selection process.

b) Statistical Metric Calculation

The statistical metric to select the optimum coefficients is calculated as follows:

Suppose m1, m2 and m3 are the mean of class1, class2, and class3, respectively and m_T is the mean of all the classes.

Let $m_T(k) = \frac{\Sigma_{i=1}^{n} m_i(k)}{n}$, where n = number of classes

and $\text{var\_mod}(k) = \frac{1}{n}\sum(m_i(k) - m_T(k))$ so that $\sum(m_i(k) - m_T(k))^2$ is not sufficient to quantify the classification contribution of the coefficients because it may give the same values in the two cases. Therefore, there is a need to introduce another metric to quantify the coefficients' contribution. We introduce another metric as follows:

$$\text{Var\_mod}(k) = \min_{i \neq j} \left| \frac{m_i(k) - m_j(k)}{\sqrt{\frac{s_i^2(k)}{n_i(k)} + \frac{s_j^2(k)}{n_j(k)}}} \right| \qquad (2)$$

where S_i is the statistical metric of class i, m_i is the mean of class i, and n_i is the number of classes. S_i is calculated using the following formula:

$$S_i^2 = \frac{\sum_{j=1}^{n_i} \left(x_j^i(k) - m_i(k)\right)^2}{n_i - 1}, \text{where } i = 1,2,3,\ldots,n_i$$

Where n_i is the number of features in class i.

If the statistical metric of any feature is less than a certain threshold, it is removed. Those kept features / coefficients are sent to the classification step.

3) Classification

Cluster-k-Nearest Neighbor (C-k-NN) classifier is utilized in this CAD system. The C-k-NN is a classifier that merges two algorithms that are K-means modified algorithm [13] and k-Nearest neighbor algorithm. The K-means algorithm is utilized to cluster the data into classes and sub-classes with a central point to represent each class and sub-class while k-Nearest Neighbor is used to classify new data by calculating the Euclidean distance between the center point of each class and the new data. With this algorithm, the advantage of k-means of calculating in less time and the advantage of k-NN, which is the accurate classification, are combined. A full mathematical explanation of the algorithm can be found in [14].

III. Experiments and Results

In this paper, the CAD system is applied in two main experiments, classification of normal from abnormal cases and benign from malignant cases. In each experiment, different scenarios of combining the coefficients of wavelets and curvelet are applied as follows:

- Wavelets transform (haar function).
- Combination of the coefficients of two levels of haar wavelet.
- Combination of haar and db1 coefficients.
- Curvelet transform at scales 2-7.
- Combination of haar wavelet and Curvelet coefficients.

For the combination of different wavelet levels or wavelet and curvelet, the coefficients were extracted separately and then combined in one matrix and randomized. After that, they were subjected to the feature selection and classification methods.

In the experiments, the sub-images (128x128) are used in the evaluation of the CAD system. Those images are divided into training and testing sets. The training set was used to train the classifier while the testing set was used to assess the classification performance.

A. Normal Vs. Abnormal

For this experiment, Tables (II - V) show the obtained results.

TABLE II. Normal vs. Abnormal Classification with Wavelet

Performance	Level	Haar wavelet
Accuracy	1	0.9829
False Negative		0
False Positive		0
Accuracy	2	0.9829
False Negative		0
False Positive		0
Accuracy	3	0.9915
False Negative		0
False Positive		0
Accuracy	4	0.9915
False Negative		0
False Positive		0
Accuracy	5	0.9915
False Negative		0
False Positive		0
Accuracy	6	0.9915
False Negative		0
False Positive		0

TABLE III. Normal vs. Abnormal Classification with Combined Wavelets

Performance	Haar_1 + Haar_6	Haar_5 + db1_6
Accuracy	0.9915	0.9915
False Negative	0	0
False Positive	0	0

TABLE IV. Normal vs. Abnormal Classification with Curvelet

Performance	Scale	Curvelet
Accuracy	2	0.9402
False Negative		0.0270
False Positive		0
Accuracy	3	0.9573
False Negative		0.0270
False Positive		0.0233
Accuracy	4	0.9658
False Negative		0.0270
False Positive		0.0233
Accuracy	5	0.9658
False Negative		0.0270
False Positive		0
Accuracy	6	0.9658
False Negative		0.0405
False Positive		0.0233
Accuracy	7	0.9658
False Negative		0
False Positive		0

TABLE V. Normal vs. Abnormal Classification with Combined Wavelet and Curvelet

Performance	Scale -Level	Curvelet with Wavelet
Accuracy	Curvelet_2 + Haar_6	0.7863
False Negative		0.1757
False Positive		0.1395
Accuracy	Curvelet_3 + Haar_6	0.9915
False Negative		0.0135
False Positive		0.0233
Accuracy	Curvelet_4 + Haar_1	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_5 + Haar_3	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_6 + Haar_4	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_5 + Haar_3	0.9829
False Negative		0
False Positive		0

In this experiment, haar wavelet reached a maximum accuracy of 99.15% while curvelet reached 96.58%. However, after combining both wavelet with curvelet, the highest accuracy achieved is 99.15 which is better than the accuracy achieved by curvelet alone. Similar increase in the accuracy is also shown in the other experiment.

B. Benign vs. malignant

For this experiment, Tables (VI-IX) show the obtained results.

TABLE VI. BENIGN VS. MALIGNANT CLASSIFICATION WITH WAVELET

Performance	Level	Haar wavelet
Accuracy	1	0.9610
False Negative		0.0200
False Positive		0.0370
Accuracy	2	0.9610
False Negative		0.0200
False Positive		0.0370
Accuracy	3	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	4	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	5	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	6	0.9610
False Negative		0.0400
False Positive		0.0370

TABLE VII. BENIGN VS. MALIGNANT CLASSIFICATION WITH COMBINED WAVELETS

Performance	Haar _2 + Haar_ 6	Haar_2 +db1_1
Accuracy	1	1
False Negative	0	0
False Positive	0	0

TABLE VIII. BENIGN VS. MALIGNANT CLASSIFICATION WITH CURVELET

Performance	Scale	Curvelet
Accuracy	2	0.8182
False Negative		0.0800
False Positive		0.0741
Accuracy	3	0.8701
False Negative		0.1200
False Positive		0.0741
Accuracy	4	0.8701
False Negative		0.0600
False Positive		0.0370
Accuracy	5	0.8571
False Negative		0.0200
False Positive		0
Accuracy	6	0.8831
False Negative		0.0800
False Positive		0.0741
Accuracy	7	0.8701
False Negative		0.1000
False Positive		0.0741

TABLE IX. BENIGN VS. MALIGNANT CLASSIFICATION WITH COMBINED WAVELET AND CURVELET

Performance	Scale - Level	Curvelet with Wavelet
Accuracy	Curvelet_2 + Haar_6	0.8571
False Negative		0.0400
False Positive		0.0370
Accuracy	Curvelet_3 + Haar_3	0.8961
False Negative		0.0600
False Positive		0.0370
Accuracy	Curvelet_4 + Haar_4	0.9091
False Negative		0.0600
False Positive		0.0741
Accuracy	Curvelet_5 + Haar_6	0.9091
False Negative		0.0800
False Positive		0.0741
Accuracy	Curvelet_6 + Haar_6	0.9091
False Negative		0.0200
False Positive		0
Accuracy	Curvelet_2 + Haar_6	0.9221
False Negative		0.0600
False Positive		0.0741

The results illustrated in the previous tables shows that, the combination of different wavelets and the combination of wavelet and curvelet have improved the classification rate. Combining different wavelets as shown in table VII increased the accuracy to reach 100%. On the other hand, combining curvelet with wavelet increased the curvelet performance as shown in Tables V and IX.

IV. CONCLUSION

The paper investigates the combination of different wavelet functions as well as the combination of wavelet and coverlet in cancer classification. To our knowledge, no previous study has examined this combination. As shown in the previous section, the combination has potential in increasing the classification accuracies. Further tests and experiments are to be carried out in different datasets as well as different cancer types.

REFERENCES

[1] *National Cancer Institute*. Available: http://www.cancer.gov/

[2] F. J, *et al.* (2010). *GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 10 [Internet]*. Available: http://globocan.iarc.fr, accessed on 01/02/2013.

[3] W. C. R. F. International. *Cancer Statistics: Worldwide*. Available: http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php#Both

[4] J. R. F. da Silva Sousa, *et al.*, "Methodology for automatic detection of lung nodules in computerized tomography images," *Computer Methods and Programs in Biomedicine,* vol. 98, pp. 1-14, 2010.

[5] S. L. A. Lee, *et al.*, "Random forest based lung nodule classification aided by clustering," *Computerized Medical Imaging and Graphics,* vol. 34, pp. 535-542, 2010.

[6] J. Zhang, *et al.*, "Lung nodule classification combining rule-based and SVM," in *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on*, 2010, pp. 1033-1036.

[7] H. M. Orozco, *et al.*, "Lung nodule classification in frequency domain using support vector machines," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, 2012, pp. 870-875.

[8] S. A. Kumar, *et al.*, "Robust and Automated Lung Nodule Diagnosis from CT Images Based on Fuzzy Systems," in *Process Automation, Control and Computing (PACC), 2011 International Conference on*, 2011, pp. 1-6.

[9] H. Chen, *et al.*, "Classification of Pulmonary Nodules Using Neural Network Ensemble," in *Advances in Neural Networks – ISNN 2011*. vol. 6677, D. Liu, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 460-466.

[10] J. Shiraishi, *et al.*, "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule," *American Journal of Roentgenology,* vol. 174, pp. 71-74, January 1, 2000 2000.

[11] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 11, pp. 674-693, 1989.

[12] E. Candès, *et al.*, "Fast Discrete Curvelet Transforms," *Multiscale Modeling & Simulation* vol. 5, pp. 861–899, 2006.

[13] B. B. Samir, "Modified k-means cluster," Universiti Teknologi PETRONAS2008.

[14] B. B. Samir, "Fast and Accuracy Control Chart Pattern Recognition using a New cluster-k-Nearest Neighbor," *Journals of World Academy of Science, Engineering and Technology,* vol. 25, 2009.