İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl: 6 Sayı:11 Bahar 2007/1 s. 17-30

RULE-BASED LUNG REGION SEGMENTATION AND NODULE DETECTION VIA GENETIC ALGORITHM TRAINED TEMPLATE MATCHING

Serhat ÖZEKES[*]

ABSTRACT

A computer-aided detection (CAD) system was developed for detecting lung nodule patterns, which generally appear as circular areas of high opacity on serial-section Computed tomography (CT) images. First of all, rule-based segmentation of lung region was performed. Then, our method detected the regions of interest (ROIs) using the density values of pixels in CT images and scanning the pixels in 8 directions by using various thresholds. Then to reduce the number of ROIs the amounts of change in their locations based on the upper and the lower slices were examined, and finally a nodule template based algorithm was employed to categorize the ROIs according to their morphologies. To calculate the parameters of the template, a genetic algorithm process was employed as an optimization method. To test the system's efficiency, we applied it to 276 normal and abnormal CT images of 12 patients with 153 nodules. The experimental results showed that the system achieved 93.4% sensitivity with 0.594 false positives per image.

Keywords: Lung Region Segmentation, Template Matching, Genetic Algorithm, Lung Nodule Detection, Computer Aided Detection

AKCİĞER BÖLGESİNİN KURAL TABANLI BÖLÜTLENDİRİLMESİ VE GENETİK ALGORİTMA KULLANILARAK EĞİTİLMİŞ ŞABLON EŞLEME YÖNTEMİYLE NODÜL TESİPİTİ

ÖZET

Bu çalışmada akciğer bilgisayarlı tomografi (BT) görüntülerindeki nodüllerin bilgisayarlı tespiti gerçekleştirilmiştir. Öncelikle akciğer bölgesinin kural tabanlı bölütlendirilmesi gerçekleştirilmiştir. Ardından BT görüntülerindeki yoğunluk değerleri ve eşik değerleri ile 8 yönlü tarama yapılarak ilgi alanları belirlenmiştir. İlgi alanlarının sayısını azaltmak amacıyla alt ve üst kesitlerdeki konum değişimleri incelenmiştir. Son olarak şablon eşleme tabanlı bir yöntem ile ilgi alanları şekilsel özelliklerine göre sınıflandırılmıştır. Şablonun değerlerinin hesaplanması için genetik algoritma yöntemi kullanılmıştır. Çalışmanın test edilmesi amacıyla 153 adet nodül bulunan 12 hastaya ait 276 normal ve anormal görüntü kullanılmıştır. Sonuçta duyarlılığın görüntü başına 0.594 yanlış pozitif oranıyla %93.4'a ulaştığı görülmüştür.

Anahtar Kelimeler: Akciğer Bölge Tespiti, Şablon Eşleme, Genetik Algoritma, Akciğer Nodül Tespiti, Bilgisayar Destekli Tespit

[*] *Istanbul Commerce University, Vocational School, Kucukyalı, Istanbul*

1. INTRODUCTION

The mortality rate for lung cancer is higher than that for other kinds of cancers around the world (Greenlee et al., 2000). Of all the types of cancer, lung cancer is the most common cause of death and accounts for about 28% of all cancer deaths. At the same time, it appears that the rate has been steadily increasing. No smoking is considered the most effective way to reduce the incidence of lung cancer in most countries, while detection of suspicious lesions in the early stages of cancer can be considered the most effective way to improve survival.

Serial section CT has been shown to increase lung tumor detection rates by 300-500%, compared with radiologists' results using only projection chest X-rays, and the average size of tumors detected has been dramatically reduced from 30 mm to 12 mm. Whereas a thoracic CT scan using a single detector scanner typically generates 40 to 100 axial image slices, the newer, multi-detector scanners typically generate 300 to 600 image slices. To read and interpret these massive amounts of image data requires significant radiologist effort and predisposes the screening process to human error and missed detection of cancerous lesions. Thus, computer-aided diagnostic (CAD) approaches are becoming increasingly necessary for both reducing radiologists' effort and improving detection sensitivity.

Various CAD methods have been proposed to detect lung nodules. Giger *et al.* obtained nodules using multiple gray-level thresholding and a rule-based approach (Giger *et al.*, 1994). Armato *et al.* introduced some 3D features, and performed feature analysis by a linear discriminant analysis (LDA) classifier (Armato *et al.*, 1999). Kanazawa *et al.* used fuzzy clustering and a rule-based method (Kanazawa *et al.*, 1998). Penedo *et al.* set up 2 Neural networks (NNs), with the first one detecting the suspected areas, and the second one acting as a classifier (Penedo *et al.*, 1998). For example, Xu *et al.* describe a system which, following proper radiogram preprocessing, utilizes a set of decision rules and a feed forward neural network to find nodular patterns (Xu *et al.*, 1997). Following a different approach, Lo *et al.* propose a two-stage system: the first one locates possible nodular patterns (thus performing a sort of attention focusing process) while the second, implemented by a convolutional neural network, discriminates nodules from non-nodules (Lo *et al.*, 1995). A prior model was developed by Brown *et al.* to find nodules on the baseline scan and located nodules in the follow up scans (Brown *et al.*, 2001).

In this study, we designed a CAD system for lung nodule detection in CT images. First, rule-based segmentation of the lung region was performed in order to decrease the number of ROIs and computation time. Using the density values of pixels in image slices and scanning these pixels in 8 directions with distance thresholds, ROIs were found. In order to classify the ROIs, a location change thresholding was used followed by a template matching based algorithm. The parameters of the template were calculated using GA. Hence, the true lung nodules were detected successfully.

2. MATERIALS AND METHODS

The primary challenge for radiologists and CAD systems alike for lung tumor detection is that, in serial sectional images, there are many objects that have the same appearance and pixel intensity as tumor nodules. In a serial-section slice, a cylindrical vessel can appear circular, and many vessels in the lung have a similar diameter to the lesions of interest. A primary failing point of all the CAD systems referenced above is that they depend upon a first-pass detection of candidates based on 2D image features, producing hundreds of first-pass candidates. The CAD systems then employ various schemes to tackle the enormous task of removing likely false positives from the vast candidate pool, with varying degrees of success. A common problem is that in filtering out the large volume of false positives, true positives are also omitted; creating a system that is prone to missing true tumors yet maintains a relatively high false positive count.

The approach described herein was motivated by the observation that experienced radiologists screen for lung lesions not by considering individual image slices independently, but by paging through the image stack looking for 3D appearance characteristics that distinguish tumors from vessels. On consecutive images, vessels maintain a similar cross-sectional size and their in-plane circular appearance appears to drift across the viewing screen from one slice to the next, following the tortuous anatomy of the vessel. True lung nodules, in contrast, appear seemingly out of nowhere as circular objects that remain at approximately the same on-screen location from slice to slice. Their size quickly increases and then just as rapidly decreases and the tumor disappears after a few slices.

We developed a CAD scheme by computer programs that could prepare quantitative values and the position of lesions to radiologists. If radiologists take into account the information obtained from our CAD system, their diagnostic performance would be higher. Figure 1 shows the overview of our CAD system.

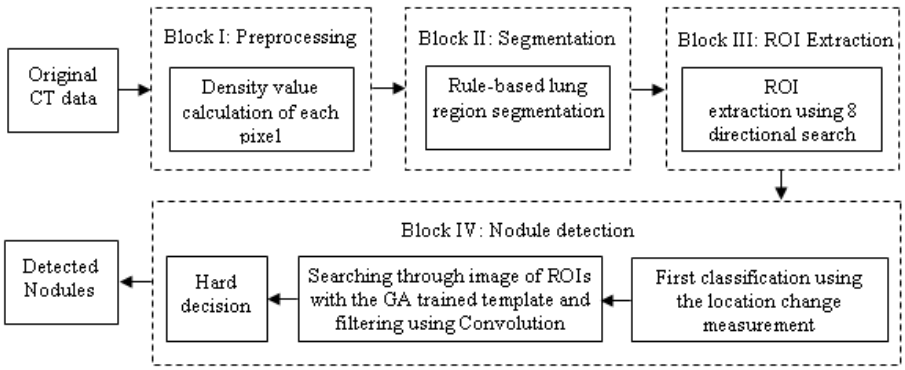

Figure 1. Procedural Flowchart for Detecting Lung Nodules

For the development and evaluation of the proposed system we used the Lung Image Database Consortium (LIDC) database (Samuel et al, 2004). In the LIDC database for each nodule, 6 expert lung radiologists provided annotations, i.e. segmentations, of all nodules using 3 different methods, one manual and two automatic, for a total of 18 possible radiologist/method combinations. Each CT slice used in this study is formed of a 16 bit ASCII coded matrix with dimensions 512 X 512. These ASCII codes are related with the density values of each pixel in slices. Using these ASCII codes density values in Hounsfield units (HU) are calculated. HU is a unit of x-ray attenuation used for CT scans, each pixel being assigned a value on a scale on which air is -1000, water is 0, and compact bone is +1000 (Hounsfield, 1980). When the dataset was examined, it was determined that density values of the nodules were between -500 HU and 100 HU, called as *"minimum density threshold"* and *"maximum density threshold"* values respectively.

2. 1. Rule-Based Lung Region Segmentation

Here are the steps of lung region segmentation:

Step 1: Thresholding. As seen in Figure 2a shapes like nodules, bones and vessels are brighter than other structures. This means they have higher HU values. Thus thresholding was performed in order to extract the lung region roughly. If *I(x, y)* is the input image seen in Figure 2a, by applying the following rule Figure 2b was achieved. Here 1 represents white and 0 represents black.

IF I(x, y) < -500 HU THEN
I(x, y) = 1
ELSE
I(x, y) = 0

Step 2: Labeling and small black shape elimination. As seen in Figure 2b the white lung region contains small black shapes representing nodules and vessels. To eliminate them we label the black shapes using connected component labeling (CCL) and analyze their sizes.

When all black shapes in the lung region seen in Figure 2b were labeled, their sizes were analyzed using the following rule. If *S(k)* representing the size of *kth* black shape in pixels is smaller than 150 pixels, we assign 1 to the *I(x, y)* which is representing the pixels of *kth* shape of the image. Thus by eliminating the small black shapes in Figure 2b, Figure 2c was achieved.

IF S(k)<150 pixels THEN I(x, y)=1

Step 3: Labeling and lung region extraction. To extract the lung region in Figure 2c, all white shapes were labeled using CCL and their morphologies were analyzed. As seen in Figure 2c the aspect ratio lung region is smaller than the other shapes. The morphologies of the shapes were analyzed using the following rule where *AR(k)* represents the aspect ratio of *kth* shape which is the ratio of height to width, *W(k)* represents the width of *kth* shape and *I(x, y)* represents the pixels of *kth* shape of the image. According to this rule the shapes, whose aspect ratios were less then 1.5 or width were less then 50 pixels, were eliminated and thus the lung region was extracted.

IF AR(k)>1.5 OR W(k)<50 pixels THEN I(x, y)=0

At the end of the third step the lung region of Figure 2a was segmented as Figure 2d and Figure 2e.

2.2. Regions of Interest Specification Methods

Pixels, which form the candidate lung nodule region, must be members of a set of adjacent neighbor pixels with densities between *"minimum density threshold"* and *"maximum density threshold"* values. It has been observed that diameters of lung nodules are between upper and lower boundaries. So, to understand whether a pixel is in the center region of the shape, first, diameter of the shape (assuming the pixel in question is the center) should be considered. In this stage, we introduce two thresholds which form the boundaries. As introduced in (Ozekes et al, 2005), one is the "*minimum distance threshold*" representing the lower boundary and the other is the "*maximum distance threshold*" representing the upper boundary.

If a pixel has adjacent neighbors that are less than "*minimum distance threshold*" or more than "*maximum distance threshold*" in 8 directions, it could be concluded that this pixel couldn't be a part of candidate lung nodule. Otherwise, it could be a part of candidate lung nodule. The values of minimum and maximum distance thresholds are dealt with the resolution of the CT image. These thresholds are used to avoid very big or very small structures such as parts of chest bones or heart and vertical vessels. Thus the black and white image of ROIs was obtained showing ROIs in black.

2.3. First Classification Using the Location Change Measurement

On serial images, vessels maintain a similar cross-sectional size and their in-plane circular appearance changes its location. But the true lung nodules, remain at approximately the same on-screen location from slice to slice.

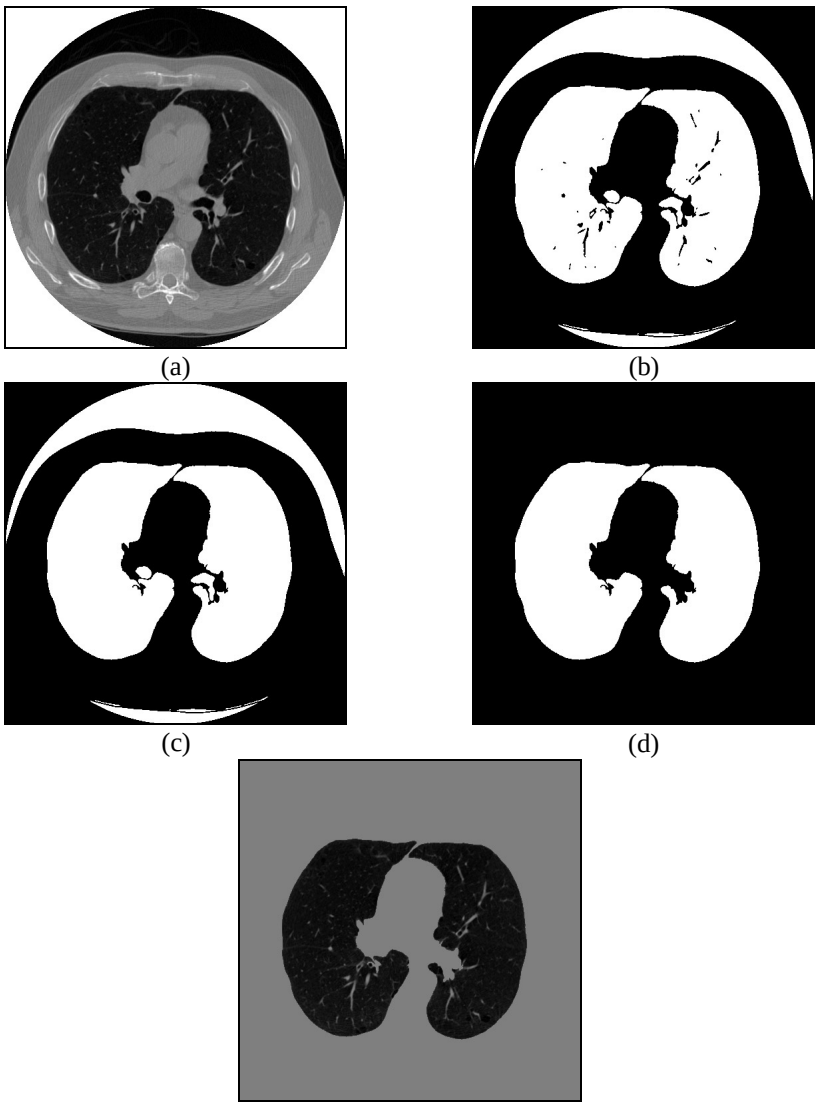

Figure 2. Steps of Rule-Based Lung Region Segmentation, a) The Original CT Image, b) Thresholding, c) Labeling and Elimination of Small Black Shapes within The Lung Region, d) Labeling and Lung Region Extraction, e) Lung Region of The Original CT Image

In this step, to classify the ROIs the proposed CAD system measures the amount of location changes of ROIs in serial sections. To specify the positions, ROIs of the serial sections are labeled using connected component labeling. When a ROI in a CT slice is analyzed, it's all Euclidian distances to the ROIs in the upper and the lower slices are calculated. The minimums of these distances ($Ed_{minupper}$ and $Ed_{minlower}$) are compared with the *"location change threshold"* and each of the candidates is classified according to the following decision rule introduced in (Ozekes ve Camurcu, 2006).

IF $Ed_{minupper} > T_{lc}$ *OR* $Ed_{minlower} > T_{lc}$ *THEN*
    normal
ELSE
    nodule candidate

where $Ed_{minupper}$ and $Ed_{minlower}$ are the amounts of location change of the ROI based on the upper slice and the lower slice respectively. T_{lc} is a threshold value. If the ROI was classified as a normal structure then it was removed. Thus the new image of ROIs was obtained with reduced number of ROIs.

2.4. Second Classification Using the Genetic Algorithm Trained Template Matching Method

To distinguish true lung nodules from normal structures by using their morphologies we used lung nodule templates. Each pixel of CT images was scanned with nodule templates and was looked for whether there was a shape similar to the nodule in the template, so too small, too thin and too long shapes were removed. If a similar one was detected then appropriate pixels of the shape are recorded as a part of a true lung nodule. The GA process was employed as an optimization method to calculate the parameters of the template.

The genetic algorithm is a method for solving optimization problems that is based on natural selection and is inspired by Darwin's theory about evolution. Algorithm is started with a random initial set of solutions (represented by chromosomes) called population. Individual solutions from the current population are taken to be parents and used to form a new population for the next generation. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until the number of populations is satisfied. Over successive generations, the population evolves toward an optimal solution. The general outline of the genetic approach used in this paper can be summarized as follows:

1. Encoding of the problem in a binary string.
2. Random generation of a population.
3. Extraction of the nodule template.
4. Reckoning of a fitness value for each subject.
5. Selection of the subjects that will mate according to their share in the population global fitness.
6. Genomes crossover and mutations.
7. And then start again from point 4.
8. If the stopping criterion has been satisfied, stop and decode the individual with the highest fitness to obtain the template.

A. Population Representation and Initialization

Individuals are encoded as strings called *chromosomes* and composed of 0's and 1's. To calculate the parameters of a nodule template the number of variables has to be determined and each variable has to be encoded with appropriate bits. Having decided on the representation, the first step in the GA was to create an initial population randomly.

B. Extraction of the Nodule Template

Chromosomes represent the binary codes of the elements of the nodule template *T*. In this step, each chromosome was decoded and the elements of the template were computed in the interval [-1, 1]. The template *T* with dimensions 8 x 8 pixels is given below;

$$T = \begin{bmatrix} t_6 & t_7 & t_8 & t_9 & t_9 & t_8 & t_7 & t_6 \\ t_7 & t_3 & t_4 & t_5 & t_5 & t_4 & t_3 & t_7 \\ t_8 & t_4 & t_1 & t_2 & t_2 & t_1 & t_4 & t_8 \\ t_9 & t_5 & t_2 & t_0 & t_0 & t_2 & t_5 & t_9 \\ t_9 & t_5 & t_2 & t_0 & t_0 & t_2 & t_5 & t_9 \\ t_8 & t_4 & t_1 & t_2 & t_2 & t_1 & t_4 & t_8 \\ t_7 & t_3 & t_4 & t_5 & t_5 & t_4 & t_3 & t_7 \\ t_6 & t_7 & t_8 & t_9 & t_9 & t_8 & t_7 & t_6 \end{bmatrix} \quad (1)$$

As seen in (1), the template is symmetrical, thus the total number of variables was 10 that are represented with *S* and each element of S is being coded in binary;

$$S = \begin{bmatrix} t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 \end{bmatrix} \quad (2)$$

C. The Objective and Fitness Functions

The objective function is used to provide a measure of how individuals have performed in the problem domain. In this step, the image that was selected as the

training image was convolved with the template belonging to the first chromosome. This convolution aims to classify the nodule from the other structures in the input image. If *T(x, y)* is the template with dimensions *n x m* and *I(x, y)* is the input image with dimensions *M x N* then the convolution of *T* with *I* is written as,

$$C(x, y) = T * I(x, y) = \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} T(i, j) I(x - i, y - i) \tag{3}$$

Here * represents the convolution operation. The shapes, which are similar to the template, become stronger at the end of the convolution computation. Therefore, the pixel values of these shapes become high positive, while the pixel values of non-similar ones become high negative. Thus, to extract the nodules from the convolved image **C**, we use the following decision rule:

IF $C(x,y) > 1$ *THEN*
$C(x,y) = 1$
ELSE
$C(x,y) = -1$

After the convolution and the decision rule, the objective function was computed between this output image **C** and the desired target image **A**. This process was repeated with the template sets belonging to each chromosome in the population. The objective function has been selected in this study as follows:

$$obj(T) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} C_{i,j} \oplus A_{i,j} \tag{4}$$

where the symbol $\oplus$ stands for the XOR operation between respective pixels of ***C*** and ***A***. After finding the objective function, the associated fitness function was evaluated for each chromosome according to the rule:

$$fitness(T) = M \times N - obj(T) \tag{5}$$

In this study for the stopping criterion the following rule was defined:

$$stopcrt = 0.99 \times M \times N \tag{6}$$

If the maximum fitness value of the chromosome was greater than stopping criterion, the algorithm was stopped and the chromosome whose fitness value was the maximum in the population was selected. The template which has been extracted from this selected chromosome was the most proper template, which satisfied the task desired to be realized.

D. Crossover and Mutation

The genetic algorithm works by randomly selecting pairs of individual chromosomes to reproduce for the next generation. The probability of a chromosome being selected is proportional to its fitness function value relative to the other chromosomes in the same generation. To reproduce, a crossover procedure is defined. For crossover, in this study, an integer position, *i*, was selected uniformly at random between 1 and the string length, *l*, minus one [1, *l*-1], for the crossing site. Then the two chromosome strings were sliced at the site, and the two tail pieces are swapped and rejoined with the head pieces to produce two progenies.

This crossover operation was not necessarily performed on all strings in the population. Instead, it was applied with a probability *Px* when the pairs were chosen for breeding. A further genetic operator, called mutation, was then applied to the new chromosomes, again with a set probability *Pm*. Mutation caused the individual genetic representation to be changed according to some probabilistic rule. In the binary string representation, mutation caused a single bit to change its state, 0 to 1 or 1 to 0.

After recombination and mutation, the individual strings were then, if necessary, decoded, the objective function evaluated, a fitness value assigned to each individual and individuals selected for mating according to their fitness, and so the process continued through subsequent generations. In this way, the average performance of individuals in a population was expected to increase, as good individuals were preserved and bred with one another and the less fit individuals die out. The genetic algorithm was terminated when the stopping criterion has been satisfied.

3. RESULTS

For the evaluation of the proposed system we used the LIDC database (Samuel et al, 2004). To test the system's efficiency, we applied it to 123 normal and 153 abnormal slice images of 12 clinical cases with 153 nodules. The diameters of the nodules were between 3.5 and 7.3 millimeters and the thicknesses were between 5.625 and 18.75 millimeters.

ROI images were obtained using 8 direction searches with *minimum distance* of 1 pixel and *maximum distance* of 8 pixels. At first, 3656 ROIs were specified by the ROI specification methods. By the first classification using the location change measurement the number of ROIs was reduced to 967. And finally the second classification was performed using a template whose diameter was 8 pixels.

To calculate the parameters of the template the GA process was performed. At the end of the training processes, genetic algorithm parameters were obtained as in Table 1 and the template was found as in (7).

Table 1. Genetic Algorithm Training Parameters for Nodule Template Optimization

Parameters	**Values**
Number of Chromosomes per Population	100
Bits per Variable	8
Number of Variables	10
Chromosome Length	80
Total Bits in the Population	8000
Crossover Probability(*Px*) for breeding	70%
Mutation Probability	1%
Generation Gap	98%
Template Parameters Range	[-1, 1]

$$T = \begin{bmatrix} -0.15294 & 0.76471 & 0.84314 & -0.92941 & -0.92941 & 0.84314 & 0.76471 & -0.15294 \\ 0.76471 & 0.41176 & -0.11373 & -0.05098 & -0.05098 & -0.11373 & 0.41176 & 0.76471 \\ 0.84314 & -0.11373 & 0.27843 & 0.019608 & 0.019608 & 0.27843 & -0.11373 & 0.84314 \\ -0.92941 & -0.05098 & 0.019608 & 0.38039 & 0.38039 & 0.019608 & -0.05098 & -0.92941 \\ -0.92941 & -0.05098 & 0.019608 & 0.38039 & 0.38039 & 0.019608 & -0.05098 & -0.92941 \\ 0.84314 & -0.11373 & 0.27843 & 0.019608 & 0.019608 & 0.27843 & -0.11373 & 0.84314 \\ 0.76471 & 0.41176 & -0.11373 & -0.05098 & -0.05098 & -0.11373 & 0.41176 & 0.76471 \\ -0.15294 & 0.76471 & 0.84314 & -0.92941 & -0.92941 & 0.84314 & 0.76471 & -0.15294 \end{bmatrix} \quad (7)$$

After convolving the image of ROIs with the template *T* 143 ROIs were classified as nodules with 164 FPs. An example of the original CT image, the segmented lung region, the extracted ROIs and the ROI classified as nodule are shown in Figure 3a, 3b, 3c, 3d and 3e respectively. The experimental results showed that the system achieved 93.4% sensitivity with 0.594 FPs per image.

4. CONCLUSION

A new scheme has been proposed to automatically detect lung nodules in CT images. Rule-based lung segmentation was performed and the morphological analysis of nodules was facilitated using template matching technique. The templates used in these techniques were trained with genetic algorithms. The results showed that the proposed CAD system was an effective assistant for human experts to detect nodule patterns and provide a valuable “second opinion” to the human observer.

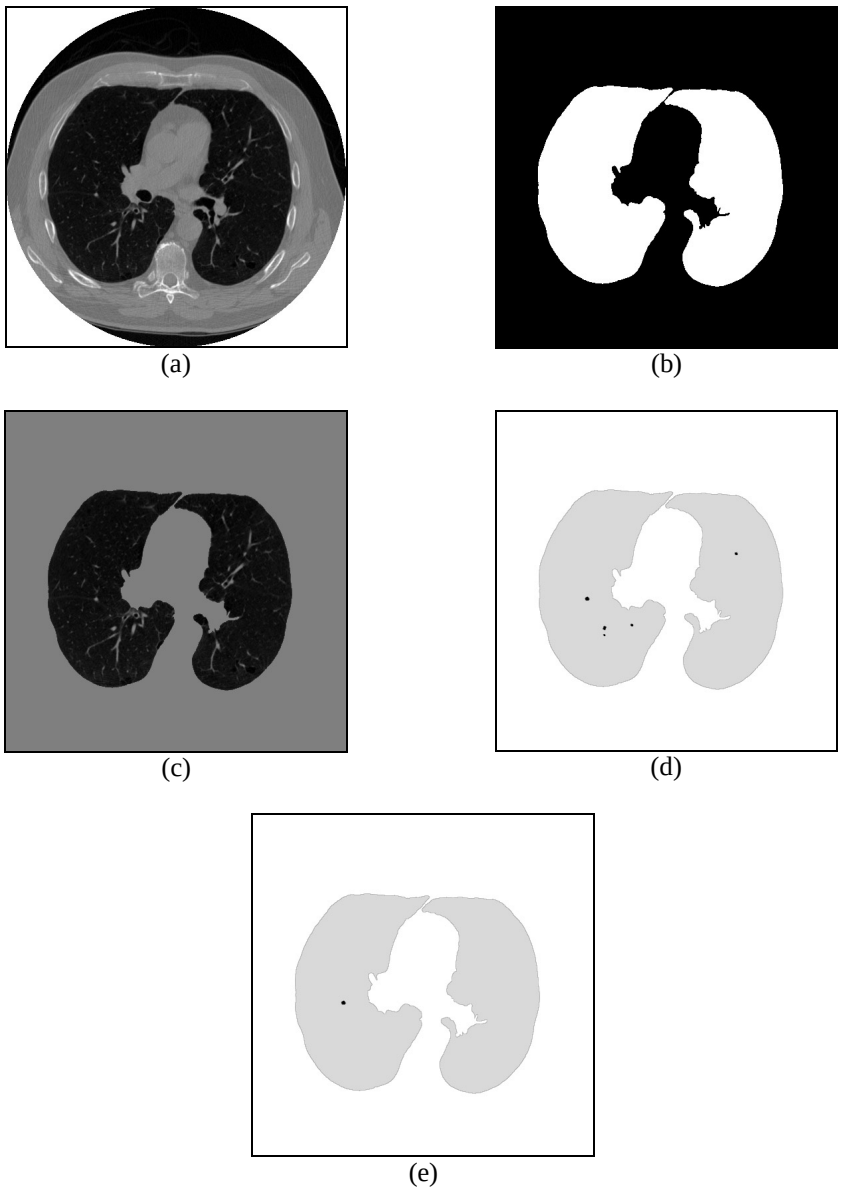

Figure 3. (a) The Original CT Image, (b) Rule-Based Segmented Lung Region, (c) Lung Region of The Original CT Image, (d) ROIs in The Lung Region, (e) Detected Lung Nodule

5. REFERENCES

Armato, S. G., Giger, M. L., Moran, C. J., Blackburn, J. T., Doi, K., and MacMahon, H., (1999), "Computerized Detection of Pulmonary Nodules on CT Scans", Radiographics, 19, 1303-1311.

Brown, M. S., McNitt-Gray, M. F., Goldin, J. G., Suh, R. D., Sayre, J. W., and Aberle, D. R., (2001), "Patient-Specification Models for Lung Nodule Detection and Surveillance in CT Images", IEEE Transactions on Medical Imaging, 20, 1242-1250.

Giger, M. L., Bae, K. T., and MacMahon, H., (1994), "Computerized Detection of Pulmonary Nodules in Computed Tomography Images", Investigate. Radiol., 29, 459-465.

Greenlee, R. T., Nurray, T., Bolden, S., and Wingo, P. A., (2000), "Cancer Statistics 2000", CA Cancer J. Clin., 50, 7-33.

Hounsfield, G. N., (1980), "Computed Medical Imaging", Med. Phys., 7, 283-290.

Kanazawa, K., Kawata, Y., Niki, N., Satoh, H., Ohmatsu, H., Kakinuma, R., et al., (1998), Computer-Aided Diagnostic System for Pulmonary Nodules Based on Helical CT Images, In: K. Doi, H. MacMahon, ML. Giger, K. Hoffmann, eds., Computer-Aided Diagnosis in Medical Imaging, Amsterdam, The Netherlands: Elsevier Science, 131-136.

Lo, S. C. B., Lou, S. L. A., Lin, J. S., Freedman, M., Chien, M. V., and Mun, S. K., (1995), "Artificial Convolutional Neural Network Techniques and Applications for Lung Nodule Detection", IEEE Trans. Med. Imag., 14, 711-718.

Ozekes, S., Osman, O., and Camurcu, A. Y., (2005), "Mammographic Mass Detection Using a Mass Template", Korean J. Radiol, 6, 3, 221-228.

Ozekes, S., and Camurcu, A. Y., (2006), "Automatic Lung Nodule Detection Using Template Matching", Lecture Notes in Computer Science, 4243, 247-253.

Penedo, M. G., Carreira, M. J., Mosquera, A., and Cabello, D., (1998), "Computer-Aided Diagnosis: A Neural-Network-Based Approach to Lung Nodule Detection", IEEE Transactions on Medical Imaging, 17, 872-880.

Samuel, G., Armato, III. Geoffrey M., Michael, F., Charles, R., David, Y., et al., (2004), "Lung Image Database Consortium-Developing a Resource for The Medical Imaging Research Community", Radiology, 739-748.

Xu, X. W., Doi, K., Kobayashi, T., MacMahon, H., and Giger, M., (1997), "Development of an Improved CAD Scheme for Automated Detection of Lung Nodules in Digital Chest Images," Med. Phys., 24, 1395-1403.