35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Overlapping Node Discovery for Improving Classification of Lung Nodules

Fan Zhang, Weidong Cai, Member, IEEE, Yang Song, Student Member, IEEE, Min-Zhao Lee, Shimin Shan, David Dagan Feng, Fellow, IEEE

***Abstract*— Distinguishing malignant lung nodules from benign nodules is an important aspect of lung cancer diagnosis. In this paper, we propose an automatic method to classify lung nodules into four different types, i.e. *well-circumscribed*, *juxta-vascular*, *juxta-pleural* and *pleural-tail*. Additionally, since the morphology of lung nodules forms a continuum between the different types, our proposed method is superior to previous methods that classify single nodules into a single type. First, a weighted similarity network is constructed based on the SVM with probability estimates, turning the 128-length SIFT descriptor to a 4-length probability vector against the four types. Then, the classification of nodules while identifying those with overlapping types is made using the weighed Clique Percolation Method (CPMw). We evaluate the proposed method on low-dose CT images from ELCAP. Our results show that there is more overlap between *well-circumscribed* and *juxta-vascular*, and between *juxta-pleural* and *pleural tail*. Also, quantitative comparisons among various methods demonstrate highly effective nodule classification results by identifying the overlapping nodule types.**

***Index Terms*— Lung nodules, Classification, SVM, CPMw, Overlap.**

I. INTRODUCTION

Lung cancer is a major worldwide problem that seriously endangers people's lives. The survival of patients with lung cancer is strongly dependent on accurate and early diagnosis [1]. Approximately 20% of medical cases with lung nodules represent cancers [2]; therefore distinguishing malignant nodules from benign nodules is essential for the detection of lung cancer.

Recently, image-based diagnosis in clinical settings has focused on the automated computation of quantitative measures to investigate the correlation among different types of lung nodules. According to the medical literature [3], lung nodules that are intra-parenchymal are more likely to be malignant than those attached to vessels or pleura, so classifying lung nodules into categories and aetiology is quite beneficial to pathologists.

To date, all works on classifying lung nodules into the four types tend to annotate an individual nodule with a single label, i.e. each nodule can only belong to one type. However, since the morphology of lung nodules forms a continuum, it may be hard, or even impossible, to clearly distinguish between different types. Although we can recognize some clear differences in shape and texture between different categories [4], there is still significant overlap between these categories. Hence, it is of great value to find these overlapping ones in order to improve the accurateness of the others.

In the light of this, we present a novel method to classify lung nodules and identify those nodules of intermediate or indeterminate type. The structure of the paper is organized as follows: Section 2 discusses the related works; Section 3 illustrates how our proposed method works; then, the experiment procedure and results are discussed in Section 4; finally, the conclusion is given in Section 5.

II. RELATED WORK

Lung nodules are typically spherical in shape; however, they are usually distorted by the surrounding anatomical structures, like vessels or pleural surface [5]. At present, the classification from [6], which divides them into four types, is the most popular criterion for lung nodule classification. The four types are: *well-circumscribed* (W) with the nodule located centrally in the lung without any connection to vasculature; *vascularized* (V) with the nodule located centrally in the lung but closely connected to the neighboring vessels; *juxta-pleural* (J) with a large portion of the nodule connected to the pleural surface; and *pleural-tail* (P) with the nodule near the pleural surface connected by a thin tail. Sample images are shown in Fig.1, with the nodule in the red circle.

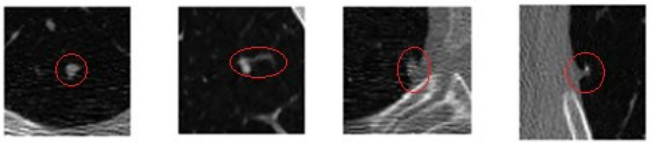


Figure 1. Sample images from the classification [5]. Well-circumscribed, vascularized, juxta-pleural and pleural-tail respectively from left to right.

To categorize individual structural or functional lung nodule images into these four groups, there has been recent interest in machine learning methods. As one of the ten classical methods in data mining, support vector machine (SVM) in particular enjoys great popularity among researchers [5]. SVMs are trained using some well-structured

*This work was supported in part by the ARC grants.

Fan Zhang, Weidong Cai and Yang Song are with the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, Australia.

Min-Zhao Lee is with the BMIT Research Group, School of Information Technologies, University of Sydney, Australia; and Royal Prince Alfred Hospital, Sydney, Australia.

Shimin Shan is with School of Software, Dalian University of Technology, China.

David Dagan Feng is with the BMIT Research Group, School of Information Technologies, University of Sydney, Australia; and the Med-X Research Institute, Shanghai Jiao Tong University, China.

data, e.g. feature vectors extracted by Scale-invariant feature transform (SIFT) algorithm [7]. Then, new images are tested against the model derived from the training process and classified as members of a particular type.

Much work has been done using SVM, demonstrating that SVM can provide more specialized and accurate solutions to our goal [8][9][10][11]. All previous work, however, has used the SVM classification as the end result without further analysis of the overlapping structure. Fortunately, the technique named Clique Percolation Method (CPM) which aims to discover the overlapping structure of the network gives us an insightful way to handle this problem. In particular, it is weighted CPM (CPMw) [12] that we are interested in when dealing with the weighted network. SVM with probability estimates [13] provides the prerequisites for constructing the network for CPMw.

III. METHODS

A. Overlapping structure

The transformation of lung nodules proceeds gradually, which makes classification difficult to clearly define. According to the structure of the above four types introduced, taking well-circumscribed and vascularized nodules as an example, there are always some nodules between 'without any connection to vasculature' and 'with significant connections to the neighboring vessels'. Fig.2 shows an instance between well-circumscribed and vascularized nodules. The nodules located in left and right are easily recognized, but the ones in the middle are difficult to call. The same situation also exists between juxta-pleural and pleural-tail nodules because of the 'large proportion' and 'thin tail'.

Identifying such intermediate nodules is beneficial to accurately classify the clear ones. It is an improvement on other rough classification methods that classify one nodule into a single type, such as the SVM classification that we use in the first stage of our proposed method, and K-means, which we use as the control algorithm in our experiment. The nodules located in the interactions among different types are usually the misclassified ones in these methods. By identifying these nodules, the classification of the others (shown in left and right in Fig.2) can be highly improved.

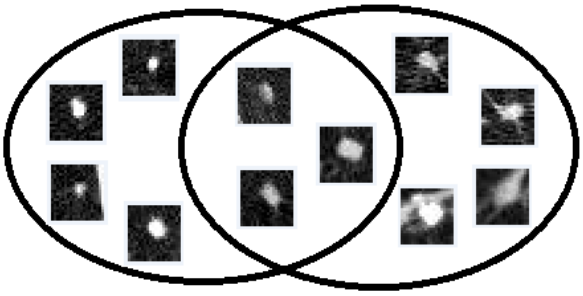


Figure 2. An example of overlapping analysis between well-circumscribed and vascularized nodules. Those located in left part are well-circumscribed, in right part are vascularized, and in the middle are the overlapping nodules.

B. Constructing weighted network with SVM

At the first stage of our proposed method, SIFT descriptors [7] are extracted for all nodules, and the one nearest the centroid is selected as representative of the nodule. After that, the usual SVM classification procedure is performed with the version that can estimate the probability [13] of descriptor *d* belonging to class *i*.

After the prediction step, for any descriptor *d*, we obtain the probability p_i against the *k* classes,

$$p_i = P(y = i \mid d),\ i = 1,..., k\ . \tag{1}$$

Finishing the SVM classification, the original 128-length SIFT descriptor is projected to a 4-length vector. This dimension-reduction process yields a shorter descriptor to represent a nodule. While filtering the noise artifacts, the projection also achieves a more meaningful characteristic to our goal of classification. Unlike other projection methods, such as PCA [14], this approach produces a 4-length vector in a space whose dimensions are nodule types. If two nodules are close together in such a space, they are more likely to be type-similar.

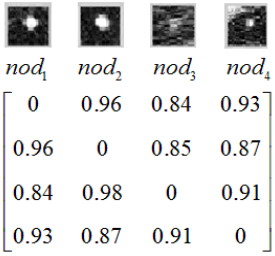
Once the new descriptors for each of the nodules are generated, cosine value as the similarity measure is implemented between two nodules, in order to test whether distinctions are in fact apparent between various nodules. For any nodule *nod*, the type-based vector can be represented as

$$V_{nod} = (p_1, p_2, p_3, p_4) \tag{2}$$

where p_i is defined in Eq.1. The similarity between two different nodules is derived by computing the cosine value, i.e.

$$Sim\,(nod_1, nod_2) = \cos(V_{nod_1}, V_{nod_2}) \tag{3}$$

Finally, the weighted similarity network, in the form of a matrix, can be constructed with the nodes representing the nodules and the weight of each link representing the similarity between the connected nodules. Fig. 3 is a sample showing the construction of a network of four nodules. Notice that we leave the self-similarity equal to 0 in order to fulfill the requirement of CPMw.



	nod_1	nod_2	nod_3	nod_4
nod_1	0	0.96	0.84	0.93
nod_2	0.96	0	0.85	0.87
nod_3	0.84	0.98	0	0.91
nod_4	0.93	0.87	0.91	0

Figure 3. A sample shows the weighted similarity network. The value is the cosine similarity between the two nodules based on the type descriptors.

C. Clustering with CPMw

The second stage of our method is to use CPMw to further cluster the nodules based on the outputs of SVM. CPMw [12] is the extension of CPM [15], which was originally used for unweighed networks.

Given the definition of *k*-clique, i.e. a subset of *k* nodes in which all $k(k-1)/2$ possible pairs are connected, CPMw computes the *k*-clique adjacency. In *k*-clique adjacency, the value indicates whether the two *k*-cliques share $(k-1)$ nodes,

called adjacent *k*-cliques. Traversing the *k*-clique adjacency, the connected *k*-cliques derived are treated as modules, also called clusters. A *k*-clique is included into one cluster if it has intensity larger than the fixed threshold value *I*. The intensity of the *k*-clique, *C*, can be written as follows based on the weight of each link:

$$I(C) = \left(\prod_{\substack{i<j \\ i,j \in C}} Sim(nod_i, nod_j) \right)^{\frac{2}{k(k-1)}} \quad (4)$$

where *Sim* is defined in Eq. 3. By proceeding in such a way, different clusters can share nodes, which are the overlapping nodules in our work, because a single node can belong to several different *k*-cliques.

CPMw has two parameters: *k* and *I*, the clique size and the intensity threshold. The optional choice of *k* and *I* gives the richest structure of weighted module. Suppose the link weight of the network ranges from w_1 to w_n (i.e. $w_1 < w_n$), then a simple suggested method in [10] is to start with the highest value of $I=w_n$, and then decrease *I* until the ratio of the two largest module sizes n_1/n_2 reaches 2. *k* is selected according to the whole structure of the achieved weighted modules. Usually, the more balanced the structure is, the better. For our experiment, we set $I=0.92$ and $k=5$. Except for the above two, CPMw does not need any other predefined parameters, such as the number of the output clusters. This makes CPMw more flexible and robust under different circumstances, especially when the network is not well organized.

Finally, further combining operations are performed in order to obtain the four type-groups. Each cluster is labeled with the type whose frequency is the highest according to the output of SVM classification. Those clusters labeled with the same type are grouped together. In this way, four groups with a more balanced classification structure than the number-fixed methods can be derived.

IV. Experiment and Results

A. Datasets and experiment procedure

In this study, we use the publicly available Early Lung Cancer Action Program (ELCAP) database [16] to illustrate the advantages of the proposed method. The ELCAP database contains 50 sets of low-dose CT lung scans with 379 unduplicated lung nodules annotated by positions, which are further divided into the four specified types (W-15.04%, V-16.09%, J-30.34%, and P-38.52% respectively).

Comparisons are done within the three methods: the proposed CPMw, SVM, and K-means. CPMw is our proposed method. SVM uses the raw SIFT descriptors. K-means is applied in two ways. The first uses K-means to classify the nodules on raw SIFT descriptors, and the other is based the probability estimates.

During preprocessing, a window of 30×30 is extracted with the annotated nodule in the center [5]. Training sets are randomly selected 10 times at a particular percentage (20% to 70%), and the average classification rates of all nodules are computed for each of the percentages.

B. Results

Samples from part of the overlapping nodules are shown in Fig. 4. Red rectangles mark out the overlapping nodules. They are ranked according to frequency of appearing in the intersections for several random training sets. The ones in the center have the highest possibility of uncertainty over which exact type it is. The possibilities of other nodules gradually become smaller to the sides. Totally, most of the overlapping nodules are located between well-circumscribed and vascularized, and between juxta-pleural and pleural-tail. Parts of this classification result are shown in the first and the second row. Also, there are many overlapping nodules between well-circumscribed and pleural-tail shown partially in the third row.

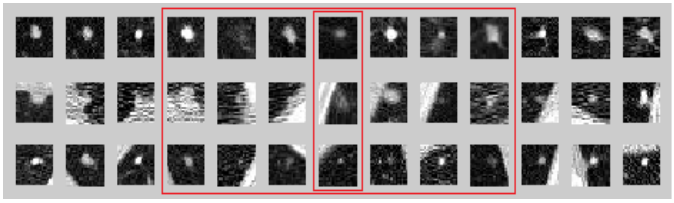


Figure 4. Samples of overlapping nodules extracted from the proposed method. The three rows (from top to bottom) show well-circumscribed and vascularized, juxta-pleural and pleural-tail, and well-circumscribed and pleural-tail respectively. The other three combinations are not shown here because fewer overlapping nodules are found. The ones in the red rectangles have the highest overlapping frequency for all training datasets.

TABLE I. Precision of each type of the nodules from SVM and the proposed method.

Training Percent	Method	Categories			
		Well-C	*Vascular*	*Juxta-P*	*Pleural-T*
20%	SVM	0.722	0.699	0.799	0.740
	CPMw	0.720	0.668	0.811	0.783
30%	SVM	0.777	0.752	0.857	0.768
	CPMw	0.759	0.740	0.798	0.813
40%	SVM	0.796	0.762	0.863	0.836
	CPMw	0.764	0.749	0.866	0.839
50%	SVM	0.850	0.806	0.896	0.858
	CPMw	0.838	0.812	0.873	0.869
60%	SVM	0.888	0.838	0.919	0.898
	CPMw	0.842	0.834	0.898	0.915
70%	SVM	0.921	0.891	0.931	0.935
	CPMw	0.904	0.889	0.943	0.927

Table I shows the precision of the four types using our proposed method and SVM based on SIFT descriptors method. The precisions for well-circumscribed nodules decrease for all training percentages. This change follows the phenomena above that more overlapping nodules are found in the well-circumscribed category due to the fact that they overlap both vascularized and pleural-tail ones. The average precisions of the two methods are shown in Table II. With the better overall recall rates (shown in Fig. 5), the precision of

CPMw is still close to that of SVM although the overlapping nodules are located in more than one type.

Fig.5 demonstrates the comparison of overall classification rates (also seen as recall rates) of nodules among CPMw, SVM based on SIFT descriptors, K-means based on SIFT descriptors, and K-means based on probability estimates. The comparisons can be analyzed from the following aspects:

TABLE II. THE AVERAGE PRECISIONS COMPARING CPMW AND SVM METHODS

Training Percent		20%	30%	40%	50%	60%	70%
method	SVM	0.740	0.788	0.814	0.852	0.886	0.919
	CPMw	0.748	0.783	0.805	0.848	0.872	0.916

(1) K-means based on SIFT descriptors v.s. K-means based on SVM probability estimates. The significant improvement using K-means based on SVM probability estimates proves that the probability vectors upon nodules types are more meaningful than the raw SIFT descriptors when we try to classify the nodules into types. Hence, by further analysis upon SVM output, we can get better classification results.

(2) CPMw v.s. K-means based on probability estimates. Following the above analysis, even though both of these methods are based on SVM probability estimates, CPMw always gives better rates than K-means. This comparison illustrates that CPMw is more suitable for the aim of classification with SVM probability estimates. Also, as the percentage of training set increases, the gap between these two methods gradually becomes smaller due to the fact that more overlapping nodules are found.

(3) CPMw v.s. all others. The higher classification rate of CPMw demonstrates that a better classification can be achieved by identifying the nodules located in between different types. In all, our proposed method does improve the classification performance by further analysis upon SVM output and identifying the overlapping nodules.

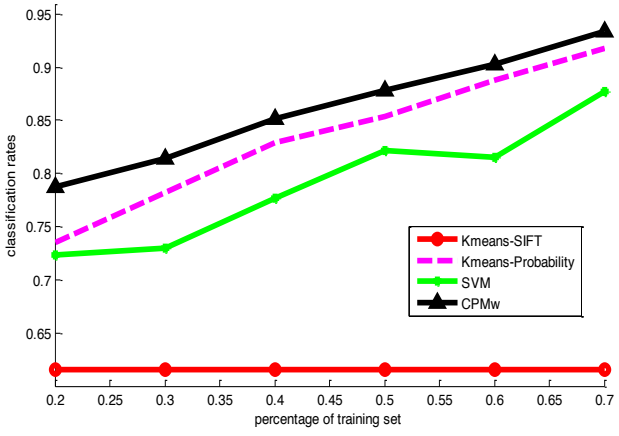


Figure 5. The overall classification rates of the four types of nodules comparing various methods.

V. CONCLUSION AND FUTURE WORKS

This paper presents a new method for lung nodule classification. Probability estimates upon the nodule type are firstly computed using the SVM for each nodule. Then, a nodule similarity network is constructed with the resulting probability vectors. Further overlapping analysis is performed using CPMw. Our evaluation on the ELCAP database shows higher performance than the rough classification.

Future directions are geared towards further utilizing more suitable descriptors for classification against types, instead of the raw SIFT descriptors. Also, we aim to incorporate other similarity techniques to the proposed approach and to obtain the best generalized method.

REFERENCES

[1] B. Zaho, G. Gamsu, M.S. Ginsbe rg, L. Jiang, and L.H. Schwartz, "Automatic Detection of small lung nodules on CT utilizing a local density maximum algorithm," *Journal of Applied Clinical Medical Physics.*, vol. 4, no. 3, pp. 248 - 260, Jun. 2003.

[2] J.J. Erasmus, J.E. Connolly, *et al*, "Solitary pulmonary nodules: part i. Morphologic evaluation for differentiation of benign and malignant lesions," *Radio-Graphics*, vol. 20, pp. 43–58, 2000.

[3] D.M. Xu, H.J. Zaag-Loonen, *et al*, "Smooth of attached solid indeterminate nodules detected at baseline CT screening in the nelson study: cancer risk during 1 year of follow-up," *Radiology*, vol. 250, no. 1, pp. 264–272, 2009.

[4] A. Farag, J. Graham, A. Farag, S. Elshazly, and R. Falk, "Parametric and Non-Parametric Nodule Models: Design and Evaluation". *Proc. of Third International Workshop on Pulmonary Image Processing in conjunction with MICCAI-10*, pp. 151-162, 2010.

[5] Y. Song, W. Cai, Y. Wang, and D.D. Feng, "Location classification of lung nodules with optimized graph construction," *in Proc. ISBI*, pp. 1439-1442, May 2012.

[6] S. Diciotti, G. Picozzi, *et al*, "3-D segmentation algorithm of small lung nodules in spiral CT images," *IEEE Trans. Inf. Technol. Biomed.* ,vol. 12, no. 1, pp. 7–19, 2008.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[8] W. Cai, Y. Song, D.D. Feng, "Regression and classification based distance metric learning for medical image retrieval," in *Proc. ISBI*, pp.1775- 1778, 2012.

[9] A. Farag, S. Elhabian, J. Graham, A. Farag, and R. Falk, "To-wards precise pulmonary nodule descriptors for nodule type classification," in *MICCAI 2010*, LNCS , vol. 6363, pp. 626–633, 2010.

[10] Y. Song, W. Cai, S. Eberl, M. Fulham, D. Feng, "Discriminative Pathological Context Detection in Thoracic Images based on Multi-level Inference", *The 14th International Conference on Medical Image Computing and Computer Assisted Intervention*, Toronto, Canada, 18-22 Sep 2011, LNCS6893, pp191-198, 2011.

[11] Y. Song, W. Cai, J. Kim, D. Feng, "A Multi-Stage Discriminative Model for Tumor and Lymph Node Detection in Thoracic Images", *IEEE Transactions on Medical Imaging*, Vol.31, No.5, pp1061-1075, 2012.

[12] J. Farkas, D. Ábel, G. Palla, T. Vicsek, "Weighted network modules," *New J. Phys.* vol. 9, June 2007.

[13] T.F. Wu, C.J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research,*, vol.5, pp. 975-1005, 2004.

[14] A. Farag, A. Ali, J. Graham, A. Farag, S. Elshazly, and R. Falk,"Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose ct scans of the chest," in *Proc. ISBI*, pp.169-172 , Mar. 2011.

[15] I. Derényi, G. Palla, T. Vicsek, "Clique percolation in random networks," *Phys. Rev. Lett.*, vol. 94, 2005.

[16] "ELCAP public lung image database," url-http://www.via.cornell.edu/databases/lungdb.html.