Computer-aided Diagnosis: A Support-Vector-Machine-Based Approach of Automatic Pulmonary Nodule Detection in Chest Radiographs

Qionghua Weng, Yan Sun, Xiaopeng Peng, Shuqin Wang, Lixu Gu*, Lijun Qian and Jianrong Xu

***Abstract*—This paper describes an automatic computer aided diagnosis system to detect pulmonary nodules in digital chest radiographs based on support vector machines. First, we design a hybrid segmentation algorithm to delineate the full lung field, specifying the visible area and the retro-cardiac lung area. The midline is also simulated for further symmetrical information evaluation. Then a multi-scale difference approach was implemented to extract initial nodule candidates. To eliminate false positives, different kinds of features are extracted to simulate a nodule's appearance, including our special designed edge radial features. These features were then selected by genetic algorithm for the optimal subset. We performed various learning experiments by changing the kernel and other parameters in SVM. Operating characteristics and accuracy statistics are applied to evaluate and compare the results. Statistics shows that in our system, it is essential to employ cost-sensitive SVM instead of standard SVM due to the huge portion difference between positive and negative examples, and the classifier yield to better performance when adding our edge radial features. Finally, with the best SVM model, we obtain 2.6 fp/image when sensitivity is 0.72; 5.21 fp/image when sensitivity is 0.86. This result is better than ANN and rule-based classifier using the same feature set.**

I. Introduction

LUNG cancer is one of the most common and lethal kinds of cancer. The five-year survival ratio of lung cancer patients can be highly improved from 14% up to 49% [1] if the lesion is detected in the initial phases, when the tumor, presented in the term of lung nodule, is solitary and localized. To detect and diagnose early nodules, X-ray is the most general and essential modality, due to its economical price and wide use in routinely examination. Therefore, the principle aim of this CAD system is to characterize suspicious objects efficiently and automatically, and that can help radiologists make reliable judgments.[2] In recent studies, there are a number of proposals dealing with this problem. Almost all of them implements in two steps [3]: 1. finding suspected nodule areas in a difference image produced by subtraction of nodule-enhanced image and nodule-suppressed image; 2. eliminating false positives from previous candidates according to selected features and results of classifiers. In this paper, we develop a hybrid nodule detection system generally followed these two steps. In the first phase, we applied a multi-scale filter instead of traditional fix-sized filter to adapt the nature of variable diameters in nodules to extract initial candidates. In the second phase, we select a set of features and use support vector machine as a classifier to reduce the number of false positives and display final detection results. We also evaluate the performance of our system compared with several other classifiers.

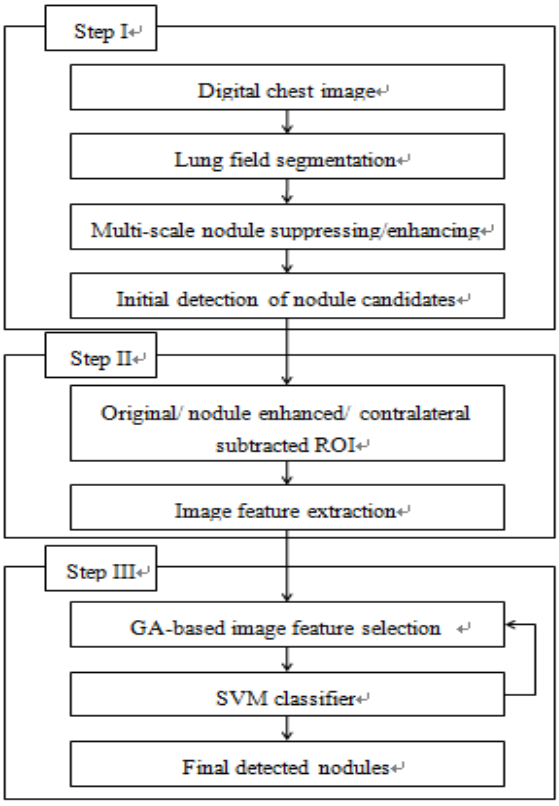

Fig.1 Overall flow chart of the CAD system

Manuscript received August 31, 2009. This work is partially supported by the National Fundamental Research Program (973) under Grant No. 2006CB504801 and 2007CB512701 as well as National Natural Science Foundation of China under Grant No. 30770608.

Qionghua Weng is a student with Digital Art lab, School of Software Engineering, Shanghai Jiaotong University, Shanghai, P.R.China (email: evline45@gmail.com).

Yan Sun, School of Software Engineering, Shanghai Jiaotong University,

Xiaopeng Peng,, Shuqin Wang, School of Software Engineering, Shanghai Jiaotong University, Shanghai, China

Corresponding author: Lixu Gu, Med-X Research Institute, Shanghai Jiaotong University, Shanghai, P.R.China (e-mail: gu-lx@cs.sjtu.edu.cn)

Lijun Qian, Jianrong Xu, Department of Radiology Renji Hospital Shanghai Jiaotong University, China

II. PREPROCESSING

A. Materials

The 53 digital chest radiographs used in this study, including 47 images with nodules and 6 normal ones, were obtained from routine cases at Shanghai Renji Hospital. All 52 pulmonary nodules have been confirmed by CT examination and/or radiologists. The original images have a 0.142 mm pixel size, 15 bit gray levels and not fixed matrix size. 247 chest x-ray images from standard public database by Japanese Society of Radiological Technology (JSRT) Database were also used as test materials.[4] They include 154 tumors whose difficulty of detection distributes almost equally from rank 1 (hard to detect) to rank 5 (easy to detect). The spatial resolution of the original x-ray image is 0.175 mm and the size of each image is 2048 x 2048 pixels with 12-bits accuracy. We converted the total of 300 images to 1024*1024 matrix size and 10 bit gray levels by sub-sampling method to reduce computational costs without worsening the performance.

B. Segmentation of lung area

In this step, we segment the whole lung field, specifying the visible lung, retro-cardiac lung and midline of the lung. First, we use a self-adaptive thresholding technique to categorize the image into three parts: visible lung, air and other human tissue. Then the left and right lung is processed independently for edge detection. By combining the results of thresholding with that of canny edge operator, the edge tracking procedure starts at the uppermost point in the binary image, heading two directions to lineate separately the final outside border and inside border of visible lung. Subsequently, the average locations of the midpoint of left and right ribcage edge can be fitted to a straight line by least square solution, and thus determine the midline of lung area. The bottom boundary is also lineated beginning from the bottom left (right) point. The costophrenic angle is calculated by the first 6 points in the bottom border. The retro-cardiac lung is approximately represented by a vertical edge paralleled to the midline and a horizontal edge connecting the bottom left point and bottom right point of the lung. This algorithm can successfully avoid the conventional problems [5] of missing the lung part above the clavicle and other problems coursed by strange screening position, patients with pacemaker and etc.

C. Nodule candidate selection

A differential approach which subtracts nodule suppressed images from nodule enhanced images is implemented to enhance potential nodule information in a chest radiograph. Since in clinic, the size of a lung nodule could vary from 3mm to 30mm in diameter, we develop a multi-scale filter to extract possible candidates. In the down-sampled 256*256 image, the resolution is approximately one millimeter per pixel. Therefore, the nodule enhanced images are obtained by consecutively convolving the original one with a sphere-shaped filter whose radius takes values from 2 to 15, while the nodule suppressed images are produced by Gaussian smooth filters whose standard deviation s takes the same value range. Then, we combine the results of 14 banarized differential images which are generated by subtraction to get all suspected nodules. Unqualified regions (e.g. too long, too small) are eliminated from the candidate sets. After this multi-scale scheme, the total number of candidates in our database is 20312. That is about 67.71 false positives per image. This scheme neglected 28 true positives, 60% are very subtle.

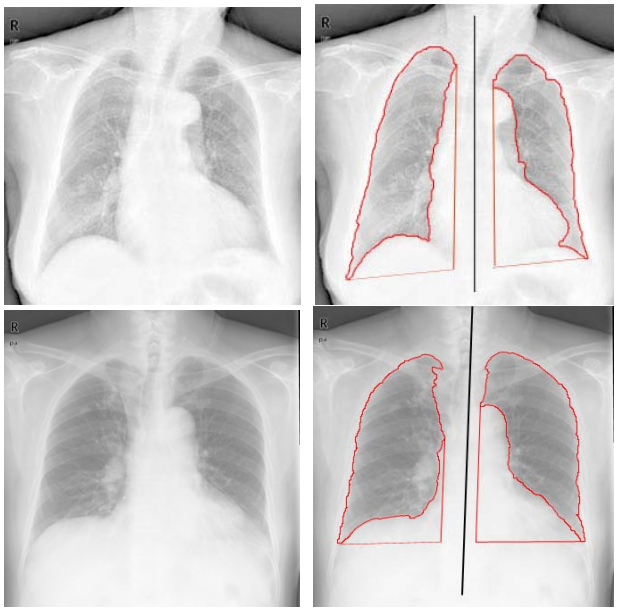

Fig.2 Two examples of lung filed segmentation

Fig.3 Extracted initial candidates

III. FEATURE SELECTION

In chest radiographs, a typical pulmonary nodule would appear as a round shaped, bright lesion and surrounded by a darker ring. Therefore, in this step we extract 80 features from three kinds of filtered image, as is shown in Fig.4: original image, nodule enhanced image and contralateral subtraction image [6]. For each candidate nodule(CN) acquired from II.C, we extract a corresponding square-shaped region of interest (ROI) as the original image which shares the same center with the CN and has a side-length that is twice the size of the CN's diameter. The nodule enhanced image is obtained from

histogram equalized followed by morphological top-hat filtering. The contralateral image is obtained by subtracting the symmetric region from the original image according to the axis. As the midline is not always strictly vertical, the corresponding symmetric region usually needs to be rotated for a slight angle. All these features could be categorized mainly into four types:

A. Geometric features

The geometric features are calculated from the binary image of ROI. A sequence of thresholds which separate the area under the histogram ranging from 5% to 50% at high pixel levels are automatically tried to determine the final value of thresholding, when the difference of two consecutive thresholds is under a certain value and the biggest circles obtained by Hough transform are steady. The features are effective diameter (the diameter of a circle with the same area as that of the nodule region), circularity (overlapping area between the circle and the nodule) [7], normalized perimeter (perimeter of the nodule/perimeter of the circle), area fraction (area of maximum bounding box/ area of minimum bounding box), diameter fraction (effective diameter/diameter of the circle obtained by Hough transform).

B. Gray-level features

In order to better simulate the appearance of pulmonary nodule, we divide the ROI into three parts: the lesion area (N) that covers the core region of the nodule, the ring area(R) that corresponds to the boundary of the nodule, and the background area (B). The width of the ring area is determined by rays casting from the centroid of CN in 24 directions. (The interval is 15 degree).

1) First-order statistics

First-order statistical features can describe the distribution of image gray-level. They are calculated from histograms of the three regions. Each histogram includes mean, standard deviation, contrast, skewness, kurtosis, energy and entropy.

2) Contrast features

The gray-level difference between N and B, N and R also provides useful information. Therefore, we define 3 features to measure the contrast status between two regions A1 and A2. Intensity difference is the contrast between their mean pixel values: $ID = I(A_1) - I(A_2)$ (1)

Square difference measures the distance between their histograms, the square of the difference between both means was divided by the sum of the variances of the two areas:

$$SD = \frac{(I(A_1) - I(A_2))^2}{Var(A_1) + Var(A_2)} \quad (2)$$

For each entry in the normalized histogram of the two areas, the absolute value of the difference was computed:

$$O = \sum_i \left| H(A_1, i) - H(A_2, i) \right| \quad (3)$$

This yields a value between 0 (total overlap) and 2 (complete separation). The main advantage of this feature is that it is independent of scaling of the intensity values of the image.

C. Edge-radial features

For each ray casted from the centroid of the lesion area, two radius can be found by gradient measures, as shown in fig. 5. The mean, max, contrast value and standard deviation of the radius series in all directions are calculated as edge radial features. The difference between R1 and R2 which represents the width information of ring area is also counted into the feature sets.

Fig.4 original nodule, enhanced image, subtracted image.

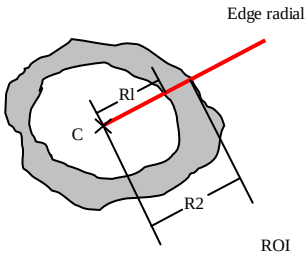

Fig.5 Sketch map of the edge radial in ROI.

D. Position features

Position features are the x,y coordinates of the centroid of the lesion region in the normalized lung field, and the Euclidean distance between the centroid and the midline.

IV. CLASSIFICATION

A. Support Vector Machine

In recent years, the support vector machine (SVM) introduced by V. Vapnik has been widely used in various classification tasks due to its good performance.[] In the literature, several authors have reported good results of SVM for candidate classification in medical image processing. Based on the structural risk minimization principle, a non-linear SVM classifier seek to construct a hyper-plane in the feature space as the decision surface so that the margin of separation between two classes is maximized and the error rate in the sample space is bounded to a certain upper limit. That guarantees theoretically the largest generalization ability of a SVM model. The SVM term comes from the fact that the points in the training set which are closest to the decision surface are called support vectors. When designing SVM classifier, the kernel used for feature mapping has to be chosen carefully since an inappropriate kernel can lead to poor performance. In our experiment, we use Gaussian radial basis kernel function. The kernel products between input vectors x and y is $K_{\text{Gaussian}}(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2})$. Since we have very

imbalanced training data sets, the positive examples are much less than negative ones, a cost-sensitive approach is necessary when using SVM. Two regularization coefficients are introduced here to minimize the number of high cost errors and total misclassification cost.[8]

B. Feature selection

In order to find the optimal combination of features and reduce computational time, a feature selection work is quite essential before training. Genetic algorithm is widely recognized as an effective approach to detect most informative variables. This heuristic algorithm starts by generating random feature sets and then refining them by evolutionary computation, such as inheritance, mutation, selection and crossover. In our system, first we use binary encoding. That is, every chromosome is a string of bits, 0 or 1, representing the presence or absence of one certain feature. Thus, the length of one chromosome is the number of features. Our fitness function is defined as

$$\text{fitness} = W_A \times Accuracy + W_F \times (\sum_{i=1}^{n_f} C_i \times F_i)^{-1}$$

in which W_A is the weight of classification accuracy, W_F is the weight of feature quantity, Ci is cost of feature I, Fi='1' represents that feature i is selected; '0'represents that feature i is not selected. [9]

V. Experiments and Results

As described in passage II.A, we initially collect 210 nodule cases from 300 chest images. These 210 nodules are randomly divided into 2 groups: 160 cases for training and 50 cases for testing. After the initial candidate extraction scheme, we have 160 positive examples and 14921 negative examples. Then we extracted 56 features mentioned above for each candidate region. The optimal subset obtained by GA consists of 16 features, including 6 geometric features, 5 gray-level features, 4 edge radial features and 1 position feature.

In the classification phase, the best coefficients are selected by 4-fold cross-validation and grid search technique. The polynomial degree d is searched from $d \in \{2.0,5.0\}$ in steps of 0.5. The σ in Gaussian kernel is in the range of {50,1000} in steps of 50. The grid search result shows that polynomial kernel performs better when $d \leq 3$, and Gaussian kernel yield better outputs when $\sigma \geq 100$. Some of the good results are shown in Table I. We can see that with the same sensitivity, the best results of two kernels are nearly the same. Finally, we reached 2.6 fp/image when sensitivity is 0.72;5.21 fp/image when sensitivity is 0.86. The best Az[10] value is 0.8542.

For comparison, we also tested the performance of standard SVM without cost parameters, SVM without feature selection phase and 12 feature sets without edge radial features, as is shown in Table II. The results show that cost-sensitive approach, feature selection technique and edge radial features have effectively improved the performance of our system. Different classifier outcome with the same 16 feature set are also evaluated. The ANN employed here has three layers with back-propagation algorithm. This comparison demonstrates a good generalization ability of our SVM classifier.

TABLE I.
RESULTS WITH COST-SENSITIVE SVM

σ	C+	C-	Sens.	Spec.	Fp/image
100	0.1	20	0.7143	0.9677	2.62
50	0.1	20	0.7857	0.9603	2.93
50	0.05	50	0.8571	0.9412	5.21
100	0.01	50	0.9286	0.9013	7.12
700	0.5	100	1	0.8892	8.34

TABLE II
RESULTS WITH DIFFERENT TYPES OF CLASSIFIERS

	Standard SVM	SVM with 56 features	SVM with 12 features	ANN	Rule based
Az	0.5613	0.7712	0.7856	0.7852	0.7612

VI. Conclusion

An automatic computerized scheme of pulmonary nodule detection is presented in this paper. In this scheme, 16 features are extracted and selected from suspected nodule regions based on GA and SVM, the final nodules are detected with an Az value of 0.8542.

Acknowledgment

The authors would like to thank experts from Shanghai Renji Hospital for their clinical advices.

References

[1] G. P. Murphy, W. Lawrence Jr., and R. E. Lenhard Jr., *Clinical Oncology*.Washington, DC: Amer. Cancer Soc., 1995

[2] B. van Ginneken, B. M. t. H. Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," *IEEE Trans.Med. Imag.*, vol. 20, no. 12, pp. 1228–1241, Dec. 2001

[3] Computer-aided Diagnosis in Chest Radiography: Results of Large-Scale Observer Tests at the 1996–2001 RSNA Scientific Assemblies *Radio Graphics* 23: 255-265, 2003.

[4] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu,M. Matsui, H. Fujita, Y. Kodera, and K. Doi. 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. *American Journal of Roentgenology*, 174: 71–74, 2000.

[5] X. Xu and K. Doi, "Image feature analysis for computer-aided diagnosis: Accurate determination of ribcage boundary in chest radiographs," *Med Phys.*, vol. 22, no. 5, pp. 617–626, 1995.

[6] M. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: Automated detection of nodules in peripheral lung fields," *Med Phys.*, vol. 15, no. 2, pp.158–166, 1988.

[7] Q. Li, S. Katsuragawa, T. Ishida, H.Yoshida, S. Tsukuda, H. MacMahon, and K. Doi, "Contralateral subtraction: A novel technique for detection of asymmetric abnormalities on digital chest radiographs," *Med Phys.*,vol. 27, no. 1, pp. 47–55, 2000.

[8] C.-C. Change and C.-J.Lin, LIBSVM—A library for support vector machines http://www.csiew.ntu.edu.tw/~cjlin/libsvm/

[9] C.-L. Huang and C.-J. Wang. A GA-based feature selection and parameter optimization for support vector machines. *Expert Systems with Applications*, 31:231 - 240, 2006