Sematic Characteristics Prediction of Pulmonary Nodule Using Artificial Neural Networks

Guangxu Li, Hyoungseop Kim, Joo Kooi Tan, Seiji Ishikawa,
Yasushi Hirano, Shoji Kido and Rie Tachibana

Abstract— **The purpose of this research is to provide a quantitative approach for finding the relationships between computer-calculated features and medical semantic concepts used for computer-aided diagnosis system. We attempt to provide an objective assessment for supporting the diagnostic characterization using Artificial Neural Networks (ANNs). We used 60 thoracic CT scans collected from the Lung Image Database Consortium (LIDC) database, in which the suspicious lesions had been delineated and annotated by 4 radiologists independently. Two experiments were performed in this study. Correlation analysis experiment was applied to explore the correlations between the image features and semantic characteristics. Agreement experiment was performed to verify the improvement of quantitative agreement, when we utilized ANNs with composite experts' opinions.**

I. Introduction

Computer-Aided Diagnosis (CAD) system is developed to detect and diagnose the suspicious lesions such as pulmonary nodules, which are often missed or misinterpreted by radiologists. In addition, offering meaningful diagnostic features to strengthen objectivity of diagnostic decisions is also one of the challenges [1]. Computer-Aided Diagnostic characterization (CADc) aims to create an intermediate step between the detection and assessment of suspicious lesions by providing a "semantic" based description of lesions' characters [2]. The pioneer work on predicting radiologists' perception of these characteristics was performed by Nakamura et al. [3], which had developed a scheme by ANN to assist radiologists in the distinction of benign and malignant pulmonary nodules. Several works for content-based image retrieval (CBIR) that retrieve similar cases from a database for a given nodule at hand have also been proposed [4]. And the first commercial system using CBIR has been published by Fujifilm Corporation. The propositions of the coincidence between the computer-based image features and the interpretation formulated by human have been paid more attention, which is known as the semantic gap problem.

Lung Image database consortium (LIDC) serves as a public part for the diagnosis of pulmonary nodules in thoracic CT scans [5]. Beside of the abundantly various thoracic CT data, LIDC also apply the semantic descriptions of suspicious lesions by four radiologists. LIDC database provides plentiful recourses to complete the intelligence of CADc. In other words, predict the human perception from computer-based features.

There are three major challenges mainly consisted to construct such a model. 1) How to delineate the relative accurate regions of interest. To define a boundary region of interest, the radiologist-drawn outlines are combined using the probability map (p-map) [6]. 2) How to predict the degree of characteristics, which are interpreted by semantics, from computer-based features. Initial work towards predicting radiologists' ratings on LIDC characteristics by Raicu et al. applied an extended a set of geometric features [7]. Later on, most studies used a combination of features to shape and geometrical structure of the nodules. 3) How to provide more objective descriptions of diagnostic characteristics according to the radiologist opinions. The prediction of the ratings of LIDC diagnostic characteristics given by individual radiologists has shown limitation. Some researches explore to combine both their ratings to predict a composite rating [8].

The essential role to complete the automation of CADc is to select a robust prediction approach. There are a number of techniques used in the nodule recognition systems: rule-based classifier; template matching; nearest cluster; neural network; Bayesian classifier and so on. For a more detailed literature review on the classification part, we recommend the recent survey by Sluimer et al. [9].

In this study, we attempt to develop a CADc system to implement the prediction of semantic characteristics as well as composite multiple radiologists' opinions. We introduce ANNs as classifier. And examines three tasks: first, predict semantic characteristics from the computer-based image features; second, predict malignant of nodule from its semantic characteristics; third, composite the experts` opinions to improve the reliability of characteristic prediction. The details are illustrated in section 3 and the experiment results will be shown in section 4.

II. LIDC Database

The LIDC has developed a lung nodule collection and reporting protocol. The construction of LIDC is described as two-stage: the first stage consists of radiologists rating nodule characteristics and the second stage applies those ratings towards predicting the radiologists' diagnoses. There are four radiologists to detect suspicious lesions independently. They are demanded to draw the outline around the suspicious lesions whose diameter from 3[mm] to 30[mm] and rate diagnostic characteristics on an ordinal scale of 1-5 (degree of

Guangxu Li, Hyoungseop Kim, Joo Kooi Tan and Seiji Ishikawa are with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan (Tel: (81) 93-884-3185; e-mail: kim@chtl.kyutech.ac.jp).

Yasushi Hirano and Shoji Kido are with Yamaguchi University, Japan.

Rie Tachibana is with Oshima National College of Maritime Technology, Japan.

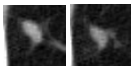
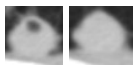
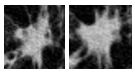
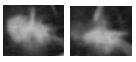
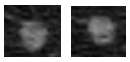
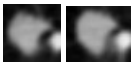
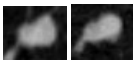
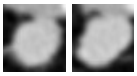
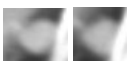
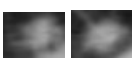
calcification characteristic is up to 6). These 9 diagnostic characteristics include: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, malignancy. Some of them are descripted in Table I. For a detailed description of these characteristics, refer the document [10].

Nakamura et al. concluded in their work [3] that there was poor predictive performance in predicting the radiologists' ratings due to the variability between radiologists. Table II lists 10 samples of the same pulmonary nodules, which were rated by all four radiologists (R1-R4) simultaneously. Table II appears two images of each sample referring to the most distinct characteristics of lesions. And the sequence of annotated characteristics in "ratings" column obeys the order described above.

TABLE I. CHARACTERISTIC DESCRIPTIONS OF NODULE IN LIDC

Characteristic	Descriptions
subtlety	in terms of its difficulty in detection. It refers to the contrast between the lung nodule and its surroundings.
texture	internal texture or composition of nodule in terms of solid and ground glass components.
margin	description of how well defined the margins of the nodule is.
sphericity	the three dimensional shape of nodule in terms of its roundness.
malignancy	subjective assessment of likelihood of malignancy of the nodule.

TABLE II. 10 SAMPLES OF RATING RESULTS OF NODULE CHARACTERISTICS BY 4 RADIOLOGISTS

Images[a]	Ratings	Images	Ratings
	5 1 6 2 4 4 5 5 4		5 1 6 5 4 5 4 5 5
	4 1 6 2 4 5 5 5 2		5 1 6 4 4 3 2 5 3
	2 1 6 3 5 5 5 5 3		4 1 6 5 5 5 5 5 3
	3 1 6 3 3 5 5 4 3		5 4 6 5 3 5 3 4 1
	3 1 6 2 4 4 3 4 4		4 1 6 3 2 3 3 3
	5 1 6 4 5 3 3 5 3		5 1 6 2 1 4 2 5 4
	2 1 6 5 5 5 4 5 2		5 1 6 3 2 4 3 3 5
	3 1 6 3 2 3 3 4 3		5 1 6 3 2 3 4 3 2
	3 1 6 3 2 2 2 2 3		5 1 6 5 4 1 2 5 5
	4 1 6 5 3 1 1 3 3		5 1 6 4 5 2 1 5 4
	3 1 6 5 2 5 5 2 2		5 1 6 5 5 1 2 5 4
	5 1 6 5 5 1 1 2 3		5 1 6 5 5 4 1 5 3
	4 1 6 3 5 1 1 5 3		5 1 6 4 4 3 4 5 5
	5 1 6 4 5 2 1 5 4		2 1 6 5 5 5 5 5 3
	4 1 6 3 4 4 5 5 4		5 1 6 3 5 4 5 5 4
	5 1 6 3 5 1 4 4 4		4 1 6 3 5 2 1 5 3
	2 1 6 3 3 5 5 5 3		4 1 6 3 2 3 5 3 4
	3 1 6 3 4 2 1 4 2		4 1 6 5 3 2 2 5 5
	4 1 6 5 5 4 5 5 4		5 1 6 3 1 3 4 4 5
	2 1 6 4 4 5 4 5 2		4 1 6 5 4 1 1 5 2

a. Two successive slices per sample; sizes of images have been adjusted to fit the table cell

III. METHODOLOGY

A. Textures Prediction

As the description above, characteristics prediction aims to apply a recognition method to support the understanding the diagnostic characteristics from numerical features that included shape features and texture features. Weimker et al. showed the robustness and good correlation between shape index features and radiologists rating [11]. In this study, we ignore the shape characteristics and devote to texture characteristics that are used to distinguish the Ground Glass Opacity (GGO) and the solid nodules. Here the "subtlety" and "texture" characteristics are set as targeted sematic characteristics. We will analyze the pixed-based textures of VOI to predict the ratings of them. We introduce statistical texture analysis computed from Grey Level Co-occurrence Matrix (GLCM) to reflect the texture measures and Zernike moments to assess the symmetrical characteristic.

The GLCM is the 2-D matrix of joint probabilities between pairs of pixels. Four measurements of GLCM are always used for detection of pulmonary nodules.

- Energy measures the homogeneity of an image. A homogeneous scene, like solid, will contain only a few gray levels, giving a GLCM with only a few but relatively high values of image histogram.
- Entropy measures the uncertainty, disorder, or statistical randomness of an image based upon the probability density of the image.
- Correlation is a measure of gray level linear dependence between the pixels at the specified positions relative to each other.
- Inverse Different Moment is also influenced by the homogeneity of the image.

Zernike moments have a number of desirable properties: rotation invariance, robustness to noise, expression efficiency. In this study, we used the normalized Zernike moments to calculate for a grey level image. The introduction about this algorithm and implementation could be found in [12].

B. Malignant Prediction

A number of publications have addressed the problem of automatic rating of the probability of malignancy (pCa). Several attempt to design the systems that explore which characteristics point toward malignancy. Since the malignancy ratings do not represent by single metric but generally depends on the human experiments, in this study, we proposed a scheme of pCa rating works using semantic conceptions. At the same time, the estimation of pCa could be implemented directly without the influence from the selection of image characters, since generally. We introduced 3 semantic characteristics (subtlety, texture and margin) and 2 geometric characteristics (sum of pixels and absolute value of mean pixel) to predict the degree of malignant.

C. Statistical Measures of Observer Agreement

The goodness of a semantic prediction is in absence of ground truth. Some paper aims to predict a composite opinion of all radiologists by combining their ratings, with the assumption that when readers agree they must be correct. Two popular methods always are considered: median and majority voting [13]. In this study, we tested two models to compare the agreement results: 1) the inter-observer agreement and 2) the machine-observer agreement. The former item purposes to

obtain the agreement of annotations among the 4 radiologists. To analyze the inter-reader variability, we have paired all inter-reader parings of ratings (6 pairings for four radiologists). As to the machine-observer agreement, since there was no direct way to infer the accuracy of reader agreement, here we employed the median voting criterion as the reference interpretation. Compare the agreement between the machine ratings and the median voting results.

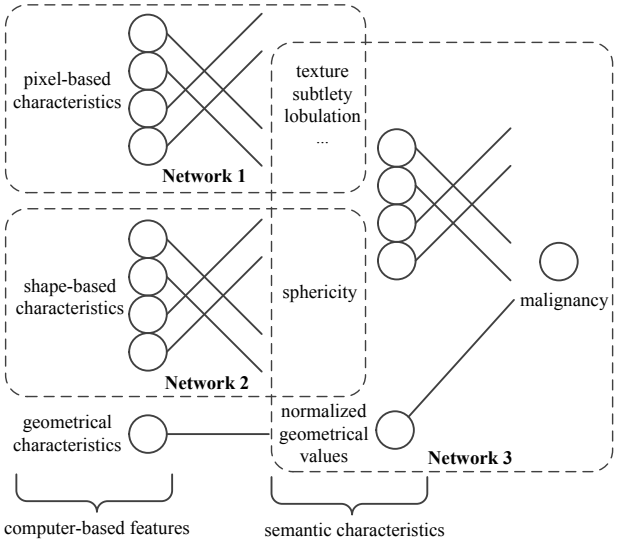


Figure 1. Neural networks structure proposed in this study. "Network 1" predicts the texture characteristics from pixel-based image textures; "Network 2" is used to predict the shape characteristics; "Network 3" estimates the malignancy property according to the semantic characteristics, which generated from human concepts.

IV. Prediction Tools

We employed ANNs to complete the prediction of semantic concepts rate as well as the detection of the pCa. Three-layer, feed-forward ANNs with back-propagation algorithms were used. The basic structure of the ANN includes five input units, three hidden units, and one output unit. The number of hidden units was empirically determined. The outline of the ANNs is demonstrated in Fig. 1.

In light of the description in section of introduction, this study tempts to deal with 3 tasks. The first task was constructing the semantic concepts rate from the pixel-based image characters. Input data are obtained from image measurements of expert delineations of volume of interest (VOI). We choose 5 computer-based textures, which have been introduced in section 3.1. The output of ANN was semantic characteristics that recorded in LIDC database by radiologists. This part is shown as the "Network 1" of Fig. 1. As to the malignancy predication, demonstrated as the "Network 3", input data are set using subjective ratings by radiologists (normalized to range from 0 to 1). As well, the output represents the likelihood of malignancy (5 levels). The third performance is constructed similarly to task one, but composite the ratings of different radiologists as the input data. This experiment tempts to check whether the observer-machine-agreement of annotation would be improved when we train the networks using composite experts' knowledge.

V. Results and Discussion

In regard to the training and testing of ANN, we performed means of a leave-one-out (or round robin) strategy. We analyzed 60 nodules semantic characteristics (subtlety, texture and malignancy) for task one and two. They were divided into 6 groups. 5 groups were used for trained network and 1 for testing. We counted the number of fall-out nodules and averaged results of 6 pairings. The fall-out radios are listed in Table III.

The data used for the third task have been shown in Table II. We tested the agreement of sematic characteristics of "Subtle" as well as "Texture". Trained 9 nodules but left 1 out for testing. The inputs to ANN training were randomly sampled among the ratings of 4 radiologists. To compare the agreement of any 2 observers and the machine-observer agreement, we introduced the kappa value and its standard error [14]. The performance results are listed in Table IV.

From Table III, we can find that the computer-based textures what we chose could reflect the specific characteristics "subtlety" and "texture" well. But the results for predicting the malignancy expressed a large variation. Generally there are not explicit demarcations for discrimination of malignancy of nodule. According to [15], the authors have stated that the most important characteristics to detect the malignancy of nodule appear to be nodule size and growth rate. However from the thoracic CT images it is difficult to provide exact estimation of that.

From Table IV, the kappa of machine-observer yielded value of 0.2857. Although this result appeared eclectic with respect to the independent result of inter-observer, comparing to the average value of inter-observer, we think that the proposed machine-observer method expressed feasibility and objectivity.

TABLE III. Results of Prediction of Semantic Characteristics

Semantic characteristics	Fall-out ratio (%)		
	Mean value	*Minimum value*	*Maximum value*
subtlety	15	10	20
texture	10	0	30
malignancy	28	0	50

TABLE IV. Results of Agreement of Inter-Observer and Machine-Observer

Observer		Observed kappa	Standard error
inter-observer[a]	group1	0.0164	0.1254
	group2	0.3529	0.1536
	group3	0.2632	0.1778
	group4	0.1818	0.1597
	group5	0.0498	0.1134
	group6	0.0619	0.1496
	average	0.1543	0.1466
machine-observer		0.2857	0.1592

a. 6 groups for 4 human observers

VI. Conclusion

As an intermediate step towards detecting and diagnosing of suspicious lesions, the CADc approach, which rates the suspicious lesions clinically, is considered one of the critical links to complete the automation of further diagnosis. In this study, we proposed a CADc approach using ANNs to pursue the objectivity of characteristic rating. The contribution of this work includes: 1) experimented the prediction of some of semantic characteristics from the computer-based features; 2) proposed an automatic approach by compositing the opinions of different experts to improve the objectivity of prediction tool.

In future works, we would like to expand the quantity of data and test a much wider range of numerical features of nodule. As well, exploring the relationship of each feature to its expected semantic characteristic, which is interpreted as the regression problem, is also meaningful. At last, since the ratings do not represent independent classes, the simple kappa statistic appeared limitation. We could like to consider weights kappa as well as other statistics, such as the accuracy and the mean squared error.

References

[1] W. H. Horsthemke, D. S. Raicu, J. D. Furst, "Characterizing Pulmonary Nodule Shape using a Boundary-Region Approach," In *Medical Imaging 2010*, *Computer-Aided Diagnosis*, Proc. of SPIE, vol. 6516.

[2] K. Doi, "Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology," *Phys. Med. Biol.*, vol. 51, 2006, pp. R5–R27.

[3] K. Nakamura et al., "Computerized Analysis of the likelihood of Malignancy in Solitary Pulmonary Nodules with Use of artificial Neural Networks," *Radiology*, vol. 214, 2000, pp. 823–830.

[4] M. Lam et al., "Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images," In *Medical Imaging 2007*, *PASC and Imaging Informatics*, Proc. of SPIE, vol. 6516.

[5] S. G. Armato III et al., "Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community," *Radiology*, vol. 232, 2004, pp. 739–748.

[6] C. R. Meyer, T. D. Johnson et al., "Evaluation of lung MDCT nodule annotation across radiologists and methods," *Acad. Radiol.*, vol. 13, 2006, pp. 1254–1265.

[7] D. S. Raicu et al., "Semantics and image content integration for pulmonary nodule interpretation in thoracic computed tomography," In *Medical Imaging 2007*, *Image Proc.*, Proc. of SPIE, vol. 6512.

[8] W. H. Horsthemke, D. S. Raicu and J. D. Furst, "Predicting LIDC Diagnostic Characteristics by Combining Spatial and Diagnostic Optinions," In *Medical Imaging 2010*, *Computer-Aided Diagnosis*, Proc. of SPIE, vol. 7624.

[9] I. Sluimer, A. Schilham, M. Prokop, B.v. Ginneken, "Computer Analysis of Computed Tomography Scans of the Lung: A Survey," *IEEE Trans. Med. Imag.*, vol. 25, 2006, pp. 385–405.

[10] S. G. Armato III 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.*, vol. 38, 2011, pp. 915–931.

[11] R. Wiemker, M. Bergtholdt, E. Dharaiya, S. Kabus, M. C. Lee, "Agreement of CAD features with expert observer ratings for characterization of pulmonary nodules in CT using the LIDC-IDRI database," In *Medical Imaging 2009*, *Computer-Aided Diagnosis*, Proc. of SPIE, vol. 7260.

[12] "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," http://www.uio.no/studier/emner/matnat/ifi/INF4300/h08/undervisningsmateriale/glcm.pdf

[13] S. G. Armato III et al., "Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence on the Definition of 'Truth'," *Acad radiol*, vol. 16, 2009, pp. 28–38.

[14] J. L. Fleiss, J. Cohen, and B. S. Everitt, "Large sample standard errors of kappa and weighted kappa," *Psychological Bulletin*, vol. 72, 1969, pp. 323–327.

[15] J. J. Erasmus, H. P. McAdams, J. E. Connolly, "Solitary Pulmonary Nodules: Part II. Evaluation of the Indeterminate Nodule," *RadioGraphics*, vol. 20, 2000, pp. 59–66.