Extraction Technique of Spicules-Based Features for the Classification of Pulmonary Nodules on Computed Tomography

Xingyi He, Jing Gong, Lijia Wang, and Shengdong Nie(✉)

Institute of Medical Imaging Engineering, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China
nsd4647@163.com

Abstract. To avoid the deformation of spicules surrounding pulmonary nodules caused by the classic rubber band straightening transform (RBST), we propose a novel RBST technique to extract spicules-based features. In this paper, the run-length statistics (RLS) features are extracted from the RBST image, in which a smooth circumference with a suitable radius inside the nodule is proposed as the border of transformed object. An experimental sample set of 814 images of pulmonary nodules was used to verify the proposed feature extraction technique. The best accuracy, sensitivity and specificity achieved based on the proposed features were 79.4%, 66.5%, 89.2%, respectively, and the area under the receiver operating characteristic curve was 87.0%. These results indicate that the proposed method of feature extraction is promising for classifying benign and malignant pulmonary nodules.

Keywords: Pulmonary nodules · Feature extraction · The rubber band straightening transform · Spicules-based features

1 Introduction

Since spiculation is one of the most important medical signs for the clinical diagnosis of benign and malignant pulmonary nodules (potential manifestation of lung cancer at early stage [1]), extraction technique of spicules-based features based on computed tomography (CT) scans has been widely studied [2–5]. The RBST technique introduced by Sahiner et al. [6] is an effective way to analyze the spiculation of nodules. According to the RBST technique, spicules that grow out radially from nodules can be transformed along approximately straight lines in the vertical direction and be tiled perfectly [6–8]. Therefore, it has to result in the deformation of the spicules, because of the rough borders of the transformed objects.

In order to suggest a solution for the deformable problem of spicules caused by the rough object border, we propose a novel RBST algorithm able to extract spicules-based features, in which the border of the transformed object is set to a smooth circumference with a suitable radius inside nodules.

M. Fei et al. (Eds.): LSMS/ICSEE 2017, Part I, CCIS 761, pp. 123–131, 2017.
DOI: 10.1007/978-981-10-6370-1_13

2 Methods

In this section, we introduce the proposed spicules-based feature extraction methodology based on a novel RBST technique. The flow diagram is presented in Fig. 1. There are four details to be improved during this process: the determination of the object border, computation of normals, computation of pixel values, and the determination of well-suited sampling interval. Finally, the RLS analysis is done in order to extract features from the RBST image.

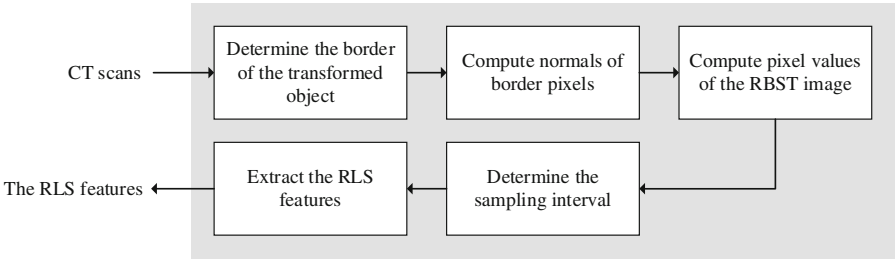

Fig. 1. The flow diagram of extracting features

2.1 Determine the Border of the Transformed Object

It is important to determine a suitable object border, because the inclusion of the RBST image will be determined by the object border. We consider a circumference with a suitable radius inside the nodule as the object border in this paper. There are two advantages: (1) avoiding the deformation of the spicules caused by the rough border; and (2) simplifying the calculation of normal vectors. The location and the radius of the circumference, which play an important role in deciding whether all of spicules can be included into the RBST image, are determined by the following strategies: firstly, the centroid of the circumference denoted by (X, Y) is considered as the center of the nodule to locate the position of the circumference; secondly, we set the circumference's radius denoted by r as the minimum value of the distance denoted by d from the centroid of the nodule to the surface of the nodule, in case the loss of spicules occurs, as illustrated in Fig. 2(a). However, when the length of spicules is too jagged, it is possible to include too much non-spicules region into the RBST image if the minimum d is equal to the r.

For this problem, we set experimentally the r as the half of the average of the d when the length of spicules is too jagged, as illustrated in Fig. 2(b). The jagged degree of spicules is measured by the standard deviation of the d and it is found experimentally that when the standard deviation of the d is equal to 3 mm, not only a number of non-spicules region can be removed from the RBST image, but also spicules can be included as many as possible. Therefore, the value of the r is determined finally according to the Eq. (1). Let (x_j, y_j) denote the position of the border pixel j on the original image, and the x_j and y_j can be calculated by Eq. (2), where the θ_j denotes the deviating degree on the Cartesian plane.

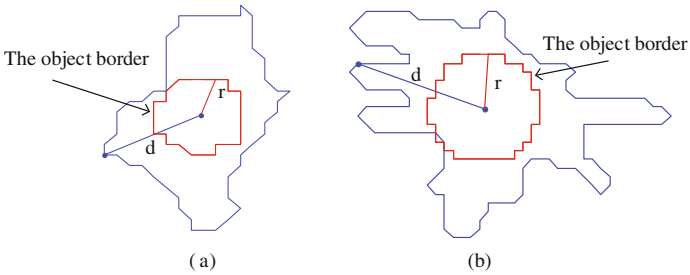

Fig. 2. Examples of the object borders inside the nodules

$$r = \begin{cases} \min(d), & \text{std}(d) \le 3; \\ \frac{1}{2}\text{mean}(d), & \text{std}(d) > 3. \end{cases} \tag{1}$$

$$\begin{bmatrix} x_j \\ y_j \end{bmatrix} = r \cdot \begin{bmatrix} \cos(\theta_j) \\ \sin(\theta_j) \end{bmatrix} + \begin{bmatrix} X \\ Y \end{bmatrix}. \tag{2}$$

2.2 Compute Normal of Border Pixels

Since the object border in this paper is a circumference as mentioned above, the normal direction to the object actually is equal to the direction from the border pixel to the centroid of the circumference. For a given border pixel which places the x_j and y_j coordinates on the original image, the normal direction $\boldsymbol{n}(j)$ through the pixel j can be calculated by Eq. (3).

$$\boldsymbol{n}(j) = \frac{1}{r}\left(\begin{bmatrix} x_j \\ y_j \end{bmatrix} - \begin{bmatrix} X \\ Y \end{bmatrix}\right). \tag{3}$$

2.3 Compute Pixel Values of the RBST Image

Afterwards, the value of each pixel on the RBST image is calculated by linear interpolation algorithm in this paper. For the pixel point (i, j) which places the *ith* row and the *jth* column on the RBST image, its normal line $\boldsymbol{n}(j)$ on the original image is through the border pixel j, and let P(x, y) denote the pixel value which places the x and y coordinates on the original image. The value denoted by $p(i, j)$ at the pixel point (i, j) is calculated according to the closest pixel on the original image, as presented in Eq. (4), where the Δx_j and Δy_j denote the sampling intervals of the normal direction $\boldsymbol{n}(j)$ in the horizontal and vertical direction, respectively.

$$p(i, j) = P\left(\lfloor x_j + \Delta x_j \cdot (i-1) \rfloor, \lfloor y_j + \Delta y_j \cdot (i-1) \rfloor\right). \tag{4}$$

Two examples of the original images and their RBST images about benign and malignant nodule were presented in Fig. 3.

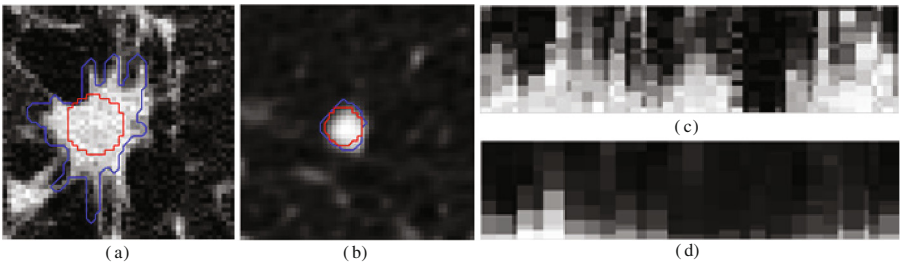

Fig. 3. The RBST images of benign and malignant nodules: in (a) and (c), the red line represents the circumference inside the nodule, and the blue one represents the nodule's boundary. The RBST image of (a) with the spicules was presented in (c), while the RBST image of (b) without the spicules was presented in (d). (Color figure online)

2.4 Determine the Sampling Intervals

It is necessary to decide appropriate sampling intervals Δx_j and Δy_j. In this paper, the Δx_j and Δy_j are determined by the pixel spacing of the RBST image, which is denoted by PX. For the circle object whose normal direction through the border pixel j is $\boldsymbol{n}(j)$, the sampling intervals along the normal direction $\boldsymbol{n}(j)$ (Δx_j and Δy_j) are calculated by Eq. (5). In addition, the number of sampling points on the circumference denoted by N and the θ_j mentioned above are also determined by PX, as presented in Eqs. (6), and (7). The number of the columns of the RBST image and PX are adjusted in this paper for the best values.

$$\begin{bmatrix} \Delta x_j \\ \Delta y_j \end{bmatrix} = PX \cdot \boldsymbol{n}(j). \tag{5}$$

$$N = \frac{2\pi r}{PX}. \tag{6}$$

$$\theta_j = j \cdot \frac{2\pi}{N}. \tag{7}$$

2.5 Extract the RLS Features

Eleven RLS features [9, 10] are extracted in this paper according to the method presented by Way T.W. et al. [8], including gray-level uniformity (GLN), high gray-level run emphasis (HGRE), low gray-level run emphasis (LGRE), long run emphasis (LRE), long run high gray-level emphasis (LRHGE), long run low gray-level emphasis (LRLGE), run length non-uniformity (RLN), run percentage (RP), short run emphasis (SRE), short run high gray-level emphasis (SRHGE), long run low gray-level emphasis

(SRLGE). In the paper, the runs which are sets of consecutive pixels with the same gray-level intensity along the vertical direction represent either the spicules or the background surrounding the nodule. Feature extraction algorithm proposed in this paper is outlined in Algorithm 1.

Algorithm1. Feature Extraction Based on the RBST technique

1) Input the image of candidate nodule in the middle layer.
2) Set parameters *PX* (increases from 0.05mm to 0.5mm) and the number of the columns of the RBST image (increases from 20 to 60).
3) Calculate the centroid of the circumference according to the center of the nodule.
4) Calculate the radius of the circumference according to the Equation (1).
5) Get the number of sampling points in the circumference according to the Equation (6).
6) Calculate the position of each sampling point according to the Equation (2), (7).
7) Calculate the normals of sampling points according to the Equation (3).
8) Calculate the pixel values of interpolating points along assigned normals according to the Equation (4), (5).
9) Extract 11 RLS features from the RBST image.

3 Results and Discussion

In this section, the classification results which were obtained by using Linear Discriminant Analysis (LDA) classifier and Random Forest (RF) classifier are presented to evaluate the proposed method. In order to assess the performance of the proposed method, some compared classifications [7, 8, 11] which were using the same sample were carried out in this paper. Four common evaluation measures were employed: Accuracy (ACC), Sensitivity (SE), Specificity (SP) and the receiver operating characteristic curve (ROC) listed separately for the following results because of its description based on two-dimensional. We collected a total of 814 lung nodules (462 benign cases and 352 malignant cases) from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) public database [12, 13]. It may be not enough to validate the effectiveness of the proposed approach when sample size is insufficient. Therefore, we applied 10-fold cross-validation method to validate the classification performance in order to make full use of these 814 samples. All classifications were based on the MATLAB 2014 platform.

3.1 The Results Under Different Parameters

The classification was performed based on the RLS features described in Sect. 2 and its area under the ROC curve (AUC) under the different parameters is presented in Fig. 4.

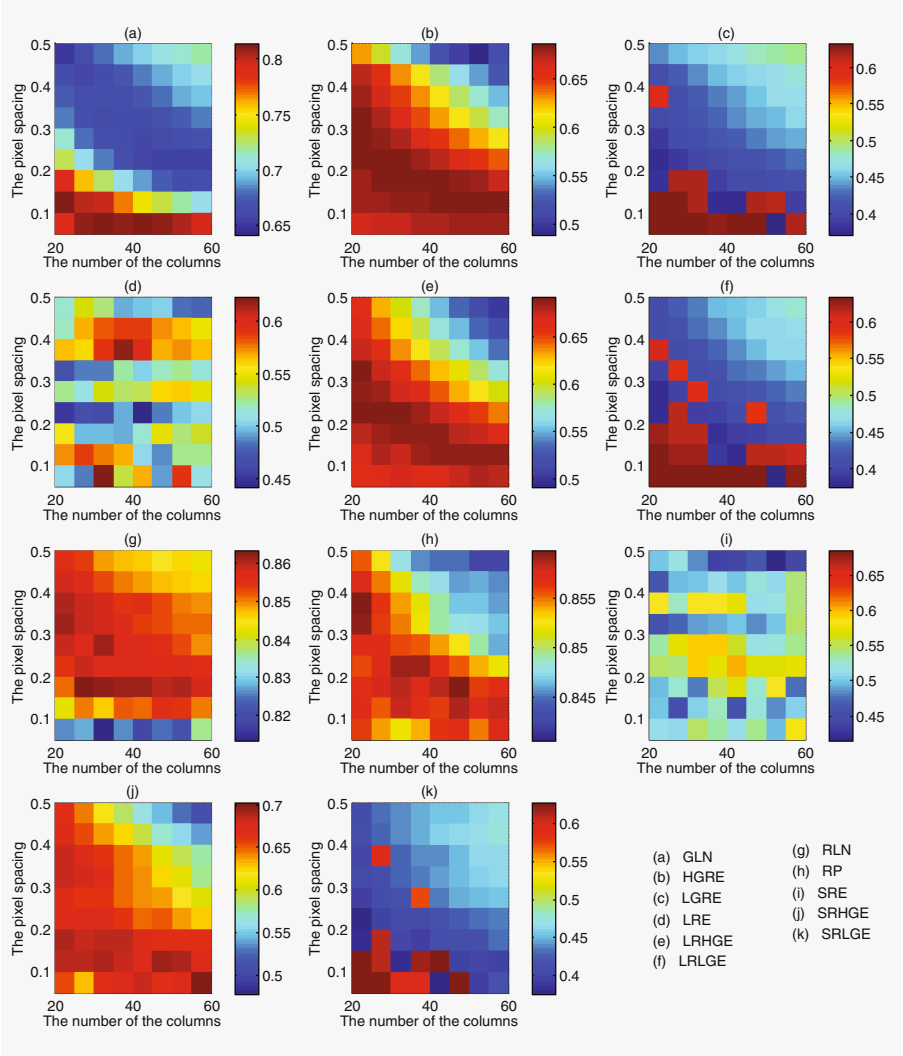

Fig. 4. The AUC values of 11 RLS features under the different parameters

The effect of two parameters on classification results was explored in this study. The one was the pixel spacing of the RBST image and the other one was the number of the columns of the RBST image denoted by *NC*. We adjusted the former ranging from 0.05 mm to 0.5 mm, and adjusted the latter ranging from 20 to 60. From these classification results presented in Fig. 4, we observed that the results are considerately different when setting different parameters. When it comes to features of GLN, HGRE, LGRE, LRHGE, LRLGE, RLN, RP, SRHGE, SRLGE, the value of AUC decreased either when the pixel spacing and the width of the band being too small or when the both being too large. It is because that it is bad for these features that too less spicules

or too much background are included into the RBST image. If the value of the pixel spacing and the number of the columns of the RBST image are too large, the RBST image will not only lose some details of spicules, but also include much useless background. On the contrary, if these both parameters are too small, then the RBST image may be not enough to include too much details. Besides, we observed that the AUC values of LRE and SRE features are irregularly with the pixel spacing and the number of the columns changing. The reason causing this may be their calculations which are according to the length of the runs. Since the short runs and the long runs complement each other in the same column of the RBST image, as a result, the calculation of these two features are affected by the complementary relationship between the short runs and the long runs. Each of the RLS feature with the maximum AUC value whose parameters were recorded in Table 1 was used to classify the nodules in the following experiments.

Table 1. The maximum AUC values and the relative parameters of 11 RLS features

Features	AUC (%)	Parameters		Features	AUC (%)	Parameters	
		PX (mm)	*NC*			*PX* (mm)	*NC*
GLN	81.1	0.1	30	RLN	86.3	0.2	25
HGRE	68.6	0.3	20	RP	86.0	0.4	20
LGRE	63.3	0.1	20	SRE	68.6	0.55	60
LRE	62.3	0.1	30	SRHGE	70.3	0.1	55
LRHGE	69.2	0.25	25	SRLGE	62.7	0.15	20
LRLGE	63.4	0.1	30				

3.2 Comparisons of the Proposed Features with Others

The proposed features were compared with other relevant features [7, 8, 11], and the classification results obtained by using LDA classifier and RF classifier were presented in Table 2. Meanwhile, the performance on the two classifiers is displayed by the ROC curves, as shown in Fig. 5. For the proposed features, the LDA classifier achieved an accuracy of 79.4%, and the RF classifier achieved an accuracy of 80.7%. Besides, by analyzing area under ROC curves, we find that the proposed features are better, no matter what classifiers are used. These results indicated that the proposed features are more promising to distinguish malignant pulmonary nodules from benign ones. It is because that the improved RBST technique in this paper is capable of avoiding the deformation of the spicules occurred during the classic RBST technique. At the same time, the RLS statistics has a more powerful ability to analyze all kinds of streaky structures on the image. Although the proposed spicules-based feature extraction method is effective to classify pulmonary nodules, its limitations still exist. For example, there features are extracted from 2-dimensional slice images, so that they cannot characterize all of spicules growing in 3-dimensional space. Therefore, promoting the proposed method to 3-dimensional space should be well considered and studied in our further study.

Table 2. The classification results obtained by using LDA and RF classifiers

Classifiers	Features	ACC (%)	SE (%)	SP (%)	AUC (%)
LDA	The proposed	79.4	66.5	89.2	87.0
	Reference [7]	77.0	63.3	86.5	86.0
	Reference [8]	71.6	58.0	77.9	82.1
	Reference [11]	73.3	65.1	80.0	80.5
RF	The proposed	80.7	76.1	84.2	86.9
	Reference [7]	76.1	76.4	83.3	85.9
	Reference [8]	75.2	68.8	80.1	80.6
	Reference [11]	72.5	65.9	77.5	79.3

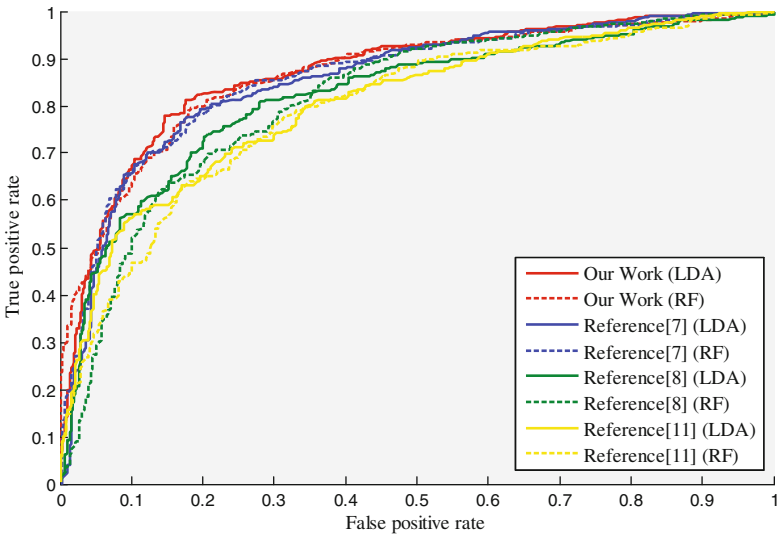

Fig. 5. The ROC curves of LDA and RF classifiers

4 Conclusions

In this study, we have developed spicules-based features based on the improved RBST algorithm. The results obtained were evaluated by the accuracy, sensitivity, specificity and the area under the receiver operating characteristic curve, and their values obtained were respectively 79.4%, 66.5%, 89.2%, and 87.0%. The comparisons with other reports indicate that the proposed features are beneficial to improve the classification performance of benign and malignant pulmonary nodules.

Acknowledgments. This work was partially funded by the National Natural Science Foundation of China under Grant (No. 60972122) and the Natural Science Foundation of Shanghai under Grant (No. 14ZR1427900).

References

1. Chowdhry, A.A., Mohammed, T.L.H.: Assessment of the solitary pulmonary nodule: an overview. In: Ravenel, J. (ed.) Lung Cancer Imaging, pp. 39–48. Springer, New York (2013). doi:10.1007/978-1-60761-620-7_4
2. Chen, H., Zhang, J., Xu, Y., et al.: Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans. Expert Syst. Appl. **39**, 11503–11509 (2012)
3. Cheng, J.Z., Ni, D., Chou, Y.H., et al.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. Sci. Rep. **6**, 24454 (2016)
4. Ao, D.C.F., Silva, A.C., de Paiva, A.C., et al.: Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a genetic algorithm, and SVM. Med. Biol. Eng. Comput. **55**, 1129–1146 (2017)
5. Silva, E.C.D., Silva, A.C., de Paiva, A.C.D., et al.: Diagnosis of lung nodule using Moran's index and Geary's coefficient in computerized tomography images. Pattern Anal. Appl. **11**, 89–99 (2007)
6. Sahiner, B., Chan, H.P., Petrick, N., et al.: Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Med. Phys. **25**, 516–526 (1998)
7. Way, T.W., Hadjiiski, L.M., Sahiner, B., et al.: Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Med. Phys. **33**, 2323–2337 (2006)
8. Zhang, G., Xiao, N., Guo, W.: Spiculation quantification method based on edge gradient orientation histogram. In: International Conference on Virtual Reality and Visualization, pp. 86–91. IEEE Press, New York (2014)
9. Dasarathy, B.V., Holder, E.B.: Image characterizations based on joint gray level—run length distributions. Pattern Recogn. Lett. **12**, 497–502 (1991)
10. Brodić, D., Amelio, A., Milivojević, Zoran N.: Classification of the scripts in medieval documents from balkan region by run-length texture analysis. In: Arik, S., Huang, T., Lai, W.K., Liu, Q. (eds.) ICONIP 2015. LNCS, vol. 9489, pp. 442–450. Springer, Cham (2015). doi:10.1007/978-3-319-26532-2_48
11. Way, T.W., Sahiner, B., Chan, H.P., et al.: Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med. Phys. **36**, 3086–3098 (2009)
12. Armato, S.G., McLennan, G., Bidaut, L., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**, 915–931 (2011)
13. Opulencia, P., Channin, D.S., Raicu, D.S., et al.: Mapping LIDC, RadLex, and lung nodule image features. J. Digit. Imaging **24**, 256–270 (2011)