Content-Based Image Retrieval System for Pulmonary Nodules: Assisting Radiologists in Self-Learning and Diagnosis of Lung Cancer

Ashis Kumar Dhara¹ · Sudipta Mukhopadhyay¹ · Anirvan Dutta² ·
Mandeep Garg³ · Niranjana Khandelwal³

Published online: 27 September 2016
© Society for Imaging Informatics in Medicine 2016

Abstract Visual information of similar nodules could assist the budding radiologists in self-learning. This paper presents a content-based image retrieval (CBIR) system for pulmonary nodules, observed in lung CT images. The reported CBIR systems of pulmonary nodules cannot be put into practice as radiologists need to draw the boundary of nodules during query formation and feature database creation. In the proposed retrieval system, the pulmonary nodules are segmented using a semi-automated technique, which requires a seed point on the nodule from the end-user. The involvement of radiologists in feature database creation is also reduced, as only a seed point is expected from

radiologists instead of manual delineation of the boundary of the nodules. The performance of the retrieval system depends on the accuracy of the segmentation technique. Several 3D features are explored to improve the performance of the proposed retrieval system. A set of relevant shape and texture features are considered for efficient representation of the nodules in the feature space. The proposed CBIR system is evaluated for three configurations such as configuration-1 (composite rank of malignancy “1”, “2” as benign and “4”, “5” as malignant), configuration-2 (composite rank of malignancy “1”, “2”, “3” as benign and “4”, “5” as malignant), and configuration-3 (composite rank of malignancy “1”, “2” as benign and “3”, “4”, “5” as malignant). Considering top 5 retrieved nodules and Euclidean distance metric, the precision achieved by the proposed method for configuration-1, configuration-2, and configuration-3 are 82.14, 75.91, and 74.27 %, respectively. The performance of the proposed CBIR system is close to the most recent technique, which is dependent on radiologists for manual segmentation of nodules. A computer-aided diagnosis (CAD) system is also developed based on CBIR paradigm. Performance of the proposed CBIR-based CAD system is close to performance of the CAD system using support vector machine.

✉ Sudipta Mukhopadhyay
smukho@ece.iitkgp.ernet.in

Ashis Kumar Dhara
dear.ashis79@gmail.com

Anirvan Dutta
anirvan10006.13@bitmesra.ac.in

Mandeep Garg
gargmandeep@hotmail.com

Niranjana Khandelwal
khandelwaln@hotmail.com

¹ Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

² Electronics and Communication Engineering, Birla Institute of Technology Mesra, Ranchi 835215, India

³ Department of Radio Diagnosis & Imaging, Post Graduate Institute of Medical Education and Research, Chandigarh 160 012, India

Keywords CT images · Content-based image retrieval · Diagnosis of lung cancer · Lung cancer · Pulmonary nodules · Self-learning tool of radiology · CBIR based CAD system

Introduction

Lung cancer accounts for the highest number of cancer-related deaths than any other cancer in both men and women

[1]. At present, there is no effective way to prevent lung cancer except campaigning against smoking. Pulmonary nodules are the potential manifestation of lung cancer [3]. The pulmonary nodules are blob-like structures with the maximum diameter 3 to 30 mm. Accurate interpretation of pulmonary nodules is essential for diagnosis of lung cancer and subsequent plan of treatment. Trainee radiologists have to depend on experienced professionals to enrich their knowledge. The lack of time of experienced radiologists is the major bottleneck for such traditional learning procedure. A content-based image retrieval (CBIR) system could assist the radiologists using the examples of similar nodules. The visual information content in the retrieved nodules could help the radiologists to interpret the corresponding query nodule and its clinical management (biopsy and follow-up scan). Large number of lung CT images are generated by hospitals and clinics everyday, and the images are stored in picture archival and communication system (PACS). The images stored in the PACS system could be used for the development of CBIR system and CBIR-based computer-aided diagnosis (CAD) system.

The works on CBIR system of pulmonary nodules considered a representative slice of the nodule [4, 5] instead of the whole nodule. Moreover, the prior works depend on radiologists for manual segmentation of nodules. The retrieval system with minimal user intervention is desired for clinical use. The performance of retrieval system is also important for its applicability. The aim of the present system is to reduce the user intervention as well as to improve the accuracy of retrieval. So far, the systems can be used by experts viz. radiologists. For the first time, it will be possible for naive users to use the CBIR system. The nodule segmentation technique used in retrieval system [6] require only a seed point from the end user. Thus, involvement of radiologists is reduced in query formation as well as feature database creation. Several shape-based, margin-based, and texture-based features are investigated to improve the accuracy of retrieval. The most relevant feature set is determined using minimum-redundancy-maximum-relevancy-based feature selection technique. The proposed system is applicable for different types of pulmonary nodules based on their internal texture (viz. solid, part-solid, and non-solid) and external attachment (viz. juxta-pleural and juxta-vascular). The most recent CBIR system of Seitz et al. [5] used the boundary information of nodules from radiologists. Therefore, the technique is considered as the gold standard. The retrieval accuracy of the proposed system is close to the gold standard technique for all configurations.

Early diagnosis of lung cancer could improve 5-years survival rate up to 80 % from 15 % [2]. The screening of lung CT images could substantially reduce the mortality rate due to lung cancer [2]. In diagnostic decision making, a radiologist typically compares the present case with

few closed cases from a medical archive. A large and ever growing number of digital images demands efficient use of relevant cases in diagnostic decision. Several CAD system have been developed in mammography using certain number of retrieved images [7–9]. A CBIR-based CAD system is also developed for diagnosis of lung cancer. The accuracy of classification of the proposed CBIR-based CAD system is close to accuracy of the CAD system using support vector machine (SVM).

The paper is organized into several sections: prior works are reviewed in “[Background](#)” section, proposed CBIR system is described in “[Methodology](#)” section, results of retrieval are provided in “[Results and Discussion](#)” section, and the conclusion and future scope of improvement are stated in “[Conclusion](#)” section.

Background

Prior Works on CBIR Systems

Müller et al. [10] reported the relevance of CBIR techniques in clinical decision making, medical education, and research. These benefits motivated researchers either to develop a general-purpose CBIR system or to develop a specialized CBIR system to support the retrieval of various kinds of medical images. Image retrieval in medical applications (IRMA) [11] presents a general CBIR approach for medical images considering various imaging modalities. IRMA combines a central database with a distributed system architecture and supports rapid prototyping and quick integration of novel image analysis methods. Development of algorithms for retrieval of similar images considering structural contents is a significant research challenge because regions of interest are commonly irregular, overlapping, partially occluded, or highly localized [10]. Therefore, the research objective is to develop algorithms that retrieve semantically and perceptually similar images to provide diagnostic decision support.

Several retrieval systems have been developed for high-resolution CT (HRCT) images of lungs. A few well-known examples are ASSERT [12], medGIFT [13], and Comparison Algorithm for Navigating Digital Image Databases (CANDID) system [14]. The CANDID system computes a global signature of each image stored in the database that can represent features such as texture, shape, and colour. ASSERT uses a physician-in-the-loop approach for retrieving HRCT images of lungs. This framework is semi-automatic and requires a physician to outline the pathology-bearing regions and identify certain anatomical landmarks. Visual contents like texture, shape, edges, and gray-scale properties of pathology-bearing regions are used for feature-based representation. The medGIFT retrieval

system adapted the open-source image finding tool integrated with a PACS system [15]. It uses combinations of textual labels and visual features for retrieval of a variety of images such as CT, magnetic resonance imaging (MRI), and color photos. The works of Lam et al. [4] and Seitz et al. [5] focused on the CBIR system of pulmonary nodules.

Lam et al. [4] developed a CBIR system named ‘BRISC’ for retrieval of CT image slices of pulmonary nodules from a lung CT image database. The pulmonary nodules are segmented using the boundary information of radiologists. Different texture features (Haralick, Gabor filter, and Markov) are extracted from each CT image slice, representing the nodule. The retrieved nodules for a query nodule are considered to be relevant if the retrieved nodules are other instances (slices) of that same nodule, or the same slice rated by a different radiologist. This method was validated on a database of 2424 images of 141 nodules and achieved a precision of 88 % considering one retrieved image. Though this work is an important milestone in the development of CBIR system of pulmonary nodules, it does not help in self-learning or differential diagnosis.

Seitz et al. [5] designed a CBIR system considering nodules with composite rank of malignancy ratings of 1 or 2 as benign, 4 or 5 as malignant, and 3 as unknown. The pulmonary nodules are segmented by the radiologists. Total 64 image features were extracted from the biggest representative slice of the nodule. The Euclidean distance was used for the similarity measure. They reported the precision of 82.77, 82.57, 81.94, 80.90, and 79.23 % considering 3, 5, 10, 20, 50 top retrieved images, respectively. This tool is important for self-learning and differential diagnosis, but its use is limited as the radiologists need to segment nodule manually. Moreover, manual segmentation is time-consuming and errors prone.

Prior Works on CBIR-Based CAD System

Tourassi et al. [7] developed a CBIR algorithm using mutual information as the similarity metric. The database consists of 1009 mammography region of interest (ROIs) with 681 ROIs containing mass and 328 ROIs containing normal mammography region. Top 302 retrieved images were considered for decision index. The area (A_z) under receiver operating characteristics (ROC) curve of their CBIR-based CAD is 0.87 ± 0.01 .

Jin et al. [8] used 24 shape and texture-based features in their CBIR-based CAD system. The database consists of 583 ROIs containing masses and 1417 ROIs containing normal breast region. The reciprocal of Euclidean distance is taken as the similarity measure between two ROI images. Class label of the query nodule is determined based on 120 top retrieved images. The A_z achieved by this CAD system is 0.84.

Jiang et al. [9] noticed that the reported that the features used in CBIR systems fall short of scalability in the retrieval stage. Scale-invariant feature transformation was used to overcome this drawback. Vocabulary tree framework was used for efficient retrieval with improved precision. Euclidean distance metric is used for similarity measure between two ROI images. The purpose of this work is to distinguish between mass and normal tissue. A query ROI image is classified according to its best matched database ROIs using majority voting. The CAD system was validated on a large data set of 11,553 ROIs (2340 ROIs containing mass and 9213 ROIs containing normal breast region) extracted from the digital database for screening mammography. Considering top 5 retrieved images, the classification accuracy achieved by the CAD system is 90.8 %.

Methodology

The purpose of the proposed CBIR system is to retrieve similar nodules from a large CT image database. The user provides a seed point for a nodule to indicate it as a query nodule. A cubic volume of interest (VOI) of size (40 mm $\times$ 40 mm $\times$ 40 mm) is selected considering the seed point as the centroid of the VOI. In the next step, nodules are segmented using the method of Dhara et al. [6]. The features are extracted from the segmented nodule, and corresponding feature vector of the query nodule is formed using these features. The feature database is created using the feature vectors of all pulmonary nodules present in the database. The creation of the feature database and the nodule database is an off-line task. The feature vector of the query nodule is compared with each feature vector in the feature database. The comparison of feature vectors is performed using a distance metrics (viz. Euclidian, Manhattan, and Mahalanobis distance). The minimum value of distance metrics indicates the highest degree of similarity. The retrieved nodules are presented to the user in the descending order of similarity measure. The block diagram of CBIR system is given in Fig. 1. Different steps of the proposed CBIR system are the segmentation of the query nodule, feature extraction

Fig. 1 Block diagram of CBIR system of pulmonary nodules

from the segmented nodule, finding similar nodules from database, and retrieval of similar nodules.

Segmentation of Pulmonary Nodules

Pulmonary nodules are segmented using the method of Dhara et al. [6], which is applicable for different types of pulmonary nodules based on their internal texture (viz. solid, part-solid, and non-solid) and external attachment (viz. juxta-pleural and juxta-vascular). In this technique, pulmonary nodules are classified into solid/part-solid or non-solid by analyzing intensity distribution of nodule core. Two separate algorithms are set for segmentation of solid/part-solid and non-solid nodules. The performance of the segmentation technique of Dhara et al. [6] was compared with the techniques of Kuhnigh et al. [16], Moltz et al. [17], and Kubota et al. [18] in terms of four contour-based metrics (mean distance, Pratt function, Hausdorff distance, and modified Hausdorff distance) and six region-based metrics (accuracy, overlap, sensitivity, specificity, similarity angle, and similarity region) [19]. The results proved that method of Dhara et al. [6] is reliable for segmentation of various types of nodules with an improved accuracy.

Representation of Pulmonary Nodules

Several imaging characteristics of nodule such as size, shape, margin sharpness, internal texture, and presence of calcification are important to the radiologists for estimation of malignancy [20]. Several 3D features are used to represent detailed structure the nodules. The 3D features are highly affected by segmentation error. The affect of nodule segmentation technique is very small on the biggest representative slice of a nodule. Therefore, several 2D features are also computed from the biggest representative slice of the nodule. Each pulmonary nodule is represented in multidimensional feature space using a feature vector. Each feature vector is linked with the corresponding subject in the CT image database. According to the nodule characteristics, the features can be grouped into shape-based, margin-based, and texture-based. A few features are not scale invariant. These features are appropriately multiplied with the voxel resolution of nodule.

3D Shape-Based Features

Several 3D shape features are computed using the binary mask of the segmented nodule. The procedure of computing these shape features are described below.

- Sphericity: Irregularity of the shape of a nodule is represented by sphericity and it is defined as [21]

$$\text{Sphericity} = 6\sqrt{\pi} V A^{-3/2}, \quad (1)$$

where V is the volume of the nodule and A is its surface area. The volume of a nodule is computed by multiplying total number of voxel with size of each voxel. The surface area of a nodule is computed by summing the area of all triangles in the nodule mesh. The nodule mesh is generated using marching cubes algorithm [22].

- Spiculation: Spicules are associated with the radial extension of malignant cells and appear as spokes of a wheel. Spiculation of a nodule is computed using the method of Dhara et al. [23]. Several steps of computation of spiculation are identification of tip of each spicule, determination of base area of each spicule, and computation of spiculation of each spicule. The net spiculation of a nodule is computed in two different ways

$$\text{Spiculation} = \sum_{i=1}^N e^{-\omega_i} h_i, \quad (2)$$

and,

$$\text{Spiculation} = \frac{\sum_{i=1}^N h_i \cos \omega_i}{\sum_{i=1}^N h_i}, \quad (3)$$

where ω_i is the solid angle subtended at peak point of i th spicule, h_i is the height of a spicule, N is the total number of spicules of a nodule under consideration.

- Lobulation: Lobulated margin of a nodule is caused by uneven growth. Lobulation is represented using the ratio of total concave surface area and total convex surface area of a nodule [23]. Several steps of computation of lobulation are triangular mesh generation and determination of concave and convex part of the triangular mesh. Lobulation of a nodule is defined as

$$\text{Lobulation} = \frac{S_{\text{concave}}}{S_{\text{convex}}}. \quad (4)$$

where S_{convex} is total convex surface area and S_{concave} total concave surface area of the nodule.

- Volume: Volume of a nodule is computed by multiplying total number of voxel with size of each voxel [23].
- Surface area: Surface area of a nodule is computed by summing area of all triangles in nodule mesh.
- Equivalent diameter 3D: Equivalent diameter of a nodule is defined as [21]

$$D = 2\sqrt[3]{\frac{3V}{4\pi}}, \quad (5)$$

where V is volume of the nodule.

- Major axis length 3D: The major axis length of the bounding ellipsoid of nodule [17].
- Minor axis length 3D: The minor axis length of the bounding bounding ellipsoid of nodule [17].

- Convex surface area: Surface area of convex hull of the nodule, where the surface area is computed by summing the area of all triangles in the equivalent triangular mesh of the convex hull of the nodule.

2D Shape-Based Features

Several 2D shape features are also computed from the biggest axial slice of the nodules. The foreground region in the biggest representative slice of the nodule is denoted as S_{biggest} . Shape features are computed based on the work of Seitz et al. [5]. The following shape features are used in the proposed retrieval system.

- Area: Number of pixel in S_{biggest} multiplied by pixel resolution.
- Perimeter: Perimeter of S_{biggest} multiplied by pixel resolution.
- Equivalent diameter 2D: It is the equivalent diameter of S_{biggest} , and represented as $\frac{2}{\pi}\sqrt{\text{Area}}$
- Convex area: Area of convex hull of S_{biggest}
- Convex perimeter: Perimeter of the convex hull of S_{biggest}
- Compactness: It is defined as $\frac{\text{Perimeter}^2}{4\pi \text{Area}}$
- Major axis length: The major axis length of the bounding ellipse of S_{biggest}
- Minor axis length: The minor axis length of the bounding ellipse of S_{biggest}
- Circularity: The circularity is defined as

$$\text{Circularity} = \frac{4\pi A}{q^2}, \quad (6)$$

where A is area of S_{biggest} and q is convex perimeter of S_{biggest} .

- Standard deviation of radial distance 2D: Standard deviation of radial distances of all boundary pixels S_{biggest} .

Margin-Based Features

Margin sharpness is the difference in intensity of the boundary region of a nodule and its surroundings. Margin sharpness is represented using the combination of acutance and histogram spread of averaged gradient of the nodule [24]. Acutance of a nodule is computed using average gradient of all boundary pixels for all the slices containing the nodule. The averaged gradient for K th boundary normal is represented as [25]

$$f_d(K) = \frac{1}{N} \sum_{k=1}^N \frac{(f(k) - b(k))}{2k}. \quad (7)$$

where $f(k)$, $b(k)$ are part of boundary normal, located in the foreground and background of the nodule. The acutance of a nodule is defined as

$$\text{Acutance of nodule} = \frac{1}{d_{\max}} \left[\frac{1}{K} \sum_{k=1}^K f_d^2(k) \right]^{\frac{1}{2}}, \quad (8)$$

where $d_{\max}$ is the maximum value of averaged gradient. Histogram spread of average gradient (HSAG) of nodule is defined as

$$\text{HSAG} = \frac{(\text{3rd} - \text{1st}) \text{ of histogram of } N_{\text{-gradient}}}{(\text{maximum} - \text{minimum}) \text{ of } N_{\text{-gradient}}}. \quad (9)$$

The histogram spread was introduced by Tripathi et al. [26] and its range is (0, 1).

2D Texture-Based Features

- Haralick features: Haralick features are calculated from gray-level co-occurrence matrix (GLCM) [5].

Fig. 2 Nine directions for computation of GLCM matrix

Table 1 Composition of pulmonary nodules based on composite rank of malignancy

Malignancy rating	Nodules of rank-1	Nodules of rank-2	Nodules rank-3	Nodules of rank-4	Nodules of rank-5
Solid	120	150	320	117	121
Part-solid	0	6	13	7	6
Non-solid	0	3	16	9	3
Total	120	159	349	133	130

Several directions (0° , 45° , 90° , 135°) and distances (1, 2, 3, 4) pixels are used to generate separate GLCM. Six Haralick's feature [27] (viz. entropy, energy, inverse difference moment, sum entropy, difference entropy, and contrast) are computed from the biggest representative slice. The mean of each Haralick feature computed from several GLCM is used in the proposed retrieval system.

- Gabor features: Gabor filter is a method which extracts texture information from an image in the form of a response image. Twelve Gabor filters are obtained by considering four orientations (0° , 45° , 90° , 135°) and three frequencies (0.3, 0.4, and 0.5) radians. Gabor filters are convolved with the biggest representative slice of the nodule. The standard deviations from the 12 response images results in 12 Gabor features per nodule [5].
- Histogram of oriented gradient (HOG): Dalal et al. [28] introduced HOG features to represent local object appearance and shape using the distribution of local intensity gradient. The HOG features are computed from the biggest representative slice of the nodule. The mean, variance, and standard deviation of HOG features are used in the proposed retrieval system.

3D Texture-Based Features

3D Haralick's features are computed using the method of Han et al. [29]. Nine GLCM was computed for each nodule considering nine directions as described Fig. 2. Fourteen Haralick features [27] were computed from each

GLCM matrix. The maximum correlation coefficient is not considered in the proposed retrieval due to large computational time [29]. The mean and range of each Haralick feature over the nine directions are computed as

$$H^k_{\text{mean}} = \frac{1}{n} \sum_{i=1}^9 H_{ik}, \quad (10)$$

$$H^k_{\text{range}} = \max_{i=1:9} \{H_{ik}\} - \min_{i=1:9} \{H_{ik}\}, \quad (11)$$

where H_{ik} is the k th Haralick feature considering i th direction. Total number of 3D Haralick's features are 26.

Feature Selection

The fewer dimension of feature vector indicates less costs of similarity calculation of every search. The relevant feature are identified by considering area (AUC) under the receiver operating characteristics and p values. The AUC is computed using ROC Toolkit [30], and p values is computed by performing two-tailed Student's t test. Minimal-redundancy-maximal-relevance (MRMR) [31] technique is used for further refinement of feature set. The list of selected features are provided in “Results of Feature Selection”.

Similarity Measure

The similarity between two images is computed by comparing their feature vectors. Several distance metrics (Euclidean, Manhattan, and Chebyshev) are used for similarity measure in feature space. Distance metrics between the feature vectors $\mathbf{A}$ ($a_1, a_2, a_3, \dots, a_n$) and $\mathbf{B}$ ($b_1, b_2, b_3, \dots, b_n$) are represented as

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}, \quad (12)$$

$$\text{Manhattan distance} = \sum_{i=1}^n |a_i - b_i|, \quad (13)$$

Table 2 Configurations used for the evaluation of the proposed classification scheme and the competing techniques

Description	Configuration	Benign	Malignant
Composite rank of malignancy 1, 2 as benign, 4, 5 as malignant, and 3 is neglected	1	279	263
Composite rank of malignancy 1, 2, 3 as benign, 4, 5 as malignant	2	628	263
Composite rank of malignancy 1, 2 as benign, 3, 4, 5 as malignant	3	279	612

Table 3 Selected features with A_z and p value

Type	Sl. no.	Feature name	A_z	p values
3D shape	1	Volume	0.88	7.73E-28
	2	Surface area	0.87	9.62E-29
	3	Convex surface area	0.85	1.11E-26
	4	Lobulation	0.81	1.02E-09
	5	Major axis length 3D	0.81	3.06E-22
	6	Spiculation ₁	0.78	1.03E-23
	7	Sphericity	0.67	2.90E-07
	8	Spiculation ₂	0.57	1.12E-13
2D shape	9	Minor axis length	0.89	5.12E-52
	10	Area	0.89	1.83E-38
	11	Convex perimeter	0.87	9.25E-43
	12	Perimeter	0.87	2.92E-41
	13	Major axis length	0.85	1.29E-33
	14	Circularity	0.71	1.02E-15
	15	Compactness	0.70	2.20E-06
	16	Roughness	0.65	0.01
Margin	17	HSAG	0.68	1.26E-10
	18	Acutance 3D	0.62	1.58E-05
Haralick 3D	19	Mean information measure of correlation ¹	0.92	2.6E-73
	20	Range sum entropy	0.92	1.16E-95
	21	Mean inverse difference moment	0.88	7.46E-40
	22	Mean angular second moment	0.87	4.16E-37
	23	Range angular second moment	0.87	4.99E-38
	24	Mean entropy	0.87	4.61E-57
	25	Range sum average	0.80	2.42E-31
	26	Mean contrast	0.79	1.21E-24
	27	Mean sum entropy	0.76	1.43E-23
	28	Range difference entropy	0.74	9.46E-18
	29	Range difference variance	0.71	5.56E-17
	30	Range sum squares of variance	0.70	7.11E-14
	31	Range inverse difference moment	0.67	5.07E-08
	32	Mean difference variance	0.64	4.33E-19
	33	Range entropy	0.60	1.5E-03
Haralick 2D	34	Entropy	0.89	7.29 E-70
	35	Energy	0.88	3.68E-29
	36	Inverse difference moment	0.88	7.83E-43
	37	Sum entropy	0.88	6.38E-64
	38	Difference entropy	0.85	4.98E-55
	39	Mean of HOG	0.89	2.70E-66
HOG	40	Variance of HOG	0.84	1.79E-52
	41	Standard deviation HOG	0.84	6.75E-53
Gabor 2D	42	GaborSD_45_03	0.67	5.60E-23
	43	GaborSD_135_03	0.67	5.60E-23
	44	GaborSD_0_03	0.67	5.82E-23
	45	GaborSD_90_03	0.67	6.26E-23
	46	GaborSD_135_04	0.67	6.19E-23
	47	GaborSD_45_04	0.67	6.22E-23
	48	GaborSD_0_04	0.67	7.71E-23

Table 3 (continued)

Type	Sl. no.	Feature name	A_z	p values
	49	GaborSD_45_05	0.67	8.60E-23
	50	GaborSD_90_05	0.67	2.66E-22
	51	GaborSD_0_05	0.67	1.37E-22

$$\text{Chebyshev distance} = \max_i(|a_i - b_i|), \quad (14)$$

where n is the dimension of each feature vector.

Retrieval of Similar Nodules

For a given query nodule, the similar nodules are retrieved by comparing the feature vector of query nodule with all feature vectors stored in the feature database. The retrieved nodules are ranked in the descending order of similarity. The number of retrieved nodules are decided by the radiologists.

CBIR-Based CAD System

Based on the work of Jin et al. [8], a CBIR-based CAD system is developed for pulmonary nodules. A query nodule results a series of retrieved nodules along with their class label. The class label of the retrieved nodules is used to determine the decision index (DI) for nodule classification. A higher DI value means a higher probability that the

query nodule is malignant. For a query nodule (Y_q), DI is computed as [8]

$$DI(Y_q) = \frac{\sum_{I=1}^M \{SIM(Y_q, X_I) \times \text{Rank}(X_I)\}}{\sum_{I=1}^M \{SIM(Y_q, X_I) \times \text{Rank}(X_I)\} + \sum_{J=1}^N \{SIM(Y_q, X_J) \times \text{Rank}(X_J)\}} \quad (15)$$

where M is the number of malignant nodules and N is the number of benign nodules among top K retrieved images. $\text{Rank}(X_I)$ is the ordering number of X_I , when the retrieved images are sorted in descending order.

Results and Discussion

Database Description

The proposed CBIR system and the competing technique are evaluated on 891 nodules of LIDC/IDRI [32].

Table 4 Rejected features with A_z and p value

Method	Sl. no.	Feature name	A_z	p values
MRMR technique	1	Range contrast	0.79	1.65E-24
	2	Range sum variance	0.68	3.67E-15
	3	Convex area	0.88	8.85E-32
	4	GaborSD_90_04	0.67	1.14E-23
	5	GaborSD_135_05	0.67	8.27E-23
	6	Equivalent diameter	0.89	1.51E-55
	7	Contrast	0.83	5.64E-25
	8	Minor axis length 3D	0.88	1.71E-55
	9	Equivalent diameter 3D	0.88	3.83E-56
	10	Mean sum variance	0.64	4.57E-19
Based on A_z and p value	1	Mean information measure of correlation ²	0.80	0.29
	2	Range information measure of correlation ²	0.80	0.29
	3	Range correlation	0.59	4.44E-5
	4	Range information measure of correlation ¹	0.59	1.25-4
	5	Range correlation	0.57	1.3E-3
	6	Mean sum squares of variance	0.54	5.36E-8
	7	Mean sum average	0.53	0.16
	8	Mean difference entropy	0.51	0.24

Fig. 3 Panels **a–d** represent the retrieved nodules corresponding to four query nodules with composite rank of malignancy: 5, 4, 1, and 2. The composite ranks of malignancy of the query and retrieved nodules are provided below the corresponding nodules for better understanding. The *top row* presents the query nodule. Top five retrieved nodules using the proposed method and the method of Seitz et al. are provided in *middle and bottom row*. The irrelevant retrieved nodules are marked with *red*

The rank of nine diagnostic characteristics of all nodules for each subject is provided in an XML file. Diameter of the nodules lies in the range of [3–30] mm. The nodules with texture index 1 are considered as non-solid, texture index 2 or 3 are considered as part-solid, and texture index 4 or 5 are considered as solid. The result of the biopsy of pulmonary nodules is not available in LIDC/IDRI. A group of four radiologists have provided the rank of malignancy of each nodule on a scale of 1:5, where rank-1 indicates the least chance of malignancy and rank-5 indicates highly suspicious for malignancy. The composite rank of malignancy is the mode of rating of four radiologists. For multiple modes, the floor of the median of rating of four radiologists is considered as the composite rank of malignancy [33]. The composition of nodules based on the composite rank

of malignancy is provided in Table 1. Han et al. [29] considered three different configurations (see Table 2) for the evaluation of their CAD system. The same configurations are used for the evaluation of proposed CBIR system and CBIR-based CAD system.

Performance Evaluation Metric

In the present work, each pulmonary nodule is labeled as benign or malignant based on the composite rank of malignancy. The proposed retrieval system focuses on the retrieval of similar nodules for a query benign or malignant nodule. The end-user of the CBIR system are radiologists and they are interested on few top retrieved nodules correspond to a query nodule. Therefore, performance of the

Fig. 3 (continued)

proposed retrieval system is evaluated in terms of precision instead of precision versus recall curve. The range of precision is $[0,1]$, and it is defined as

$$\text{Precision} = \frac{\text{Number of relevant nodules}}{\text{Number of retrieved nodules}}. \quad (16)$$

Results of Feature Selection

Out of 68 features, a set of 60 features is considered as relevant because A_z is ≥ 0.60 and p value is less than 0.05 for those 60 features. Minimal-redundancy-maximal-relevance [31] based feature selection technique is used to refine feature list by removing the uncorrelated features. Total number of selected features by this technique is 50, and these features are used in proposed CBIR system. The list of selected features and rejected features are provided in

Tables 3 and 4, respectively. The spiculation of nodule is an important discriminating feature as it provides A_z of 0.78 using Spiculation₁ (2). The values of spiculation are more reproducible using Spiculation₂ (3) and hence, it is used in the proposed retrieval system.

Qualitative Results of Retrieval

The results of retrieval for the proposed system and the gold standard technique are provided in Fig. 3 for four query nodules. The composite rank of malignancy of the query and top five retrieved nodules are displayed for better understanding. The proposed system and the gold standard technique provides relevant retrieved nodules for a malignant query nodule with composite rank of malignancy: 5 (Fig. 3a) and a benign query nodule with composite rank of malignancy: 1 (Fig. 3c). The third retrieved nodule by the proposed

Table 5 Performance of the proposed CBIR system, automated Seitz et al., and the gold standard technique for different configuration

Config- uration	Metric	Method	$P@1$	$P@3$	$P@5$	$P@7$	$P@9$	$P@11$	$P@13$	$P@15$
C_1	Euclidean	Proposed	83.76	82.41	82.14	82.46	82.30	81.96	82.23	82.12
		Automated	83.21	82.35	81.85	81.31	81.02	81.05	80.83	80.84
		Seitz et al.	85.61	84.69	84.28	83.95	83.91	83.33	82.90	82.60
	Manhattan	Proposed	84.31	82.28	81.80	82.20	82.67	82.50	82.60	82.55
		Automated	82.84	81.06	81.44	81.29	81.02	81.18	81.01	80.80
		Seitz et al.	85.79	84.62	83.95	83.16	83.05	82.66	82.37	82.25
	Chebyshev	Proposed	83.94	82.65	82.36	81.94	81.46	81.16	81.29	81.00
		Automated	79.89	79.77	78.75	78.44	77.98	78.08	77.87	77.71
		Seitz et al.	82.29	80.44	79.59	79.18	78.80	78.40	78.10	77.69
C_2	Euclidean	Proposed	76.54	76.73	75.91	75.88	75.68	75.49	75.40	75.21
		Automated	75.08	75.46	75.35	74.78	74.98	74.71	74.68	74.66
		Seitz et al.	78.68	76.28	76.18	78.89	76.14	75.84	75.74	75.82
	Manhattan	Proposed	77.10	76.05	75.96	75.93	75.84	75.67	75.72	75.68
		Automated	76.30	74.97	75.44	74.75	74.82	74.90	74.72	74.48
		Seitz et al.	76.54	76.66	76.27	76.03	75.72	75.44	75.52	75.50
	Chebyshev	Proposed	77.53	76.35	75.06	75.24	75.18	75.23	75.15	74.85
		Automated	73.96	74.41	74.39	73.74	73.74	73.42	73.13	72.90
		Seitz et al.	76.77	74.64	74.52	74.12	74.16	73.83	73.69	73.51
C_3	Euclidean	Proposed	74.29	74.82	74.27	74.33	74.09	74.02	73.66	73.54
		Automated	75.53	73.96	73.40	73.54	73.25	73.43	73.71	73.36
		Seitz et al.	78.10	76.51	75.51	75.32	74.74	74.60	74.51	74.32
	Manhattan	Proposed	76.99	74.22	74.19	73.62	73.89	74.05	73.84	73.87
		Automated	75.42	74.49	73.83	73.74	73.65	73.81	73.65	73.55
		Seitz et al.	77.10	76.06	75.35	75.34	74.69	74.14	73.69	73.72
	Chebyshev	Proposed	74.18	74.22	73.75	72.99	72.84	72.71	72.40	72.30
		Automated	75.08	73.21	72.46	72.44	72.33	72.05	71.67	71.63
		Seitz et al.	71.94	73.06	73.54	73.14	72.98	73.02	72.88	72.52

$P@n$ represents the average precision considering “n” top retrieved images. The values of “n” are 1, 3, 5, 7, 9, 11, 13, and 15. The configuration-1, configuration-2, and configuration-3 are denoted as C_1 , C_2 , and C_3 , respectively

system and second retrieved nodule by the gold standard technique are not relevant to the malignant query nodule with composite rank of malignancy: 4 (Fig. 3b). The second retrieved nodule by the proposed system and fifth retrieved nodule by the gold standard technique are not relevant to the benign query nodule with composite rank of malignancy: 2 (Fig. 3d). Base on the retrieval results of Fig. 3, it is seen that

the retrieval performance of the proposed system is similar to the gold standard technique.

Quantitative Results of Retrieval

The precision for each query nodule is noted for different number of retrieved images. All nodules present in the data

Table 6 Variation of A_z with number of top retrieved images

K	A_z for configuration-1	A_z for configuration-2	A_z for configuration-3
100	0.9253	0.8609	0.8019
150	0.924	0.8526	0.8006
200	0.9247	0.8502	0.8068
250	0.9141	0.8496	0.8007
300	0.9091	0.8495	0.8005
350	0.9038	0.8514	0.7921
400	0.7745	0.8445	0.7940

Here, K is the number of top retrieved images

The ‘bold’ is used to represent the maximum value of each column

set are used as the query and corresponding precision value for the different number of retrieved nodules are noted. This procedure is repeated for various distance metrics such as Euclidian, Manhattan, and Chebyshev distances. The method of Seitz et al. is automated using the segmentation technique of Dhara et al. and named as automated Seitz et al. The average precision of the proposed system, automated Seitz et al., and the gold-standard technique are computed for different distance metrics as provided in Table 5.

In the proposed system, the pulmonary nodules are represented using 50 features, whereas the gold standard technique and automated Seitz et al. considered 64 features. The results in Table 5 depict that the gold-standard technique outperforms automated Seitz et al. and the proposed retrieval system. The gold-standard technique makes use of the segmented result provided by the experienced radiologists. On the other hand, the proposed technique and automated Seitz et al. use an established segmentation technique, and they are vulnerable to the segmentation error. Hence, the gold standard technique outperforms automated Seitz et al. though both of them use the same set of features. The proposed retrieval system considers several 2D shape-based, 3D shape-based, margin-based, and texture-based features of nodules. Though the proposed retrieval system is affected by same automated segmentation scheme, it outperforms the automated Seitz et al. due to the inclusion of more relevant features. The precision is

Table 7 Results of classification for CBIR-based CAD system and CAD system using SVM for configuration-1, configuration-2, and configuration-3

Method	A_z in configuration-1	A_z in configuration-2	A_z in configuration-3
CBIR-based CAD (considering $K = 100$)	0.9253	0.8609	0.8019
CAD using SVM	0.9459	0.8801	0.8328

Fig. 4 ROC plot for **a** configuration 1, **b** configuration 2, and **c** configuration 3

higher in configuration-1 as compared to configuration-2 and configuration-3 due to absence of uncertain nodules in configuration-1. The precision of configuration-2 is higher as compared to configuration-3. The fact shows that nodules with composite rank of malignancy “3” have much similarity towards benign category.

The retrieval results using Euclidean and Manhattan distance metrics are comparable for all retrieval techniques. The retrieval performance using Chebyshev distance is lower as compared to Euclidean and Manhattan distance metrics. Euclidean and Manhattan distance metrics give equal weight to each component of the feature vector, whereas Chebyshev distance is determined by the maximum absolute difference among the individual components. Therefore, the retrieval performance is more affected by noisy features, if Chebyshev distance metric is used. It is observed that the performance of the gold standard technique is lower than the proposed technique for Chebyshev distance, unlike the other distance metrics (see Table 5). It indicates that some of the features of the gold standard technique are more noisy than the features used in the proposed technique.

The retrieval results using Euclidian and Manhattan distance metrics are comparable for all retrieval techniques. The retrieval performance using Chebyshev distance is lower as compared to Euclidian and Manhattan distance metrics. Euclidian and Manhattan distance metrics give equal weight to each component of the feature vector, whereas Chebyshev distance is determined by the maximum absolute difference among the individual components. Therefore, the retrieval performance is affected by noisy features, if Chebyshev distance metric is used. It is observed that the performance of the gold standard technique (see Table 5) is lower than the proposed technique for Chebyshev distance, unlike the other distance metrics. It indicates that some of the features of the gold standard technique are more noisy than the features used in the proposed technique.

Results of CBIR-Based CAD System

Each nodules in the database is chosen as a query image. Considering K retrieved images, DI is calculated for the

query nodule using Eq. 15. Performance of the proposed CBIR-based CAD system is evaluated in terms of A_z . The change of A_z with number of top retrieved images is provided in Table 6 for three configurations. The accuracy of classification decreases with the increment of top retrieved images. The classification accuracy is higher in configuration-1 as compared to configuration-2 and configuration-3 due to the absence of uncertain nodules in configuration-1. The classification accuracy in configuration-2 is higher than configuration-3, and the result is consistent with the retrieval accuracy in Table 5. The classification results also suggest that nodules with composite rank of malignancy “3” have much similarity towards benign category.

The performance of CBIR-based CAD system is compared with the CAD system using SVM and the results are provided in Table 7. The ROC plots for CAD system using CBIR technique and SVM classifier are provided in Fig. 4. The classification accuracy of CBIR-based CAD system is close to the classification accuracy of SVM for three configurations.

Computational Time

For the proposed and compering techniques, the mean and standard deviation (SD) of time required for segmentation of pulmonary nodules, computation of features, and retrieval of similar nodules are provided in Table 8. The computational time is determined using a personal computer with Intel quad-core processor of 2.66 GHz and 8 GB RAM. The time of segmentation of nodule is not available for the gold standard technique, as the segmentation was performed by radiologists. In the case of gold standard technique, the time of nodule segmentation is not included. Hence, the retrieval time of the gold standard technique is lower than automated Seitz et al. and the proposed technique. Due to inclusion of segmentation time, the retrieval time of automated Seitz et al. is higher than the gold standard technique. The proposed retrieval system uses the same nodule segmentation scheme as automated Seitz et al. and delayed by the computation of 3D features. The time of retrieval of the proposed

Table 8 Computational time for the proposed system and the competing techniques

Method	Time of segmentation in second (Mean $\pm$ SD)	Time of feature computation in second (Mean $\pm$ SD)	Total time of retrieval in second (Mean $\pm$ SD)
Seitz et al.	Not available	1.94 ± 0.95	1.98 ± 0.95
Automated	2.96 ± 6.05	1.95 ± 0.95	4.96 ± 6.06
Seitz et al.			
Proposed	2.96 ± 6.05	25.33 ± 19.37	28.34 ± 20.82

The retrieval time is computed considering ten top retrieved nodules

system is much higher than the automated Seitz et al. The 3D features have already proved their worth to improve the precision of retrieval. However, speed of retrieval could be reduced by use of GPU.

Conclusion

A retrieval system is developed to assist the budding radiologists in self-learning and differential diagnosis of lung cancer. The proposed system may reduce the inter-observer variability during estimation of malignancy using the information of similar nodules. The performance of the proposed retrieval system is close to the gold standard technique where radiologists have drawn the nodule boundary. While reported CBIR system depends on user for nodule segmentation, the proposed system would be fast and easy to use as the user need to provide only a seed point on the query nodule for retrieval of similar nodules. The CBIR system should be updated from time to time to keep it relevant. Availability of dedicated experienced radiologists is the bottleneck for database up-gradation. The semi-automated nodule segmentation technique paves the procedure of easy database up-gradation. Further work is required to improve the performance of nodule retrieval system. The CBIR-based CAD system could also be used for practice as its performance is close to the CAD system using SVM. To improve the performance of CBIR system and CBIR-based CAD system, more research work is required on segmentation of pulmonary nodules, improvement of feature set, and learning based retrieval.

Compliance with Ethical Standards

Conflict of interests This study was funded by the Department of Electronics and Information Technology, Govt. of India, Grant number 1(2)/2013-ME &TMD/ESDA. The authors declare that they have no conflict of interest. This work is done using a public lung CT image data set and for this type of study formal consent is not required. This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

References

1. Siegel R, Naishadham D, Jemal A: Cancer statistics. *CA Cancer J Clin* 63(1):11–30, 2013
2. Diederich S, Wormanns D, Semik M, Thomas M, Lenzen H, Roos N, Heindel W: Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. *Radiology* 222(3):773–781, 2002
3. Ko JP, Naidich DP: Computer-aided diagnosis and the evaluation of lung disease. *J Thorac Imaging* 19(3):136–155, 2004
4. Lam MO, Disney T, Raicu DS, Furst J, Channin DS: BRISC— an open source pulmonary nodule image retrieval framework. *J Digit Imaging* 20(1):63–71, 2007
5. Seitz Jr KA, Giuca AM, Furst J, Raicu D: Learning lung nodule similarity using a genetic algorithm. *Proceedings of SPIE Medical Imaging 2012*. San Deigo, USA, 2012, pp 831,537–831,537–7
6. Dhara AK, Mukhopadhyay S, Das Gupta R, Garg M, Khandelwal N: Erratum to: a segmentation framework of pulmonary nodules in lung CT images. *J Digit Imaging* 29(1):148–148, 2016
7. Tourassi GD, Vargas-Voracek R, Floyd Jr CE: Content-based image retrieval as a computer aid for the detection of mammographic masses. *SPIE Medical Imaging 2003*, 2003, pp 590–597
8. Jin R, Meng B, Song E, Xu X, Jiang L: Computer-aided detection of mammographic masses based on content-based image retrieval. *SPIE Medical Imaging 2007*, 2007, pp 65,141w–65,141w
9. Jiang M, Zhang S, Li H, Metaxas DN: Computer-aided diagnosis of mammographic masses using scalable image retrieval. *IEEE Trans Biomed Eng* 62(2):783–792, 2015
10. Müller H., Michous N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *Int J Med Inform* 73(1):1–23, 2004
11. Lehmann TM, Schubert H, Keysers D, Kohlen M, Wein BB: The IRMA code for unique classification of medical images. *Proceedings of SPIE Medical Imaging 2003*, 2003, pp 440–451
12. Shyu C, Brodley CE, Kak AC, Kosaka A, Aisen A, Broderick L: Assert: a physician-in-the-loop content-based retrieval system for hrct image databases. *Comput Vis Image Underst* 75(2):111–132, 1999
13. Florea F, Müller H, Rogozan A, Geissbuhler A, Darmoni S: Medical image categorization with MediC and MedGIFT. *Netherlands: Maastricht*, 2006, pp. 3–11
14. Kelly PM, Cannon TM, Hush DR: Query by image example: the comparison algorithm for navigating digital image databases (candid) approach. *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, 1995, pp 238–248
15. Müller H, Lovis C, Geissbuhler A: The MedGIFT project on medical image retrieval. *Medical Imaging and Telemedicine* 2, 2005
16. Kuhnigk JM, Dicken V, Bornemann L, Bakai A, Wormanns D, Krass S, Peitgen HO: Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Trans Med Imaging* 25(4):417–434, 2006
17. Moltz JH, Kuhnigk JM, Bornemann L, Peitgen H: Segmentation of juxtapleural lung nodules in ct scan based on ellipsoid approximation. *Proceedings of First International Workshop on Pulmonary Image Processing 2008*. New York, 2008, pp 25–32
18. Kubota T, Jerebko AK, Dewan M, Salganicoff M, Krishnan A: Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. *Med Image Anal* 15(1):133–154, 2011
19. Silva JS, Santos JB, Roxo D, Martins P, Castela E, Martins R: Algorithm versus physicians variability evaluation in the cardiac chambers extraction. *IEEE Trans Inf Technol Biomed* 16(5):835–841, 2012
20. Kligerman S, White C: Imaging characteristics of lung cancer. *Semin Roentgenol* 46(3):194–207, 2011
21. Sladoje N, Nyström L, Saha PK: Measurements of digitized objects with fuzzy borders in 2D and 3D. *Image Vis Comput* 23(2):123–132, 2005
22. Lorensen WE, Cline HE: Marching cubes: a high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics* 21(4):163–169, 1987

23. Dhara AK, Mukhopadhyay S, Saha P, Garg M, Khandelwal N: Differential geometry-based techniques for characterization of boundary roughness of pulmonary nodules in CT images. *Int J Comput Assist Radiol Surg* 11(3):337–349, 2016
24. Dhara AK, Mukhopadhyay S, Chakrabarty S, Garg M, Khandelwal N: Quantitative evaluation of margin sharpness of pulmonary nodules in lung CT images. *IET Image Process* 10(9):631–637, 2016
25. Rangayyan RM, El-Faramawy NM, Desautels JL, Alim OA: Measures of acutance and shape for classification of breast tumors. *IEEE Trans Med Imaging* 16(6):799–810, 1997
26. Tripathi AK, Mukhopadhyay S, Dhara AK: Performance metrics for image contrast. *Proceedings of IEEE International Conference on Image Information Processing*. Simla, India, 2011, pp 1–4
27. Haralick RM, Shanmugam K, Dinstein IH: Textural features for image classification. *IEEE Trans Syst Man Cybern* 6:610–621, 1973
28. Dalal N, Triggs B, Schmid C: Human detection using oriented histograms of flow and appearance. *Computer Vision–ECCV 2006*. Springer, 2006, pp 428–441
29. Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H, Liang Z: Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *J Digit Imaging* 28(1):99–115, 2014
30. Noessner J, Niepert M, Stuckenschmidt H. ROCKIT: Exploiting parallelism and symmetry for map inference in statistical relational models. arXiv:1304.4379. 2013.
31. Peng H, Long F, Ding C: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238, 2005
32. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Beek EJR, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DPY, Roberts RY, Smith AR, Starkey A, Batra P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Castele AV, Gupta S, Sallam M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY, Clarke LP: 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(2):915–931, 2011
33. Dasovich GM, Kim R, Raicu DS, Furst JD: A model for the relationship between semantic and content based similarity using LIDC. *Proceedings of SPIE Medical Imaging 2010*. San Diego, USA, 2010, pp 762,431–762,431–10