Sci.Int. (Lahore), 23(3),201-203, 2011 ISSN 1013-5316; CODEN: SINTE 8

AN AUTOMATED SYSTEM FOR CLASSIFYING COMPUTED TOMOGRAPHIC (CT) LUNG IMAGES USING MARKOV MODEL

Hanan M. Amer[1], Fatma E.Z. Abou-Chadi[2], Marwa Obayya[3]
e-mail: hanan.amer@yahoo.com
[1]Facility of Engineering Communication Department, Mansoura University, Egypt
[2]Dept. of Electronic and Communication Engineering, Facility of Engineering, Mansoura University, Egypt
[3]Electronic and Communication Engineering, Facility of Engineering, Mansoura University, Egypt

ABSTRACT: *In this paper, the computed topographic (CT) images were utilized to discriminate different lung abnormalities. The performance of traditional feature extraction techniques and wavelet-domain based estimators was compared using Hidden Markov Model (HMM). This was done by analyzing data recorded for healthy subjects and patients suffering from lung cancer and emphysema diseases. The techniques utilized included statistical, intensity, and morphological features as well as features derived from texture analysis, and wavelet domain-based features. Results have shown that using wavelet domain estimators gives higher rates for classifying lung abnormalities. Classification rate reaches about 99.3%.*

Keywords: Computed Tomography (CT); Texture Feature, and Morphology Feature; Feature Extraction and Feature Selection.

INTRODUCTION

In recent years great advances have been made in Computer Aided Diagnosis (CAD) systems for detecting disease from Computed Tomography (CT) scans, mainly due to the advances made in the scanning machines which allow a greater amount and quality of information to be extracted during a single breath of the patient. The use of feature extraction, textural analysis and pattern recognition techniques for classification is most suited to the evaluation of global conditions (e.g. cancer, Emphysema and normal), which we will be concern in this paper. This recent progress in CAD in Chest Radiology has been discussed in a Guest Editorial in IEEE Transaction on Medical Imaging in 2001 Giger M.L. (2001) [1], where it has been noted that the amount of 3-D image data from thoracic CT scans greatly increases the number of images that much be reviewed by the radiologist and therefore a search aid may be a great benefit.

The most familiar cancer that occurs usually for men and women is lung cancer. According to the report submitted by the American Cancer Society in 2003, lung cancer would report for about 13% of all cancer diagnoses and 28% for all cancer deaths. The survival rate for lung cancer analyzed in 5 years is just 15%. If the disease is identified while it is still localized, this rate increases to 49%. However, only 15% of diagnosed lung cancers are at this early stage. Where emphysema begins with the destruction of air sacs (alveoli) in the lungs where oxygen

from the air is exchanged for carbon dioxide in the blood. As air sacs are destroyed, the lungs are able to transfer less and less oxygen to the bloodstream, causing shortness of breath. Air pollution, dust or chemicals and smoking can be consider the main factor that causes the last lung diseases.

Previous work in this area has involved training a machine learning algorithm using statistical, textural and/or morphological features of common patterns extracted from CT scans and documenting the success of classifying these patterns correctly. Uppaluri et al. [2] developed the Adaptive Multiple Feature Method (AMFM) to recognize honeycombing, ground glass, bronchovascular, nodular, emphysema like, and normal tissue patterns on the basis of their textures. A training and test set were assembled from 72 patients scans, 20 normal, 13 with emphysema, 19 with IPF, and 20 with sarcoidosis, broken down into a grid of 31 x 31 pixel blocks so 22 independent texture features could be extracted from each block. This system produced an overall accuracy of 93.5% for the test set. More recently, Uchiyama et al. [3] selected regions in 315 HRCT images from 105 patients, relating to six different patterns, i.e., ground-glass opacities, reticular and linear opacities, nodular opacities, honeycombing, emphysematous change, and consolidation, labeled by 3 radiologists. The lungs were first segmented, using standard technique, then divided into many contiguous regions of interest (ROIs) with a 32x32 matrix and classified using artificial neural networks. The accuracy varied from 88 to 100%, with specificity in detecting a normal ROI of 88.1%. Most recent Sluimer et al. [4] presented a CAD system to automatically distinguish normal from abnormal tissue in HRCT chest scans of 116 patients, producing 657 ROIs labelled as containing normal or abnormal tissue. The circular ROIs with an 80-pixel diameter were extracted from the peripheral lung region in slices at the height of the aortic arch, with each ROI required to contain at least 75% abnormal tissue. An accuracy of 86.2% as obtained, comparable to those of a radiologist when evaluating only the ROIs, i.e. without seeing the whole scan.

The aim of the present study is to develop an automated system for classifying CT lung images by utilizing a number of traditional feature extraction techniques and the wavelet-domain estimators as well as features derived from texture analysis in the design of Hidden Markov Model [5].

II. Data Collection

In the present study 762, 512×512×8 bit, images extracted from low-dose documented whole-lung CT scans (242 normal, 242 cancer, 242 emphysema). The CT scans were obtained in a single breath hold with a 1.25 mm slice thickness. They were obtained from the early lung cancer action project (ELCAP) association [6]. This database was made possible by collaboration between the ELCAP and VIA research groups. It was created to make available common dataset that may be used for the performance

evaluation of different computer aided detection systems. This database was first released in December 2003 and is a prototype for web-based image data archives.

Three cases of human lung diseases; these cases are described in the following table. All images are available in digital images and communication in medicine (DICOM) format; or shortly (DCM), as shown in table-1. An example of the acquired images is shown in Fig.1.

Table 1. the CT lung images cases

Case	Size	Resolution	Image Format
Normal	512×512×242	0.76×0.76×1.25	DCM
Cancer	512×512×242	0.76×0.76×1.25	DCM
Emphysema	512×512×242	0.76×0.76×1.25	DCM

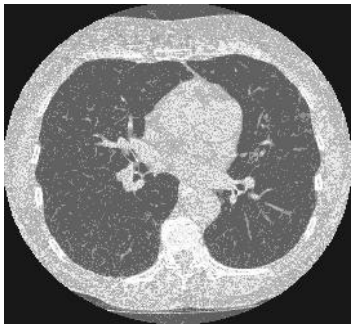

Fig.1. Raw chest CT image

III. METHODS

The proposed system that processes and classifies automatically the digital CT images of the human lung consists of four main parts as shown in Fig.2.

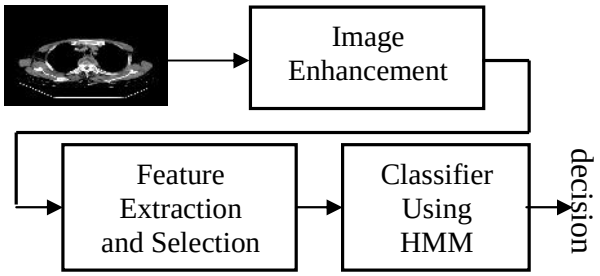

Fig. 2, The block diagram of the proposed system

A. Image Enhancement.

The first step needed in human lung CT images is to increase the quality and contrast. This was performed using a Wiener filter with mask size 7×7 that gives the best enhancement result visually [7].

B. Feature Extraction.

The second step is used to extract the most salient features of digital CT images and to reduce the dimensionality of acquired data. There are several reasons for performing feature extraction: (i) to reduce the bandwidth of the input data (with the resulting improvements in speed and reductions in data requirements); (ii) to provide a relevant set of features for a classifier, resulting in improved performance, particularly from simple classifiers; (iii) to reduce redundancy; (v) to recover new meaningful underlying variables or features that structure in the data identified [8].

Five sets of features were utilized and their performance was compared. The techniques used to calculate Features are: (1) statistical-based features, where six parameters were calculated mean, variance, standard deviation, skewness, median, and kurtosis, (2) intensity-based features, which give a set of features in the spatial domain calculating first ten maximum pixel intensity and their location, smallest ten minimum pixel intensity and their location, and number of zero-crossing. (3) morphological-based features, which give a set of features based on the object shape (area, centriod, euler number, major-axis-length, mainor-axis-length, bounding box, extreme, orientation, eccentricity, convex hull and solidity), (4) texture-based features that refers to the characterization of regions in an image by their texture content. These are entropy of grayscale image, and contrast, correlation, energy and homogeneity of the create gray-level co-occurrence matrix from image. And, (5) wavelet domain-obtained by using (Daubchies-4) wavelet function.

C. Feature selection

The third step is used to select those features that contain the most discriminatory information. Alternatively, we may wish to limit the number of features we make, perhaps on grounds of cost, or we may want to remove redundant or irrelevant information to obtain a less complex classifier, this is described by using Principal Component Analysis (PCA). The purpose of principal component analysis is to derive new features vectors (in decreasing order of importance) that are linear combinations of the original features vectors (or matrixes) and are uncorrelated [9].

The technique has three effects: (i) it orthogonalizes the components of the input vectors (so that they are uncorrelated with each other), (ii) it orders the resulting orthogonal components (principal components) so that those with the largest variation com first, (iii) and it eliminates those components that contribute the least to the variation in the data set [10],[11].

D. Classifier Using Hidden Markov Model.

Hidden Markov Model (HMM) is a tool to statistically model a process that varies in time [6] the main characteristics of the HMMs are as follows:

1. The set of the possible hidden states s=$\{ s_1,....,s_l \}$, where *l* is the number of hidden states in the model.
2. The transition matrix *A* whose elements a_{ij} represent the probability to go from state s_i to state s_j.
3. $V = \{v_1,...,v_M\}$, where *M* is the number of distinct observation (emission) symbols per state.
4. The emission matrix *B* whose elements b_{jk} indicate the probability of emission of symbol v_k when the system state is s_j.

5. The set of initial state probability distribution $\Pi = \{\pi_1,...,\pi_l\}$ whose elements π_i represent the probability for s_i to be the initial state. For convenience, we denote an HMM as a compact notation $\lambda = \{A, B, \Pi\}$.

The Baun-Welch algorithm was used to train HMMs, one for each lung diseases, using the training data set [12]. For these calculations the HMM toolbox for Matlab was used. Thus based on the different number of hidden states l= {4,8,12}, the statistical modeling resulted in three models for each lung disease.

In order to use the foregoing algorithms for pattern classification, it is simply necessary to define a separate model for each class patterns. Maximum likelihood classification of an unknown observation sequence can be achieved by calculating the probability of the observations given the model $P(O/\lambda)$ for each model in turn. The unknown pattern is assigned to the class of the model that has the highest probability of generating the observed data; that is for E classes C= $c_1, c_2,..., c_E$, where c_E is represented by model λ_E, then O is assigned to class c_E if $P(O/\lambda) = \max_{m=1}^{E} P(O/\lambda_m)$ [12].

IV. RESULTS

To evaluate the performance of the proposed classifier, three measures are used and defined as follows

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$Specifficity(\%) = \frac{TN}{TN + FP} \times 100 \quad (2)$$

$$Accuracy(\%) = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \quad (3)$$

where TP, TN, FN, FP stands for true positive, true negative, false positive and false negative respectively.

The results of the classification procedure are presented in table-2, table-3, table-4, table-5 and table-6. We have 726 images (242 images for normal, 242 images for cancer and 242 for emphysema). We trained HMM with the first 363 images and tested it with the other 363 images. It has been shown that HMM with no., of states twelve was able to characterize differences between normal patients and patients suffering from cancer and emphysema diseases.

Table 2. Sensitivity, Specificity and Accuracy of the classifier using statistical parameters

Case	*Sensitivity*	*Specificity*	*Accuracy*
Normal	100%	97.18%	98.15%
Cancer	89.5%	93.67%	95.50%
Emphysema	96%	98.28%	96.25%
Average	95.16%	96.37%	96.63%

Table 3. Sensitivity, Specificity and Accuracy of the classifier using intensity parameters

Case	*Sensitivity*	*Specificity*	*Accuracy*
Normal	100%	96.18%	97.15%
Cancer	90.5%	94.67%	96.50%
Emphysema	97%	97.28%	98.25%
Average	95.83%	96.04%	97.3%

Table 4. Sensitivity, Specificity and Accuracy of the classifier using morphological parameters

Case	*Sensitivity*	*Specificity*	*Accuracy*
Normal	100%	96.18%	97.15%
Cancer	90.5%	94.67%	96.50%
Emphysema	97%	97.28%	98.25%
Average	95.83%	96.04%	97.3%

Table 5. Sensitivity, Specificity and Accuracy of the classifier using texture parameters

Case	*Sensitivity*	*Specificity*	*Accuracy*
Normal	100%	97.18%	98.15%
Cancer	95.36%	97.67%	96.50%
Emphysema	96%	95.28%	98.25%
Average	97.12%	96.58%	97.63%

Table 6. Sensitivity, Specificity and Accuracy of the classifier using wavelet parameters

Case	*Sensitivity*	*Specificity*	*Accuracy*
Normal	100%	97.18%	98.15%
Cancer	95.36%	97.67%	96.50%
Emphysema	96%	95.28%	98.25%
Average	97.12%	96.58%	97.63%

IIV. CONCLUSION

This presents the development of an automated system for classifying CT lung images. The system consists of four main parts: image enhancement, feature extraction, feature selection and classification. Image enhancement was carried out using wiener filter. Six sets of features were extracted using six different feature extraction techniques. These are: statistical-based, intensity-based features, morphological-based features, and the usage was achieved using the principal component analysis approach. HMM with different orders (4, 8 and 12) was presented as a diagnostic tool to aid physicians in the classification of lung diseases. It is shown that the HMM best distinguishing between lung

abnormalities was a model based on wavelet-domain features with twelve hidden states. With this setting average classification rate reaches 99.3%.

REFERENCES

[1] Giger, M.L., N. Karssemeijer, and S.G. Armato, Guest editorial computer- aided diagnosis in medical imaging, pp. 1205-1208.

[2] R. Uppaluri, E. A.Hoffman, M. Sonka, P. G.. Hartley, G. W. Hunningshake and G. McLennan, *Computer recognition of regional lung disease patterns.*, Pp. 648-654, 1999.

[3] Y. Uchiyama, S. Katsuragawa, H. be, J.Shiraishi, F.Li, Q. Li, C.-T. Zhang, K. Suzuki and K. Doi. Quantitive computerised analysis of diffuse lung disease in high-resolution computed tomography, *Med. Phys*., **30**: (9) September, pp. 2440-2453, 2003.

[4] Sluimer IC, van Waes PF, Viergever MA, van Ginneken B. Computer-aided diagnosis in high resolution CT of the lungs, *Med Phys*., **30** (12) December, pp. 3081-90, 2003.

[5] Eickeler,S., Kosmala, A., Rigoll,G., " Hidden Markov Model based online gesture recognition", *Proc. Int. Conf. on Pattern Recognition* (ICPR), PP.1755-1757,1998.

[6] http://www.via.cornell.edu/lungdb.html.

[7] Gonzales R.C. and Woods R.E., Digital Image Processing, 2nd Edition, New Jersey, Prentice Hall, 2004.

[8] Moulin P. and Liu J., "analysis of multi-resolution image denoising schemes using generalized Gaussian and complexity priors", *IEEE Information Theory*, Vol. **45**, No. 3, PP: 909-919, April-1999.

[9] Webb, A.R., "*Statistical Pattern Recognition*", John Wiley & Sons Ltd., 2002.

[10] www.mathworks.com.

[11] Chen, X., and A. Yuille, "*Detecting and reading text in natural sense*". CVPR. Volume:**2**, PP. 366-373, 2004.

[12] Eickeler, S., Kosmala, A., Rigoll, G., "Hidden Markov Model Based online gesture recognition", *Proc. Int. Conf. on Pattern Recognition(*ICPR), PP.1755-1757,1998.