CARS 2002 – H.U. Lemke, M.W. Vannier; K. Inamura, A.G. Farman, K. Doi & J.H.C. Reiber (Editors)

Optimal image feature set for detecting lung nodules on chest X-ray images

Jun Wei, Yoshihiro Hagihara, Akinobu Shimizu, Hidefumi Kobatake
Graduate School of Bio-Applications and Systems Engineering
Tokyo University of Agriculture and Technology
2-24-16, Naka-cho, Koganei, 184-8588, Tokyo, Japan

Abstract

The performance of a computer-aided diagnosis system depends on the feature set used in it. This paper shows the results of image feature selection experiments. We evaluated 210 features to look for the optimum feature set. For the purpose, a forward stepwise selection approach was employed. The area under the receiver operating characteristic (ROC) curve was adopted to evaluate the performance of each feature set. Analysis of the optimally selected feature set is given and the experiments using 247 chest x-ray images are also shown.

Keywords: Computer aided diagnosis, lung cancer, feature selection

1. Introduction

Lung cancer is one of the most serious cancers in the world. Survival from lung cancer is directly related to its growth at its detection. The earlier the detection is, the higher the chances of successful treatment are. Chest X-ray image has been used for detecting lung cancer for a long time. The early detection and diagnosis of pulmonary nodules in chest X-ray image are among the most challenging clinical tasks performed by radiologists. Computer-aided diagnosis (CAD) has been proven to be a very effective approach as assistant to radiologists for improving diagnostic accuracy. Numerous systems were reported for detecting lung nodules on chest X-ray images [1-3]. However, the strong concern of almost all of them is that the false positives per image are too large. How to reduce the number of false positives while maintaining a high true positive detection rate is the most important work in realizing a chest CAD system[4].

Most of the proposed computer-aided diagnosis systems (CAD systems) adopt a two-step pattern recognition approach, which is a combination of a feature extraction process and a classification process using neural network classifier or statistical classifier. The performance of the classifier depends directly on the ability of characterization of candidate regions by the adopted features. Many kinds of features have been proposed for discriminating between normal tissues and abnormal ones. However, there have been a few researches on comparing the effectiveness of those features[5-9]. The numbers of features used in those researches are not sufficiently large. The purpose of our research is to find the optimal feature set from a large number of features which enable a CAD

system for lung cancer screening to take large step toward a practical application. And this paper shows the results of the preliminary experiments of this project.

2. CAD system and the optimal feature set

Our CAD system consists of four processing steps: 1) location of tumor candidates by using adaptive ring filter, 2) extraction of the boundaries of tumor candidates, 3) extraction of feature parameters, and 4) discrimination between the normal and the abnormal regions. Fig. 1 shows the configuration of the CAD system. Adaptive ring filter, which is a kind of convergence index filter (CI filter) [10], is employed to extract tumor candidates. It evaluates the degree of convergence of gradient vectors to the pixel of interest. Its output does not depend on the contrast of the region of interest to its background. Actually, we have found highly ranked local peaks of the outputs of the adaptive ring filter correspond to the summit of tumors. In this work, the top 25 peaks on each X-ray image are detected as the tumor candidate location.

At each tumor candidate location, the boundary of the candidate is estimated by using a two-step process. In the first step, Iris filter, which is another kind of CI filter, is used to estimate the fuzzy boundary [11]. Then, SNAKES algorithm is applied to the output image of the Iris filter to obtain the boundary of the tumor candidate. It is called a suspicious region (SR) in the following. Feature parameters are calculated for each SR.
The discrimination between the normal and the abnormal regions is performed using a statistical method based on the Maharanobis distance measure. Fig. 2 shows the whole story of the project of our research. Features are extracted from each of multi-resolution images. Various kinds of filtering or transformation such as Fourier transform, Wavelet transform, spatial difference, Iris filtering, adaptive ring filtering, et al. can be applied to each image. Those transformed images give another various kinds of features. The total number of features extracted from the multi-scale images and their transformed ones can be well over one thousand. Among them we can expect to derive the optimal feature set by which the performance of the CAD system can be vastly improved. This project is the work-in-progress and this paper shows the results of the preliminary study using the original images reduced by 1/4. Spatial resolution of the original image is 0.175mm per pixel and that of the reduced image is 0.7mm per pixel.

3. Feature selection

3.1 Method

Feature selection is a very important step in organizing a classifier. Theoretical approach cannot be applied to determine the optimal combination of features, and the only way to select the optimal feature subset is to evaluate all possible combinations of the features. Moreover, sufficient numbers of test materials are necessary to evaluate the performance of

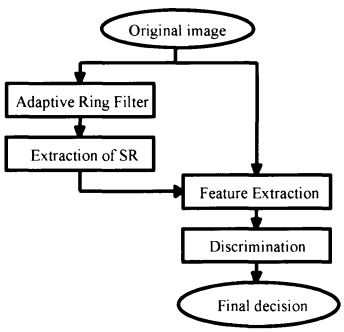

Fig. 1 Processing flow of the CAD system

each feature combination. It means that the number of combinations and the total amount of computation time become impractically huge. Therefore, it is acceptable to adopt heuristic algorithms such as a genetic algorithm, a forward stepwise and a backward stepwise selections which need much smaller computational loads. These algorithms give not a really optimal feature set but a sub-optimal one because only a part of possible combinations of features are evaluated. The sub-optimal feature set is referred to as the optimal feature set for simplicity in the following. In this work, we adopted the forward stepwise selection method to obtain the optimal feature set. The area under ROC curve was adopted as the criterion to evaluate the performance of feature sets.

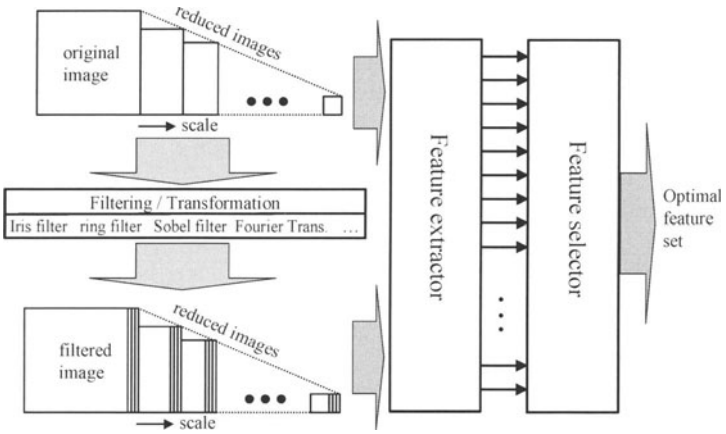

Fig. 2 The scheme to obtain the optimal feature set

3.2 Features

Four kinds of filtered images, that is, the original image reduced by 1/4, the output images of the iris filter and the adaptive ring filter, and the difference image obtained by applying the Sobel filter were used to extract features, and the number of features examined in the following experiments is 210. These features can be categorized into four types as follows.

(1) Geometric Features

Six geometric features are calculated from the binary SR region. They are Spreadness, Circularity, Area, Equivalent radius, Distance from the candidate point to the pulmonary hilum, and Flatness.

(2) Contrast Features

Generally, tumor region is brighter than its background on x-ray image. So, the contrast information can be used as features. In this work, nine kinds of features [12] are calculated from SR regions on each of 4 kinds of filtered images. The total number of features related to the contrast is 36.

(3) First-Order Statistics

First-order statistics are calculated from two histograms of the grey-scale values. They are obtained from two regions. One is called the inner region which covers the core region of SR. The other one is called the outer region which corresponds to the boundary area of SR. The histograms are obtained from 4 kinds of filtered images. Features calculated from

each histogram include mean, standard deviation, contrast, skewness, kurtosis, energy and entropy. The total number of first-order statistical features is 56.

(4) Second-Order Statistics

Co-occurrence matrix method has been adopted to extract features of second-order statistics. They are obtained by using Haralick transformation. Co-occurrence matrix is the two-dimensional histograms of the frequency of the joint occurrences of two pixels with a displacement and an orientation. In this work, they were set to 2 pixels and 90°, respectively. Co-occurrence matrices are obtained from the inner and the outer regions of each SR. The fourteen scalar statistical properties were calculated from a co-occurrence matrix. Four kinds of filtered images were used to calculate it. Therefore, the total number of features related to the second-order statistics is 112.

4. Experiments

4.1 Experimental materials

Two hundred and forty-seven chest x-ray images included in JSRT Database were used as test materials. They include 154 malignant 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. Using our CAD system, 6175 candidate regions (SR) were detected. Several malignant tumors split into two or three candidate regions and 187 SR's correspond to malignant tumors. The other 5988 SR's correspond to normal tissues. The forward stepwise selection method was adopted to find the optimal combination of features among 210 features. The leave-one-out method was employed to evaluate the performance of each combination of features. In reference [13], CAD system based on the optimal feature set obtained from 55 features is described. These features are included in 210 features. For comparing the performance changes by increasing the number of features, experiments using 55 features have been also performed.

4.2 Experimental results

Fig. 3 shows the relationship between the area under the ROC curve (Az) and the number of features (the dimension of the feature vector) selected by the forward stepwise selection method. Among 210 features there were 8 features which are highly correlated with other ones and they were excluded from the experiments. Therefore the total number of features is 202. The area Az increases gradually as the increase of the number of features and it keeps almost the maximum (Az = about 0.85) where the number of features is over 30. The maximum of Az was attained by the combination of 98 features. Table 1 shows the contribution of 4 kinds of filtered images to the optimal feature set with 98 features. We can say that the 4 kinds of filtered images almost equally contain information effective in discriminating between malignant and normal tissues. The relationship between Az and the number of features for the case of 55 features is also shown in Fig. 3. Experimental results showed that the performance attained by the optimal subset of 202 features is much better than that of 55 features.

The relationship between the number of false positives per image and the number of features was also analyzed, where the true positive detection rate is 80%. We can say that by using optimally selected feature set the number of false positives per image can be as

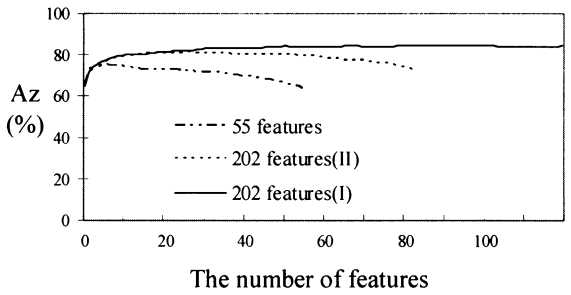

Fig. 3 The relationship between the number of features and the area Az.

low as 5.4 per image.

Another optimal feature selection procedure has been applied to features belonging to the types (2), (3) and (4). First, the optimal feature set for each of 4 kinds of filtered images was determined using the forward stepwise selection method. The numbers of the features consisting of the optimal feature sets are 25, 24, 13 and 16 for the original image, the output image of the iris filter, the output image of the adaptive ring filter and the difference image, respectively. Totally 78 features were selected. By adding six geometrical features of the type (1), the total number of features becomes 84. Then, the forward stepwise selection method was applied to obtain the optimal feature set among 84 features. Experimental results as shown in Fig. 3 showed that the area Az attained by the feature set obtained by this procedure is slightly smaller than that of the optimal feature set among 202 features.

5. Conclusion

The optimal feature set among 210 features has been identified using the forward stepwise selection method. The average number of false positives per image attained by the optimal feature set is as low as 5.4 per image where the true positive detection rate is 80%. The spatial resolution of the X-ray images used in this paper is 0.7mm per pixel. It is low enough to remove fine structures of candidate regions. However the experimental results shown in this paper is promising. Experiments using images with higher spatial resolution and much more features are now in progress.

Table 1 The contribution of 4 kinds of filtered images to the optimal feature set. Arabic numerals show the number of selected features.

Filtered image	Geometric Features	Contrast Features	First-Order Statistics	Second-Order Statistics	Total
-	3	-	-	-	3
Reduced original image	-	6	4	8	18
Output image of the Iris filter	-	7	5	15	27
Difference image	-	8	6	8	22
Output image of the ring filter	-	6	6	16	28
Total	3	27	21	47	98

Acknowledgements

This work was supported in part by the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan and the Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare, Japan.

References

1. H. Suzuki, N. Inaoka, H. Takabatake, M. Mori, H. Natori and A. Suzuki, "An experiment system for detecting lung nodules by chest x-ray image processing," SPIE. Biomedical Image Processing II, Vol.. 1450, pp. 99-107, 1991.
2. J. Lin, S. 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 Trans. On Med. Imag., Vol.. 15, pp. 206-217, 1996.
3. Xin-Wei. Xu, K. Doi, T. Kobayashi, H. MacMahon and M. L. Giger, "Development of an improved CAD scheme for automated detection of lung nodules in digital chest images," Med. Phys., Vol..24, No. 9, pp. 1395-1403, 1997.
4. H. Yoshida and K. Doi, "Computerized detection of pulmonary nodules in chest radiographs: reduction of false positives based on symmetry between left and right lungs," Proc. SPIE in Medical Imaging 2000, pp. 97-102, 2000.
5. Y.Wu, M.L.Giger, K.Doi, C.J.Vyboorny, R.A.Schmidt, C.E.Metz, "Artificial Neural Networks in Mammography: Application to Decision Making in the Diagnosis of Breast Cancer," Radiology, Vol..187, pp.81-87, 1993.
6. M.Kupinski , M.L.Giger , P.Lu, and Z.Huo, "Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction, " Proc. of SPIE, Vol..2434, pp.598-605, 1995.
7. B.Sahiner, H P Chan, D.Wei, N.Petrick, M.A.Helvie, D.D.Adler, and M.M.Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue, " Med. Phys. Vol..23, No.10, pp.1671-1683, 1996.
8. M.A.Kupinski, M.L.Giger, "Feature selection and Classifiers for the Computerized Detection of Mass Lesions in Digital Mammography," IEEE International Congress on Neural Networks, Houston, Texas, June, pp. 2460-2463, 1997.
9. G.D.Tourassi, E.D.Frederick, M.K. Markey, C.E.Floyd, Jr., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," Med. Phys., Vol..28, No.12, pp.2394-2402, 2001.
10. Jun Wei, Yoshihiro Hagihara and Hidefumi Kobatake: Detection of Cancerous Tumors on Chest X-ray Images - Candidate Detection Filter and Its Application -, Proc. of ICIP, Paper No. 27AP4.2, 1999.
11. H. Kobatake and S. Hashimoto, "Convergence Index Filter for Vector Fields," IEEE Trans. on Image Processing, Vol. 8, No. 8, pp.1029-1038, 1999.
12. Guido M te Brake, Nico Karssemeijer, et al, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," Phys. Med. Biol., 45, pp. 2843-2857, 2000.
13. J. Wei, Y. Hagihara, H. Kobatake, "Detection of lung nodules on digital chest radiographs," Med. Imag. Tech., Vol. 19, No. 6, pp. 468-476, 2001. (in Japanese)