A Multi-Stage Discriminative Model for Tumor and Lymph Node Detection in Thoracic Images

Yang Song, *Student Member, IEEE,* Weidong Cai, *Member, IEEE,* Jinman Kim, *Member, IEEE,* and David Dagan Feng, *Fellow, IEEE*

***Abstract*—Analysis of primary lung tumors and disease in regional lymph nodes is important for lung cancer staging, and an automated system that can detect both types of abnormalities will be helpful for clinical routine. In this paper we present a new method to automatically detect both tumors and abnormal lymph nodes simultaneously from PET-CT thoracic images. We perform the detection in a multi-stage approach, by first detecting all potential abnormalities, then differentiate between tumors and lymph nodes, and finally refine the detected tumors for false positive reduction. Each stage is designed with a discriminative model based on support vector machines and conditional random fields, exploiting intensity, spatial and contextual features. The method is designed to handle a wide and complex variety of abnormal patterns found in clinical datasets, consisting of different spatial contexts of tumors and abnormal lymph nodes. We evaluated the proposed method thoroughly on clinical datasets, and encouraging results were obtained.**

***Index Terms*—Lung tumor, abnormal lymph node, detection, multi-stage, discriminative, spatial feature**

I. Introduction

LUNG cancer is the most common cause of cancer-related death in men and women, and is responsible for 1.3 million deaths annually, as of 2008 [1]. In particular, non-small cell lung cancer (NSCLC) is the most prevalent type of lung cancer, accounting for about 80% of all cases [2]. Staging, which assesses the degree of spread of the cancer from its original source, is the most important factor affecting the prognosis and potential treatment of lung cancer. For NSCLC, the tumor node metastasis (TNM) staging is the internationally agreed system, which involves analysis of the primary lung tumor, regional lymph nodes and distant metastases [2]. The size and spatial extent of the primary lung tumor and the locations of the abnormal regional lymph nodes indicate a stage IA to IIIB NSCLC, while any distant metastases suggest a stage IV NSCLC.

Positron emission tomography – computed tomography (PET-CT) with ^{18}*F-fluoro-deoxy-glucose* (FDG) tracer is now accepted as the best imaging technique for non-invasive staging [3]. While the CT scan provides anatomical information, it has relatively low soft tissue contrast causing difficulties in separating abnormalities from the surrounding tissues. On the other hand, the PET scan has high contrast and reveals increased metabolism in structures with rapidly growing cancer cells, but their localization is limited by the low spatial resolution in PET images. The integrated PET and CT scan thus provides complementary pathological and anatomical information. In current clinical routine, the localization and characterization of abnormalities need to be performed manually by examining all PET-CT slice pairs. To assist this time-consuming process and potentially provide a second opinion to the reading physicians, an automated system that can provide fast and robust detection is highly desirable.

In this work, our objective is to design a fully automatic methodology for simultaneous detection of primary lung tumors and disease in regional lymph nodes from PET-CT thoracic images. The problem exhibits two main challenges. First, although PET indicates areas with high uptake activities, it can also highlight non-pathological areas (e.g. in myocardium), and the standard uptake value (SUV), which is a semi-quantitative measure of normalized radioactivity concentration, normally exhibits high inter-patient variances. Second, separations between lung tumors and abnormal lymph nodes are difficult. Although they may be differentiated by segmenting the lung fields from CT images, if tumors extent to the surrounding organs especially the mediastinum, such segmentations may not be reliable. For complex cases involving tumors invasion into the mediastinum or lymph nodes abutting the lung field, the ability to differentiate between the two types of abnormalities are more challenging.

In our prior work [4], we proposed a discriminative model with local-, spatial- and object-level features for detecting tumor and abnormal lymph nodes. Whereas good detection performance was observed, several issues should be addressed: (i) the method required the surrounding regions of tumors and abnormal lymph nodes to be accurately classified, which involved a heuristic-based grouping operation to separate the surrounding regions from the mediastinum; (ii) the regions belonging to a tumor or lymph node volume were classified individually, which could result in inconsistent labeling of the set of regions within a 3D volume; and (iii) the high-uptake myocardium introduced false positives in the detection results.

Therefore, we now propose a new method, and the main distinctive characteristics of our method are: (i) a multi-stage detection is designed, where stage-1 detects all ab-

Manuscript received December 16, 2011; revised January 11, 2012; accepted January 13, 2012. *Asterisk indicates corresponding author.*

*Yang Song is with Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, Sydney 2006, Australia (e-mail: yson1723@uni.sydney.edu.au).

Weidong Cai and Jinman Kim are with BMIT Research Group, School of Information Technologies, University of Sydney, Sydney 2006, Australia.

David Dagan Feng is with BMIT Research Group, School of Information Technologies, University of Sydney, Sydney 2006, Australia, and Center for Multimedia Signal Processing (CMSP), Department of Electronic & Information Engineering, Hong Kong Polytechnic University, Hong Kong, and also with Med-X Research Institute, Shanghai Jiao Tong University, 200030 Shanghai, China.

normalities from a 3D image set, stage-2 differentiates the detected abnormalities into tumors and abnormal lymph nodes, and stage-3 reduces the false positives of tumors; (ii) each stage is optimized with a structural discriminative approach, specifically we choose to employ support vector machine (SVM) [5] for classifying individual regions due to its high performance in classification without a generative model, and the conditional random field (CRF) [6] for its capability in exploring the contextual information between multiple regions for a simultaneous classification of all regions in a 3D volume; and (iii) a number of new feature sets are designed, including various types of intensity, spatial and contextual features, for each stage of the detection method.

We test our method on clinical PET-CT image sets, where multiple tumors and abnormal lymph nodes may co-exist and introduce extra complexities. For example, the multiple abnormalities may exhibit large differences in uptake activities; and lymph nodes may reside very closely to the lung tumor and become particularly difficult to separate. The datasets also show a wide variety of abnormal patterns, with tumors of various shapes and spatial extents, and lymph nodes of different sizes and locations.

A. Related Work

Up to now, the amount of publications on simultaneous detection of lung tumors and disease in regional lymph nodes is limited. In our recent work [7], a region-based approach with spatial information was reported, which however, proposed a detection method mainly to facilitate image retrievals, rather than focusing on optimization of the detection performance. Furthermore, the method required a separate class of tumor border, to workaround the issue that the surrounding areas of tumors were often confused with the mediastinum. Such a tumor-border class complicated the training process, which was quite unnatural for the clinical process. Our later work [4] avoided such an issue with a multi-level discriminative model and more comprehensive spatial features, but also posed several improvement opportunities, which we explained in the previous section.

A similar type of work is on lung tumor detection, which first detects all abnormalities, then extracts only those that are highly representative of lung tumors. By first segmenting the lung field, a threshold and fuzzy-logic based approach is then used to detect the lung tumors [8], but the detection performance is quite sensitive to the delineation accuracy of the lung field. Another approach attempts to handle tumors lying close to the edge of lung fields by incorporating the location, intensity, and shape information [9], but the method could potentially result in a large number of false positives with the predefined SUV thresholds. To reduce the false positives detected in the mediastinum, learning-based techniques with tumor-specific features were proposed [10], [11], but the methods were based on empirical studies of SUV distributions and tumor sizes, and did not seem to consider abnormal lymph nodes in the thorax.

Another category of abnormality detection is to detect all instances from PET images, regardless of their types. Such approaches include a texture-based classification method [12], a water-shed based algorithm integrated with morphological measures [13], and a region-based SUV threshold computed based on the object and background ratio [14]. While the former two techniques operate on user-selected volume-of-interest (VOI) or potential lesions, the last one assumes a large portion of the mediastinum to be normal. It was also shown that the detected abnormalities could be used to infer the cancerous status of a patient [15], which however, did not assess the detection performance of tumors or lymph nodes.

There are also a number of existing works on lymph node detection, mostly on CT images. Most of these methods utilize the segmentation of the anatomical structures in mediastinum, such as airways, aorta and pulmonary artery [16], [17], [18], [19]. A Hessian matrix for detecting the blob-like shaped lymph nodes is also used [18], [20]. A deformable registration approach has been recently proposed to restrict the search area of the blob detectors [20], using a probabilistic mediastinal lymph node atlas created by combining all database images with manually delineated lymph nodes. A different discriminative method was also proposed to detect lymph nodes based on comprehensive appearance and spatial features [19]. Detection methods for other types of lymph nodes include the directional difference filtering for abdominal nodes [21] and the marginal space learning for axillary nodes [22]. The detection performances of these approaches are usually highly related to the segmentation accuracy of anatomical structures, which is however, hard to avoid for CT images. These approaches also focus on the lymph nodes only, not considering cases with tumors, especially if they affect the appearances of the anatomical structures in the mediastinum.

Another often studied area for lung tumor and regional lymph nodes is segmentation, including a number of different methods for tumor volume delineation on PET images with a comprehensive review in [23], those for CT [24], [25], [26], and PET-CT images [27], [28], [29], and lymph node segmentations on CT images [30], [31], [32]. While segmentation techniques normally assume a prior knowledge of presence of abnormalities with user annotated initial seeds or bounding boxes, detection algorithms aim to determine such presences and focus on optimizing the detection recall and precision.

B. Outline

The paper is structured as follows. Section II gives an overview of our proposed method. Section III, IV and V describe the three stages of detection – abnormality detection, tumor and lymph node differentiation and tumor region refinement. Section VI introduces the materials and evaluation methods. The experimental results and discusses are presented in section VII and section VIII concludes the paper.

II. Summary of Our Proposed Method

An intuitive idea of detecting tumors and abnormal lymph nodes in a discriminative construct is to assign one most probable label to each voxel. Specifically, let $V = \{v_i : i = 1, ..., N_v\}$ be the 3D image set of a thoracic scan with N_v voxels. Define the set of labels $\{L, M, T, N\}$, representing the

lung field, mediastinum, tumor and abnormal lymph nodes, and a label set $Y = \{y_i : i = 1, ..., N_v\}$ with one label y_i for each voxel $v_i \in V$. Then, the solution is to compute the maximum a posteriori estimates using various types of classifiers:

$$Y^* = \underset{Y}{\operatorname{argmax}}\ P(Y|V) \tag{1}$$

However, it is difficult to design a set of features sufficient for discriminating the four voxel classes in a single classification step. For example, the main distinctive feature between T and N is the spatial context, i.e. being within the lung field or mediastinum; but how to describe this spatial feature without first labeling the regions of L and M is a challenge. Furthermore, in complex cases where the tumors are adjacent to or invading into the mediastinum, the surrounding areas of tumors might be classified as mediastinum, hence causing more difficulties in differentiating T from N.

We thus propose a multi-stage discriminative model to detect tumors and abnormal lymph nodes. First, we made no distinction between the two types, and classified the voxels into the lung field (L), mediastinum (M), and abnormal region of interest (ROI), with two levels of features and SVM-based soft labeling. Next, we designed a CRF model with unary and pairwise features in 3D space to differentiate the detected abnormal volumes into tumors (T) and abnormal lymph nodes (N). Last, we formulated another 3D CRF model to refine the detected tumors for false positive reductions.

III. Stage-1: Abnormality Detection

Each transaxial PET-CT slice of a 3D image set (e.g. Fig. 1a and 1b) was first preprocessed to remove the background and soft tissue areas outside of the lung and mediastinum automatically, with Otsu thresholding and connected component analysis. The preprocessing was based on a simplified lung field estimation method [33] without excluding the mediastinum (Fig. 1c and 1d). Some unnecessary surrounding tissues could remain in the resultant images, but it would not affect the detection.

The preprocessed images were then clustered into regions using quick-shift clustering [34], which was chosen for its edge preserving capability and computational efficiency. The clustering was performed separately for PET and CT slices, and the regions formed from co-registered PET and CT slices were fused into one set of regions (Fig. 1e). To describe the fusing step briefly: consider a voxel v_i, belonging to region r_a in PET and r_b in CT; if such $\langle r_a, r_b \rangle$ combination was newly encountered, a new region was then created in the fused region set; otherwise, v_i would be assigned to the existing region representing this combination. Each 3D image set was then represented by N_r regions from all slices $V = \{r_i : i = 1, ..., N_r\}$. Note that this clustering step would produce a large number of regions, separating the small variations in intensities with region contours delineating closely the actual image edges. The resulted regions, therefore, would be highly homogeneous to facilitate accurate region detections.

Each region r_i was then classified into L, M or ROI categories in a two-step process based on low-level and high-level features. The ROI regions were then the detected abnormalities, and a soft label was also produced for r_i indicating its probability of belonging to each category, The example outputs are shown in Fig. 1f and 1g.

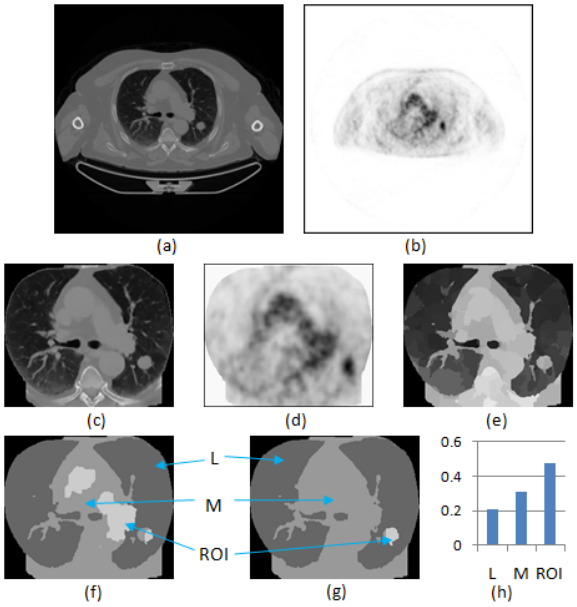


Fig. 1. Illustration of stage-1 on abnormality detection. (a) A transaxial CT slice. (b) The co-registered PET slice. (c) The CT slice after preprocessing. (d) The PET slice after preprocessing. (e) The regions created using quick-shift clustering. (f) The detection output based on low-level features, with blue arrows indicating the correspondence between grayscale values and region types. (g) The second detection output based on high-level features. (h) The soft labeling vector of the detected ROI region.

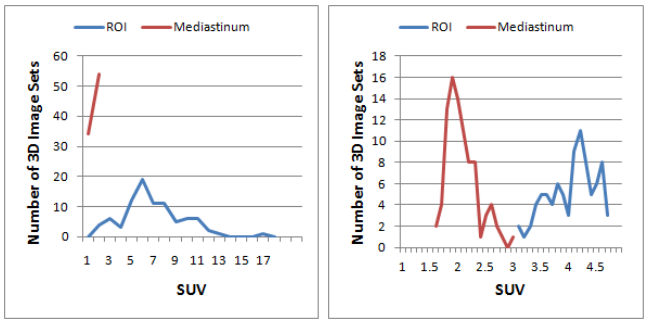
A. Low-level Features

In the first step, intensity and neighborhood features were extracted for r_i, and a three-class SVM (the LIBSVM package [35]) was trained for labeling $y_i = \{L, M, ROI\}$.

1) Intensity: A two-dimensional intensity vector was computed: (i) average CT density and (ii) average normalized SUV of r_i. While the average CT density is based on the raw values, for PET, we performed an extra SUV normalization $\|u_i\|$ with a sigmoid function (besides the standard SUV computation based on the injected dose and patient weight):

$$\|u_i\| = \frac{C_1}{1 + \exp(-(u_i - \theta_V)/\theta_V)} \tag{2}$$

where u_i was the average SUV of r_i, θ_V was the adaptive reference value computed for each 3D image set V, and C_1 was a scaling constant controlling the range of the normalized SUV. The motivation of this normalization step was that, we observed large differences in SUVs between patients with lung tumors, as shown in Fig. 2a, which would lead to high intra-class (e.g. between ROIs) variations hindering effective classification. We thus rescaled the SUVs across patients within a similar range, and in the process, boost the separations between the ROIs and the mediastinum (including other structures), as shown in Fig. 2b. The actual value of C_1 was not important, and was set to 5 only so that its half (i.e. 2.5) would match the most often used SUV threshold, for an easier visualization.



(a) (b)

Fig. 2. Empirical distribution of the average SUVs of the annotated ROIs and mediastinum areas, with one key sample taken from each 3D image set, based on (a) the original SUV; and (b) the normalized SUV.

The reference value θ_V was derived by:

$$\theta_V = \max_s\{C_2 \times \max_i(u_i^s) + l_s\} \tag{3}$$

where s denoted the slice number in the 3D image set V, and l_s was the average SUV of the lung field, and C_2 was a constant. With the $\max_s$ iteration, θ_V was derived based on the 3D image set and served as a global context, so that all regions in V could be normalized with the same reference value. The lung field was estimated by Otsu thresholding the CT image into foreground (the mediastinum and ROIs) and background (the lung field), and the estimated background mask was then mapped to the PET image to compute its average SUV.

The constant C_2 was obtained using a learning-based approach with the following steps: (i) C_2 was initialized to a certain value; (ii) the low-level features were computed for the training set; (iii) region labelings based on the low-level features were conducted on the training set; and (iv) the total number of mislabelings was computed. These four steps were repeated for $C_2 \in [0.1, 0.2]$, and the value producing the smallest error counts was selected. The range $[0.1, 0.2]$ was chosen as motivated by our previous study [14], in which a fixed value 0.15 was used.

Such computation (3) was similar to our previously proposed SUV threshold [14], which instead of using l_s, was based on mediastinum approximation, which we found not robust enough if the lung tumor invaded a large portion of the mediastinum. Furthermore, because of the use of l_s, θ_V was no longer a threshold and ROIs should exhibit a considerably higher SUV than θ_V. We thus incorporated a learning-based approach (SVM) for ROI detection.

2) Neighborhood: The average CT density and normalized SUV of the neighboring area of region r_i in the adjacent slices (one above and one below) were also computed, to incorporate 3D information.

B. High-level Features

The low-level features could achieve high discriminative power if all three classes exhibited clear separations in their ranges of CT densities and normalized SUVs. In some cases, however, the SUVs of ROIs and the mediastinum were relatively close, and some false positive ROIs could be detected in the mediastinum, e.g. the extra ROI detected in the mediastinum as shown in Fig. 1f. Nevertheless, such false positive areas still had lower SUVs comparing to the real ROIs. We thus exploited the high-level features, by computing the contrast between the detected ROIs and the lung field and mediastinum, based on the labeling results obtained with the low-level features.

To do this, we first formed the lung field, mediastinum and ROI – $\{R_L, R_M, R_O\}$, by grouping 3D-connected regions that were classified as the same category into large 3D volumes. Note that multiple isolated volumes of category M might exist in one image, and the largest one approximated the mediastinum R_M. If there were multiple ROIs detected in an image, R_O represented the collection of all ROIs. Let u_L, u_M and u_O be the average normalized SUVs of R_L, R_M and R_O, a four-dimensional high-level feature was then computed for each ROI region r_i:

$$\{u_r/u_L, u_r/u_M, u_r/u_O, u_r\} \tag{4}$$

where u_r was the average normalized SUV of r_i. Note that u_L, u_M and u_O represented the average SUVs of the lung fields, mediastinum and abnormalities of a 3D image set, not for individual slices, so that the ratios, especially u_r/u_O, would not be sensitive to the local features to cause inconsistent representations for regions at different slices. A binary SVM was then used to classify r_i as M or ROI, and ROI regions then represented the detected abnormalities.

C. Soft Labeling

Since any misclassification at this stage would be propagated to later stages, to reduce the impact of possible misclassifications, rather than labeling every region with a single category (L, M or ROI), we used soft labeling to create a vector of probabilities that the given region belonged to each of the three categories (e.g. Fig. 1h). Such soft labeling allowed more discriminative information within the same category, and a fuzzy distinction between different categories.

The soft labeling vector, denoted as $p_i = \{p_i^L, p_i^M, p_i^O\}$ for region r_i, was first obtained from the probability estimates of the SVM classification based on low-level features, indicating the probability of r_i belonging to L, M or ROI. Then, based on the probability estimates of the SVM classification with the high-level features, denoted as $p_i^{M,H}$ and $p_i^{O,H}$, the soft labeling vector was recomputed as:

$$p_i = \{p_i^L, (1-p_i^L)p_i^{M,H}, (1-p_i^L)p_i^{O,H}\} \tag{5}$$

As described in later sections, the soft labeling vectors were used for feature computations for further differentiations between tumors and abnormal lymph nodes.

D. Design Motivation

In this section, we summarize our motivations of the method design for stage-1. First, while voxels were normally the processing units, the amount of information describable by a voxel was limited; and voxel-based processing would identify small and unnecessary details (e.g. vessels). Therefore, we chose to use the region-based approach to incorporate more information and produce more spatially smoothing labeling.

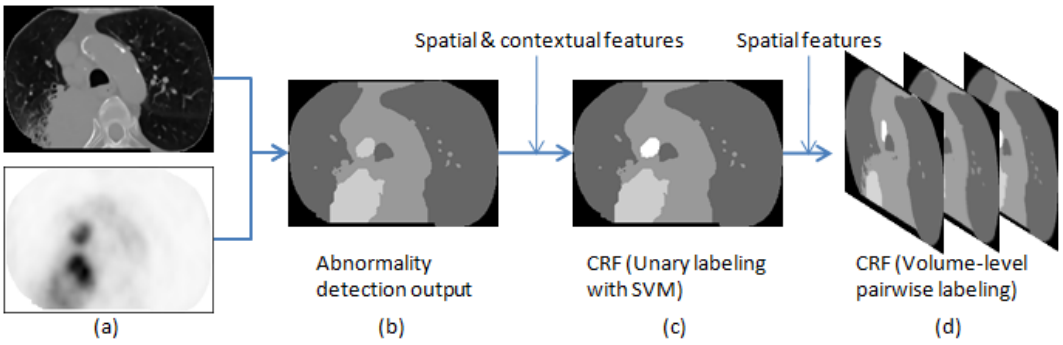


Fig. 3. Illustration of stage-2 on tumor and lymph node differentiation. (a) A transaxial PET-CT slice. (b) The abnormality detection output – two ROIs detected. (c) Region-level labeling with the unary term based on spatial and contextual features – the two ROIs labeled as one tumor (gray) and one abnormal lymph node (white). (d) Volume-level labeling with the pairwise term based on 3D spatial features.

Second, although the regions were in 2D space, the features computed for each region actually incorporated the global context pertaining to the 3D volume, such as the normalized SUV and the high-level feature set. Such global information was particularly useful to accommodate inter-subject variations and cases with low contrasts between the mediastinum and ROI. Therefore, we could classify effectively at the region level whether a region represented an abnormality based on its feature set.

IV. Stage-2: Tumor and Lymph Node Differentiation

At the second stage of our method, the detected abnormalities were differentiated as tumors or abnormal lymph nodes, as illustrated in Fig. 3. A CRF model integrated with SVM and a comprehensive set of features were designed to achieve an accurate discrimination between the two types of abnormalities (SVM), and minimize any misclassification by exploiting 3D correlations (CRF). The use of CRF allowed us to incorporate the structural information in addition to the region-based features, so that a 3D ROI volume could be classified collectively.

A. CRF Formulation

Based on the outputs of stage-1, the abnormalities detected from a 3D image set V were represented as N_O regions $\{r_i : i = 1, ..., N_O\}$. Rather than individual regions created at the clustering step, r_i here represented a large ROI region created by merging regions that were labeled as ROI and spatially connected in the same slice. A set of 3D connected $\{r_i\}$ (i.e. across slices) then formed a 3D ROI volume, and V could contain multiple ROI volumes, e.g. a primary lung tumor and several abnormal lymph nodes.

The objective was then to assign each r_i a binary label $a_i \in \{T, N\}$; and the probability of a labeling set $A = \{a_i : i = 1, ..., N_O\}$ was modeled as a conditional distribution in the CRF framework:

$$P(A|V) = Z^{-1} \exp(-E(A|V)) \quad (6)$$

where Z was the partition function.

We defined the energy $E(A|V)$ as a linear combination of a set of unary features $F_k(a_i, V)$ and a pairwise feature $G(a_i, a_{i'}, V)$:

$$E(A|V) = \sum_i \sum_k \lambda_k F_k(a_i, V) + \sum_{i,i'} \mu G(a_i, a_{i'}, V) \quad (7)$$

where λ_k was the weight of the kth feature, i and i' indexed the 3D connected regions (in different slices), and μ was the weight of the pairwise feature. The unary features were computed for each r_i and were the most decisive factor for labeling $a_i \in \{T, N\}$, while the pairwise features were to exploit the 3D structural information for a consistent labeling throughout an ROI volume.

As often suggested [36], to reduce the computational complexity for training of the parameters, we employed a piecewise approach to obtain λ_k and μ separately. The model parameter λ_k was learned in the unary term, and μ was assigned constant 1 for equal weights between the unary and pairwise terms. We will describe our design details of the unary and pairwise features, and the model parameters in the following sections.

Basically, each ROI region r_i could be labeled as either T or N category, with each possible labeling associated with an energy cost $\sum_k \lambda_k F_k(a_i, V)$; and lower costs imply higher probabilities. However, regions should not be treated totally independently from other regions that were spatially connected, since it would be highly likely that connected regions belonging to a 3D ROI volume would be assigned the same label. Therefore, the labeling difference between pair of connected regions r_i and $r_{i'}$ also contributed to a cost value $\mu G(a_i, a_{i'}, V)$; and higher costs implied higher penalties for inconsistent labelings of region pairs.

The idea was then to search for a labeling combination for V, so that the total energy cost was minimum, leading to a labeling set that would be optimized for the whole 3D volume. Graph cut [37], [38], [39] was used to derive the most probable labeling A^* that minimized the energy function:

$$A^* = \underset{A}{\operatorname{argmin}} \; E(A|V) \quad (8)$$

B. Unary Term

The unary term $\sum_k \lambda_k F_k(a_i, V)$ (denoted as $\psi(a_i)$) indicated the labeling preference of individual region r_i. Specifically, given label a_i for region r_i, a higher $\psi(a_i)$ meant a higher cost (i.e. lower probability) of r_i belonging to a_i. The unary term could thus be considered as a binary classifier for r_i. We designed a highly discriminative feature set describing the spatial and contextual features, and derived the unary cost $\psi(a_i)$ from the classifier output (SVM).

1) Spatial and Contextual Features: The main distinctive feature between T and N is the location information. In particular, it is generally true that T is in the lung field while N is in the mediastinum. One could then attempt to distinguish T and N by analyzing the spatial locations of ROIs relative to the detected lung fields and mediastinum.

However, quite often T might invade into the mediastinum and appear to be outside of the lung field. Furthermore, T's boundary areas (i.e. areas surrounding T with similar CT density but lower SUV than T) were usually labeled as M; for T that was near to the mediastinum, such labeling would cause merging of the pleural with mediastinum, rendering T to appear outside of the lung field. In addition, N could be adjacent to the lung field, exhibiting similar features to certain types of T. These factors implied that more comprehensive features were necessary for differentiating T and N effectively.

From our analysis and experiments on the dataset, we designed a set of spatial and contextual features, which were extracted for each region r_i that was detected as ROI, as described in the following.

(i) The quad-radial global histogram: four radial lines were drawn at $\pm 45°$ and $\pm 135°$, from the geometric center of r_i, extending to the edges of the image slice. The radials represented the areas to the medial, anterior, posterior and lateral portions of the thorax, relative to r_i (Fig. 4). The spatial distribution of different voxel categories among the four radials was expected to be different for T and N. A 12-dimensional histogram H_g was thus created to compute the distribution of L, M and ROI in the four radials:

$$H_g = \{H_{g,i} : i = 1, ..., 4\} \tag{9}$$

$$H_{g,i} = \frac{1}{S} \sum_v p_v,\ p_v = \{p_v^L, p_v^M, p_v^O\} \tag{10}$$

where i indexed the radials, and v indexed the voxels in each of the four radials, S was the size of r_i, and p_v was the soft labeling vector of v, which was propagated from the region-level soft labeling vector (5). The normalization of histogram values with the size of r_i effectively magnified the differences in H_g between a tumor and lymph nodes, which was especially useful for differentiating a tumor that invaded into the mediastinum and the lymph nodes.

(ii) The surrounding contour histogram: a closed contour was drawn outside of r_i, following the same shape as r_i, with a displacement of d from the boundary of r_i. The belt area that was enclosed by the boundary of r_i and the closed contour was then the surrounding contour of r_i (Fig. 5). The width of the belt area d was chosen as half of the minor axis length of r_i. A three-dimensional histogram H_s was then created to

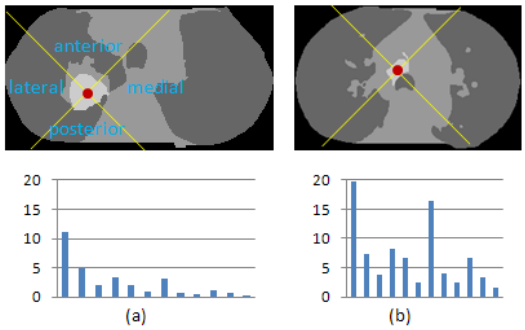


Fig. 4. Illustration of the quad-radial global histograms on (a) a tumor, and (b) an abnormal lymph node. Upper row: the radial structures drawn on stage-1 outputs. Lower row: the histogram representations. The radials are divided with yellow lines originated from the region centroid (red dot).

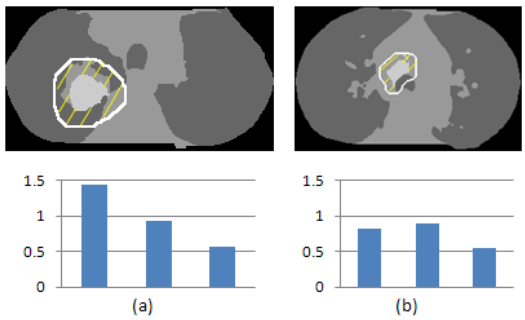


Fig. 5. Illustration of the surrounding contour histograms on (a) a tumor, and (b) an abnormal lymph node. Upper row: the contour structures drawn on stage-1 outputs. Lower row: the histogram representations. The white contour indicates the border of the surrounding contour, shaded with yellow lines.

count the percentages of L, M and ROI in the surrounding contour:

$$H_s = \frac{1}{S} \sum_v p_v,\ p_v = \{p_v^L, p_v^M, p_v^O\} \tag{11}$$

where v indexed the voxels in the surrounding contour. The use of soft labeling vector instead of the most probable label helped to reduce the problem with the tumor boundary areas that were misclassified as M, since now these voxels would also contribute to the histogram bin representing ROI. The choice of the contour width d was also to make sure the contour was larger than the ambiguously labeled boundary area and covered some lung field surrounding the tumor, while not too large to extend to further areas and diminish the discriminative power of the feature.

(iii) Pleural distances: the distances between r_i and the lateral (LL), medial (LM), anterior (LA), and posterior (LP) sides of the nearest lung field were computed (Fig. 6):

$$D_1 = \frac{1}{Y} \sum_{y=1}^{Y} d_{y,LL},\ D_2 = \frac{1}{Y} \sum_{y=1}^{Y} d_{y,LM} \tag{12}$$

$$D_3 = \frac{1}{X} \sum_{x=1}^{X} d_{x,LA},\ D_4 = \frac{1}{X} \sum_{x=1}^{X} d_{x,LP} \tag{13}$$

where Y and X was the height and width of r_i, and d denoted the signed distances between the yth row or xth column of

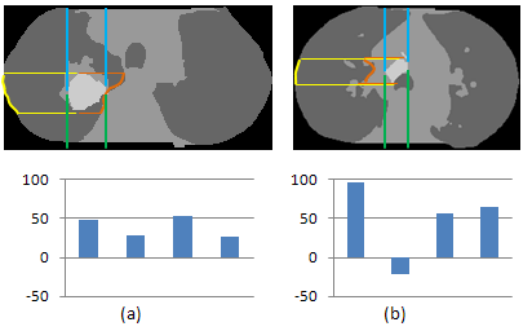


Fig. 6. Illustration of the pleural distances on (a) a tumor, and (b) an abnormal lymph node. Upper row: the feature structures drawn on stage-1 outputs. Lower row: the histogram representations. Yellow and orange lines depict the region for computing D_1 and D_2, and blue and green lines for D_3 and D_4.

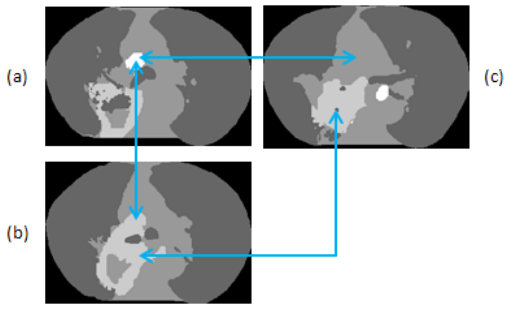


Fig. 7. Illustration of the spatial overlapping between regions based on three slices of the same 3D image set. (a) A slice showing T and N. (b) A slice with connected T and N. (c) A slice without the N region as shown in (a). In this example, the N region in (a) would be connected to the T region in (c) via the T region in (b). With the overlapping factor, (a) and (c) can be then uncorrelated.

r_i and the sides of the lung field, normalized by the slice length or width. The lung field was extracted based on the labeling at stage-1, which excluded the tumor areas and was thus smaller than its actual size. Although the sign of D_2 normally indicated if r_i was in the lung field, a tumor that invaded into the mediastinum would exhibit a similar D_2 to the lymph nodes, yet the other distances would help for their differentiations.

2) Unary Cost: The unary cost $\psi(a_i)$ indicated the cost of labeling r_i with $a_i = \{T, N\}$, and a higher $\psi(a_i)$ implied a lower probability of r_i belonging to a_i. We used a binary SVM to classify r_i to either T or N category based on its feature vector $\{H_g, H_s, D\}$, with a probability estimate p_{a_i} for each category. The unary cost was then computed as:

$$\psi(a_i) = 1 - p_{a_i} \tag{14}$$

to produce two cost values for each r_i. Note that we did not explicitly compute the model parameter λ_k, which essentially resembled the feature weights in the SVM classifier, but used the probability estimates from SVM directly for the cost value.

Furthermore, we observed that the regions nearer to the boundary of the ROI volume were more prone to mislabeling; and by reducing their contributions to the total unary energy, the graph-cut solution (8) for the labeling A^* would be less affected by such regions and thus more optimized. Therefore, the unary cost was refined with a Gaussian weight ω_i based on the distance between r_i and the volume center (i.e. center in z direction):

$$\psi(a_i) = \omega_i(1 - p_{a_i}) \tag{15}$$

$$\omega_i = \exp(-\frac{(z_i - z_c)^2}{2\sigma^2}) \tag{16}$$

where z_i and z_c were the z coordinates of r_i and the center of the volume, and σ was calculated as $1/2$ of the size of the volume (in z direction). The ROI volume was created using 3D connected component analysis.

C. Pairwise Term

The pairwise term $\mu G(a_i, a_{i'}, V)$ (denoted as $\phi(a_i, a_{i'})$) was useful in promoting spatial consistency between spatially connected regions r_i and $r_{i'}$. Specifically, a cost was assigned as $\phi(a_i, a_{i'})$ if r_i and $r_{i'}$ were labeled differently. The regions r_i and $r_{i'}$ were considered spatially connected if they were part of the same 3D ROI volume. Recall that r_i (and $r_{i'}$) represented a large ROI region, which was formed by grouping the connected ROI-category regions in the same slice. The pairwise term thus explored the inter-slice and volume-level information for refined labeling.

The model parameter μ was set to 1 to have equal weights between the unary and pairwise terms, and the pairwise cost was defined as:

$$\phi(a_i, a_{i'}) = \delta(a_i - a_{i'}) \cdot x' \tag{17}$$

$$x' = \frac{1}{1 + \exp(-C(x - 0.5))} \tag{18}$$

where $\delta(a_i - a_{i'})$ was 0 or 1 indicating the same or different labelings of r_i and $r_{i'}$, and x' was the cost value, which was a sigmoid normalization of the actual cost x:

$$x = \alpha(r_i, r_{i'}) \cdot \beta(r_i, r_{i'}) \tag{19}$$

where $\alpha(r_i, r_{i'})$ and $\beta(r_i, r_{i'})$ were the distance and overlapping factor, as described below. Since $x \in [0, 1]$, the sigmoid normalization was useful for magnifying the differences between higher and lower costs, and the constant C was chosen as 20 empirically for a steep curve. Larger or smaller C was also acceptable; but if C was too large, the region labelings would be overly smoothed, and if C was too small, many inconsistent labelings across the 3D volume would be evident.

The factor $\alpha(r_i, r_{i'})$ measured the spatial distances between the two regions:

$$\alpha(r_i, r_{i'}) = \exp(-\frac{\|r_i - r_{i'}\|^2}{2\langle\|r_i - r_{i'}\|^2\rangle}) \tag{20}$$

where $\|r_i - r_{i'}\|^2$ was the L2-norm of the 3D spatial distance between the geometric centroids of r_i and $r_{i'}$, and $\langle\cdot\rangle$ indicated the average spatial distance between all pairs of regions in the 3D ROI volume. The value of $\alpha(r_i, r_{i'})$ was in $[0, 1]$ range, and was larger if the distance between r_i and $r_{i'}$ was smaller. Such a computation was similar to the usual CRF formulation, but was based on the spatial distances, rather than intensity differences; and the pairwise cost was computed between

all pairs of regions of a 3D ROI volume, rather than only for those neighboring regions. Such a design was motivated by our objective to assign a single labeling to the 3D ROI volume; and hence all regions in the volume were considered correlated, with higher correlations between regions that were nearer spatially.

The factor $\beta(r_i, r_{i'})$ was computed as the degree of overlap in the xy plane between the two regions in different slices:

$$\beta(r_i, r_{i'}) = \frac{1}{\tau} \log(\frac{\cap(r_i, r_{i'})}{s_i + s_{i'}} + 1) \tag{21}$$

where $\cap(r_i, r_{i'})$ was the size of the overlapping area, s_i and $s_{i'}$ were the sizes of r_i and $r_{i'}$, and τ was the maximum value of $\log(\cdot)$ in the 3D volume for normalization. The value of $\beta(r_i, r_{i'})$ was also in $[0, 1]$ range, and a larger overlap would lead to a larger $\beta(r_i, r_{i'})$. This factor was introduced because in some cases, adjacent T and N volumes could actually form into one 3D volume. This usually happened when T and N regions were connected in some slices, so the two volumes then became joint at these places, as shown in Fig. 7. With the $\alpha(r_i, r_{i'})$ factor alone, the regions in T and N would be all correlated, and the lowest energy solution would tend to produce a single label for the joint volume. Therefore, the $\beta(r_i, r_{i'})$ factor was incorporated to adjust the correlations. The regions that were further from the joint slices usually exhibited smaller overlap in the xy plane, hence producing smaller $\beta(r_i, r_{i'})$. They then became much less correlated, and more probable to obtain different labelings correctly.

D. Design Motivation

In comparison with volume-level processing, our design had three main merits. First, it would be difficult to create a representative and discriminative feature set for the 3D ROI volume. The characteristics across a 3D volume were usually highly complex, and choosing a suitable scale of feature descriptions would then be an issue. Lots of feature details would lead to very high feature dimensions, with large intra-class variations preventing effective classifications. Reducing feature details would help to manage the complexity, but it might then affect the discriminative power of features. Breaking a 3D ROI volume into a set of regions would thus reduce such difficulties and the slice-by-slice subdivision was a natural choice.

Second, labeling based on a set of regions was more effective in error tolerance. More specifically, in our formulation, even if some regions were misclassified, the volume-level labeling could still be correct, with the weighted unary costs, pairwise terms and total energy optimization. However, if the 3D ROI volume was treated as a single entity, only a single label would be assigned without an opportunity of refinement. Third, as described in the previous section, abnormalities of different types could be connected into a single 3D volume; and labeling the entire volume would thus cause problems. However, with region-level classifications and volume-level constraints, it became more likely to obtain both consistent and more accurate labeling across a 3D ROI volume.

We also experimented with adding more inter-slice features for the unary term, particularly for the quad-radial global histogram and surrounding contour histogram. However, lower performance was observed, especially for tumors adjacent to the mediastinum and abnormal lymph nodes close to the lung field. This suggested that features from adjacent slices would actually introduce more contradicting information.

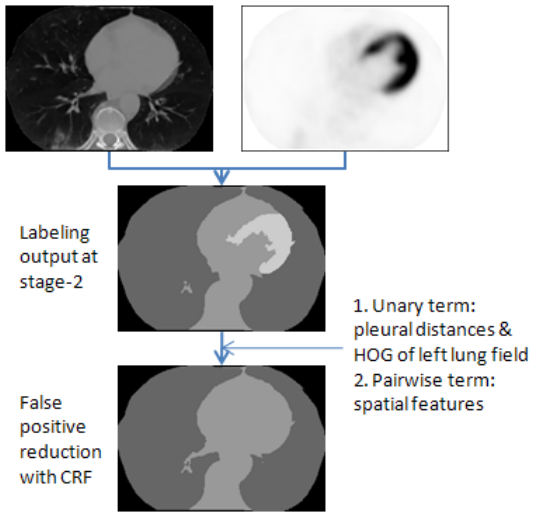


Fig. 8. Illustration of stage-3 on tumor region refinement.

V. Stage-3: Tumor Region Refinement

Sometimes patients could exhibit high uptake activities in the myocardium, which often leaded to tumors detected at the myocardium areas. The usual assumption was that images showing high SUVs in the myocardium should be considered normal, and the detected tumors in such areas should be ignored [15]. We thus designed a method to identify such false positive tumor volumes and update their labelings to M, as shown in Fig. 8.

A. Problem Formulation

Given a detected tumor volume T_q, we first checked that if it was at the left half of the thorax; if not, T_q could not be at the myocardium. For those passing the check, a CRF model was employed to classify T_q to either M or T category, depending on the likelihood of T_q representing a high-uptake myocardium or a lung tumor. Defining T_q as a series of regions $\{r_i : i = 1, ..., N_O\}$, with each r_i representing a set of connected T regions in a slice, the CRF model was designed based on the same construct as (6) and (7) and the same pairwise term as (17), but with a different set of features for the unary term.

B. Unary Term

At the unary level, a T-type region r_i was labeled as M or T. The main design concern at this step was that any real tumor region should not be misclassified as M, even if at the expense of possibly leaving some false positive detections (i.e. unfiltered myocardium regions). Therefore, we designed a rather rigid feature set so that only r_i that was highly probable of depicting the myocardium would be classified as M, as described in the following.

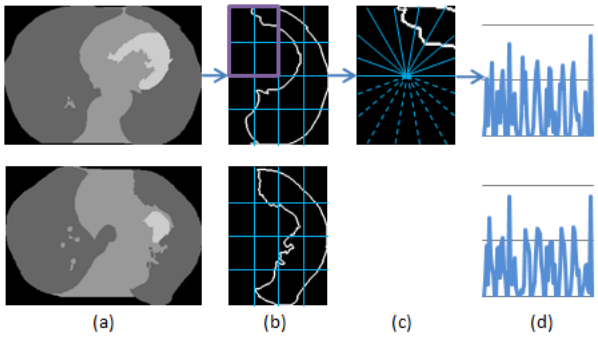


Fig. 9. Illustration of the HOG descriptor for describing the shape of the left lung field. Upper row: a myocardium case. Lower row: a real tumor case. (a) The detection output at stage-2. (b) The edge of left lung field, with blue lines showing the subdivision for creating 3×3 half-overlapping cells, and the purple box indicating the first cell. (c) The structure of the 9 unsigned radials on the first cell. (d) The complete histogram representation.

1) Pleural Distances: The signed distances between r_i and the four sides of the left lung field was computed, in the same way as (12) and (13). The distances represented the location of r_i, and r_i would only be M if it was outside and near to the anterior side of the lung field. The region should also be relatively further from the lateral side, to differentiate from tumors that were adjacent to the mediastinum.

2) Shape of Lung Field: A HOG descriptor [40] was used to describe the shape of the left lung field that r_i was adjacent to, as illustrated in Fig. 9. The image slice was first filtered in x and y directions to generate the edge of the left lung field, and the gradient orientations along the edge were computed. The bounding box of the left lung field was then divided into 3×3 half-overlapping cells; and for each cell, a normalized histogram of 9 unsigned radial channels binning the edge orientations was constructed. The combined feature was thus 81-dimensional. The parameters (3 and 9) were chosen based on the original design of the HOG descriptor, and were empirically tested to achieve a good discriminative power with a low feature dimension. This shape feature was assigned to r_i (although the feature actually described the lung field); and it was the main representative feature for r_i being the myocardium, since the shape of the left lung field adjacent to the myocardium was normally quite different from the other parts of the lung. The use of histogram allowed some spatial variations of the shape of the lung field, and the rigid subdivision of cells and channels helped to identify differences in shapes of lung fields. Note that the shape of r_i was not used, since normally only a varying portion of the myocardium would display high SUVs, and r_i would then exhibit a large variety of shapes.

A binary SVM was then used to classify r_i to either M or T, and the unary cost was computed from the probability estimates of the classifier using (15).

C. Design Motivation

The initial check of whether the detected tumors were in the left thorax was important, so that the feature design could concentrate on differentiating myocardium from tumors at similar locations. If the labeling was performed for all detected tumors, a more complicated feature set would be required, and simply adding the spatial location would not work due to the variations of tumor locations. This was also the reason why stage-2 and stage-3 were designed as sequential stages, rather than combining them into a three-class labeling approach.

Furthermore, a combined approach would require a single training set; but with a small number of myocardium samples available, a good training result would be difficult to obtain. From method design point of view, with a modularized design, changing the feature set of one stage would not affect the other. We could thus better focus on a well-defined objective with reduced complexity. In addition, since the graph cut algorithm would only produce approximations for multi-class problems, it was more accurate to employ the binary formulations.

TABLE I
SUMMARY OF OUR DATASET. (A) TUMORS. (B) ABNORMAL LYMPH NODES.

Category	# volumes
Well-within lung field	21
Adjacent to pleural	26
Adjacent to mediastinum	27
Invasion into mediastinum	19

(a)

Category	# volumes
Well-within mediastinum	27
Adjacent to left lung field	19
Adjacent to right lung field	19

(b)

VI. MATERIALS AND EVALUATION METHODS

Our dataset used in this study consists of image scans from 85 patients diagnosed with NSCLC, acquired using a Siemens TrueV 64 PET-CT scanner at the Royal Prince Alfred Hospital, Sydney. By selecting image slices covering all abnormalities in the thorax from each patient scan, the dataset contained 85 3D image sets comprising 2480 transaxial PET-CT slice pairs. The reconstructed matrix size of each transaxial CT slice was 512×512 voxels with a slice thickness of 3mm. For PET images, the matrix size was 168×168 with a slice thickness of 5mm. During the preprocessing, the PET images were linearly interpolated to the same voxel size as the CT images, and FDG uptake normalized into SUV based on the injected dose and patient's weight. We will simply refer to a slice pair as a slice in the following.

For each 3D image set, a senior expert indicated the quantities of lung tumors and abnormal lymph nodes, with descriptions of their locations and characteristics. This senior expert has read over 8000 PET-CT lung cancer studies. To encode these ground truths for training and testing, we also created the following for each tumor and abnormal lymph node: (i) an approximate 3D bounding box indicating its span of x, y and z coordinates; and (ii) a key slice showing its most prominent feature (e.g. size and spatial extent) with a corresponding mask depicting the region labelings. A total of 93 lung tumors and 65 abnormal lymph nodes were annotated, and grouped into several categories, as shown in Table I.

Linear-kernel SVMs was used as the classifiers in this work. Based on our experiments, the linear-kernel construct was more suitable than the polynomial and Gaussian radial basis functions. The training was performed by first selecting 10 3D image sets as the training pool, which roughly comprised of two annotated volumes for each category of tumors and abnormal lymph nodes. From this training pool, an initial set of training samples was chosen manually to represent the typical patterns of different classes. A bootstrapping approach was then conducted to include training samples incrementally until no further improvements could be observed based on the training pool. No training samples were replaced and only additional samples were added during this bootstrapping procedure. The training pool was limited to 10 3D image sets to minimize the risk of over-fitting; note that since 10 3D sets actually contained a large number of image regions that could be used as training samples, the bootstrapping procedure was effective in reducing the size of training samples to about only 10% from the training pool.

We evaluated the recall (R), precision (P) and F-score (F) of the detection results at each stage:

$$R = TP/(TP + FN) \tag{22}$$

$$P = TP/(TP + FP) \tag{23}$$

$$F = 2 \cdot R \cdot P/(R + P) \tag{24}$$

where TP, FN and FP were the numbers of true positive, false negative and false positive detections, which were all object (i.e. volume) based. In the PASCAL standard [41], an object detected with at least 50% overlap with the ground truth volume would be considered TP. Since in our dataset the ground truth volume was an approximate 3D bounding box, TP was determined based on visual inspections of all slices, with the criteria that (i) the detected volume depicted the actual abnormality relatively closely, with imprecise delineations allowed around the boundaries; and (ii) a volume was labeled (T or N) consistently throughout the stack of slices, without any mislabeling within the detected volume. Inclusion of the latter criterion implied a more stringent requirement than the PASCAL standard.

In addition, the detection performance was also measured with receiver operating characteristics (ROC) curves. The ROC curve was a plot of true positive rates (TPR) versus false positive rates (FPR), by varying the classification thresholds based on the probability estimates of the labeling outputs. The probability estimates of all regions from a volume were averaged as the object-level measure, which was gathered for all TP, FN and FP volumes. The area under the curve (AUC) was then computed to quantify the detection performance. Furthermore, the discriminative power of each type of feature was evaluated by analyzing the probability estimates of different classes. For features with low dimensions, the separation of feature spaces were visualized based on the data distribution of our dataset.

TABLE II
RESULTS OF ABNORMALITY DETECTION WITH LOW-LEVEL FEATURES ONLY OR INCLUDING THE HIGH-LEVEL FEATURES. (A) THE NUMBERS OF TRUE POSITIVES (TP), FALSE NEGATIVES (FN) AND FALSE POSITIVES (FP). (B) THE RECALL, PRECISION AND F-SCORE (%).

	Low	High
TP	157	155
FN	1	3
FP	51	20

(a)

	Low	High
Recall	99.4	98.1
Precision	75.7	88.6
F-score	85.9	93.1

(b)

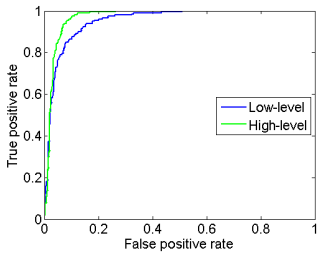


Fig. 10. ROC curves of abnormality detection, for low-level features only or including the high-level features.

VII. RESULTS AND DISCUSSIONS

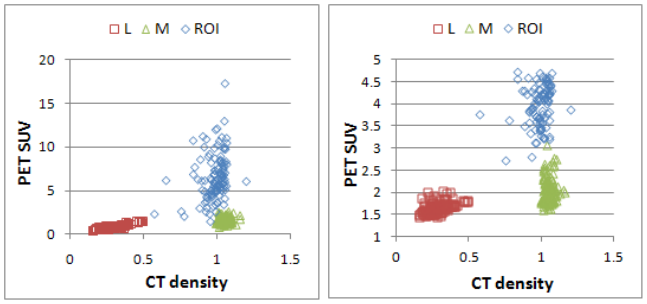
A. Detection of Abnormalities

We first report the recall, precision and F-score of abnormality detections on our dataset. A 3D ROI volume (regardless of tumor or abnormal lymph node) that was successfully detected was considered as true positive. A missed detection was false negative. A detected ROI volume that was actually a normal thoracic area was then false positive. The constant C_2 in (3) was determined as 0.18 using the learning-based procedure.

As shown in Table II, with low-level features only, our method achieved 99.4% recall of abnormalities, with only one false negative detection, but at the expense of 51 false positives. With the inclusion of high-level features, the false positives were largely reduced and the precision of detection increased by about 13%. Although two more false negatives were produced with the high-level features, the overall performance was higher (93.1% F-score) with a better balance between the recall and precision.

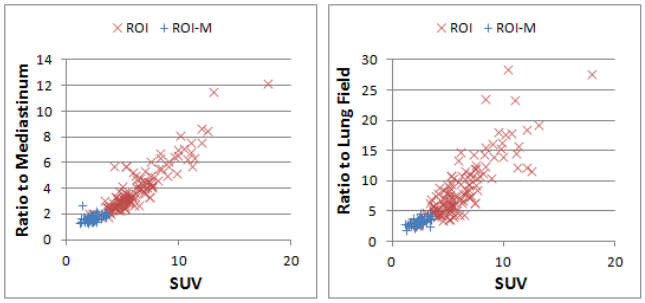
The false negatives were all at abnormal lymph nodes, which were hard to detect due to their relatively low SUVs close to the mediastinum. The false positives were detected at the high-uptake regions in the mediastinum, because of either reasons: (i) the region represented high-uptake myocardium; or (ii) the tumor in the same image set exhibited quite low SUV and hence the high-uptake region showed similar SUV to the tumor. For case (i), the stage-3 of our method targeted the detection of high-uptake myocardium, and the results will be presented in the later section. For case (ii), since the high-level feature worked based on the SUV contrast between ROIs and the mediastinum, if the contrast level was really low, such false positives could then remain.

Fig. 10 shows the ROC curves of the abnormality detection results, based on low-level features only or also including the high-level features. The AUC values of both curves were 0.9512 and 0.9691, respectively. Although the difference in



(a) (b)

Fig. 11. Scatter plots for the lung field, mediastinum and ROI areas based on (a) the original SUV; and (b) the normalized SUV.



(a) (b)

Fig. 12. Scatter plots of for ROI (the real ROI regions) and ROI-M (the M regions initially detected as ROIs at low-level), showing the SUV ratios between the detected ROI and (a) the mediastinum, and (b) the lung field.

AUC was small, it can be seen from the curves that at near 100% true positive rates, there was large reduction in false positive rates with the high-level features.

The discriminative power of the low-level intensity features can be seen from Fig. 11, showing a clearer separation between the three types of structures with the normalized SUV than the original SUV. We also evaluated the feature space based on the third quartile of SUV for L and M and the first quartile for ROI to measure the overlaps between the upper SUV range of L/M and the lower SUV range of ROI; and a similar difference in the feature separation was observed.

Fig. 12 shows the feature space of a subset of the high-level features: u_r/u_M – ratio of average SUV between ROI region R_r and the mediastinum, u_r/u_L – ratio between R_r and the lung field, and u_r – average SUV of R_r. The feature space was computed for both the real ROIs and the false positives detected with the low-level features. The two clusters were well separated for each feature dimension, but quite close at the separation boundary and hence the remaining false positive or negative ROIs after the labeling with high-level features. The discriminative power was enhanced when all four feature dimensions were integrated with different feature weights.

B. Differentiation of Tumors and Lymph Nodes

Based on the abnormality detection outputs, we then present the performance of tumor and lymph node differentiation. As shown in Table III, our CRF model was able to identify most tumors correctly, especially with the volume-level pairwise

TABLE III
THE CONFUSION MATRIX BETWEEN TUMORS AND LYMPH NODES. (A) UNARY TERM ONLY. (B) UNARY + PAIRWISE TERMS.

Ground Truth	Prediction (%)	
	T	N
T	76.3	23.7
N	11.7	88.3

(a)

Ground Truth	Prediction (%)	
	T	N
T	97.9	2.1
N	10.0	90.0

(b)

TABLE IV
RESULTS OF LYMPH NODE CLASSIFICATION BASED ON DIFFERENT CATEGORIES: WELL-WITHIN MEDIASTINUM (CAT-1), ADJACENT TO LEFT LUNG FIELD (CAT-2), AND ADJACENT TO RIGHT LUNG FIELD (CAT-3).

	# of Errors		
	Cat-1	Cat-2	Cat-3
Unary only	1	4	2
Pairwise - corrections	0	2	0
Pairwise - new errors	2	2	0
Pairwise - final	2	4	0

term. Only two tumors were misclassified, both being adjacent to the mediastinum exhibiting very similar characteristics to the lymph nodes. Although our dataset contained 25 more tumors that were also adjacent to the mediastinum and correctly classified, these two sets were particularly difficult because of their small sizes and locations at the hilar area.

The improvement with the pairwise term on the lymph nodes was relatively small, and a detailed breakdown of the classification errors is shown in Table IV. With unary term only, a total of 7 abnormal lymph nodes were partially mistaken as tumors, which meant a portion of the lymph node volume (i.e. some slices) were mislabeled. Most of these problems occurred for lymph nodes that were adjacent to the lung fields and were quite large in some slices, hence appearing very similar to tumors. With the pairwise term, such partial incorrectness were largely reduced, with only 2 errors remained for Cat-2 lymph nodes. However, with the pairwise term, a side effect was also introduced because of the spatial smoothing of labelings, resulting in 4 new errors. If a tumor and an abnormal lymph node were spatially connected at some slices, the two volumes would become joint and the pairwise term would impose higher penalties if the two sets of regions were labeled differently. Since the lymph node usually comprised fewer regions and thus lower influence on the total energy optimization, the final labeling would usually become T for both volumes. Although such an issue was addressed with the overlapping factor, which indeed helped to assign correct labelings for 55.6% volumes having such problems, some errors remained for lymph nodes that were very near to the tumors in large portions of slices.

In the above measures, we only assessed the differentiation capability of our method, given the accurately detected ROI volumes. Next, we took into account the false negative and false positive ROIs, and assessed the recall, precision and F-score of tumor and abnormal lymph nodes detections. As shown in Table V, 97.9% recall and 82.7% precision were obtained for tumors, and 86.2% recall and 88.9% precision for abnormal lymph nodes. The false negatives and false positives could be further divided into two groups based

TABLE V
RESULTS OF TUMOR (T) AND ABNORMAL LYMPH NODES (N) DETECTION. (A) THE NUMBERS OF TRUE POSITIVES (TP), FALSE NEGATIVES (FN) AND FALSE POSITIVES (FP). (B) THE RECALL, PRECISION AND F-SCORE (%).

	T	N
TP	91	56
FN	2	9
FP	19	7

(a)

	T	N
Recall	97.9	86.2
Precision	82.7	88.9
F-score	89.7	87.5

(b)

TABLE VI
THE NUMBERS OF INCORRECT DETECTIONS OF TUMORS AND ABNORMAL LYMPH NODES. (A) FALSE NEGATIVES (FN). (B) FALSE POSITIVES (FP).

	T	N
FN ROI	0	3
Mislabel	2	6

(a)

	T	N
FP ROI	13	5
Mislabel	6	2

(b)

on the causes (Table VI): (i) false negative or positive ROI detections; and (ii) incorrect differentiation between tumors and abnormal lymph nodes. Both causes have been discussed in previous paragraphs. We will present the results on false positive reduction in the following section.

The ROC curves of detection results are shown in Fig. 13. For both types of abnormalities, the improvements with the pairwise terms were evident. The AUC values for tumor detection with unary terms only or including the pairwise terms were 0.8402 and 0.9583, respectively; and for abnormal lymph nodes they were 0.8556 and 0.9078. With the pairwise terms, true positive rates of tumors quickly raised near to 100% with low false positive rates. At the low range of false positive rates for abnormal lymph nodes, the true positive rates tended to saturate at around 90% only, mainly due to some volumes particularly difficult to detect for their either quite low SUVs or very similar characteristics to tumors.

To further evaluate the discriminative power of the spatial and contextual feature set, the probability estimates of the T and N classification based on the unary term were summarized for all key slices. As shown in Fig. 14, with the quad-radial global histogram only, about 83.9% tumors and 77.8% abnormal lymph nodes were correctly identified. The surrounding contour histogram mainly improved the detection for abnormal lymph nodes to about 85.7%, since lymph nodes were surrounded mostly by regions of the mediastinum. By

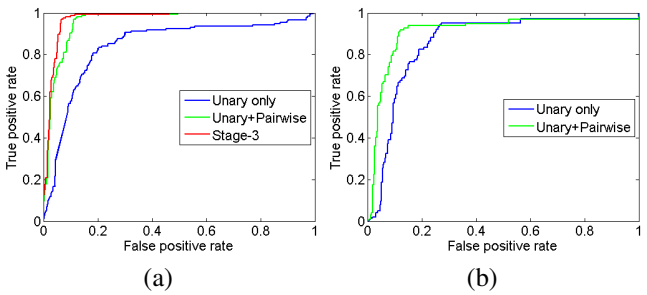


Fig. 13. ROC curves of detection for (a) tumors and (b) abnormal lymph nodes, with unary terms only or including the pairwise terms. The final detection results after stage-3 refinements are also shown for tumors.

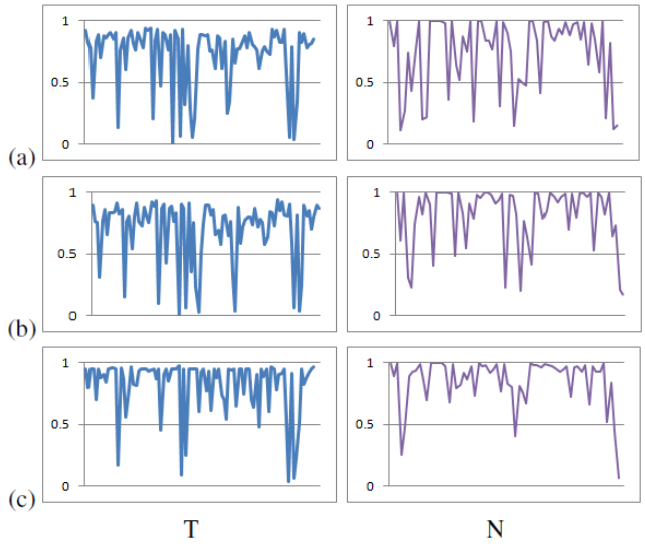


Fig. 14. Distributions of the probability estimates of tumors (T) and abnormal lymph nodes (N) on key slices, where x axis is the slice index and y axis is the probability value. A probability larger than 0.5 indicates a correct labeling, obtained with different combinations of features: (a) the quad-radial global histogram only; (b) also including the surrounding contour histogram; (c) also including the pleural distances.

incorporating spatial location information of the ROI relative to the lung field, the pleural distances were effective in further improving the classifications of tumors and abnormal lymph nodes to about 90.3% and 93.7%.

Fig. 15 shows examples of successful differentiations between tumors and abnormal lymph nodes, depicted using transaxial slices that were sampled from the classified 3D volumes. The presented datasets illustrate typical cases of different categories of tumors and abnormal lymph nodes, and were not included in the training data. While the first two examples have tumors within the lung field or attached to the pleural, the other five examples display more complex cases where the tumors are adjacent to the mediastinum and lymph nodes are adjacent to the lung fields making them difficult to be differentiated. Since the thoracic appearance was quite different for each case, basic features such as shapes and locations would not be robust enough; our method, however, handled these cases successfully.

C. Refinement of Tumor Regions

Among all the false positive ROIs, 9 were actually high-uptake myocardium volumes and labeled as tumors (Table VIIa). The refinement stage then reduced the number of false positive tumors to 2, by labeling the other 7 as mediastinum (Table VIIb) with the CRF model. The unary term, i.e. SVM classification based on the unary features, was able to identify 5 such high-uptake myocardiums; the other 4 volumes were partially labeled as M, as normally only the regions belonging to the lower portions of the myocardium would exhibit distinctive myocardium features. By only labeling those regions that were highly representative of myocardium as M, the refinement stage was successful in keeping the labeling of all true positive tumors unchanged. The

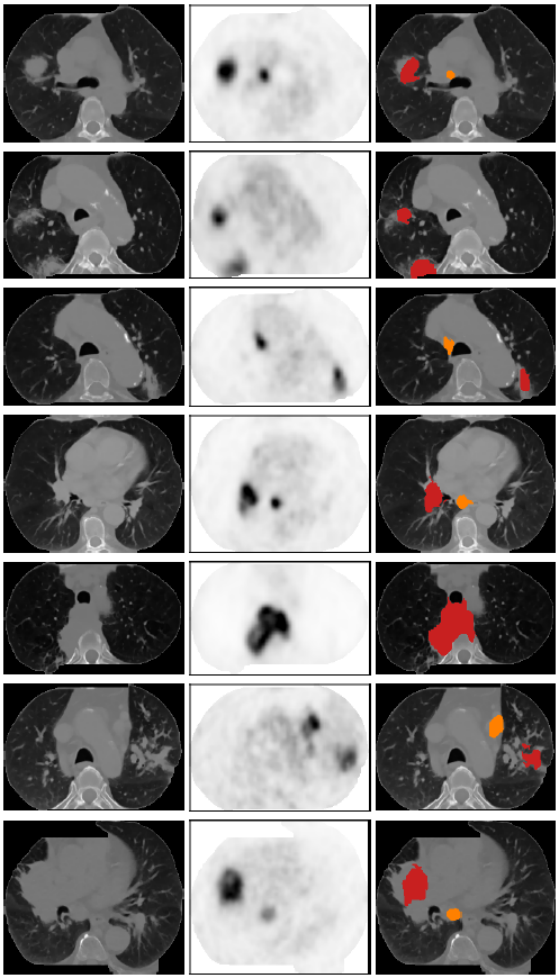


Fig. 15. Examples of detected tumors and abnormal lymph nodes of different categories. Each row shows one example, using a transaxial slice view for easier visualization. The left column shows the CT slices, the middle column shows the PET slices, and the right column depicts the detection results highlighted on the CT images (red for T and orange for N).

TABLE VII
THE NUMBERS OF FALSE POSITIVES AND REFINEMENT RESULTS. (A) FALSE POSITIVE ROIS DETECTED DUE TO HIGH-UPTAKE REGIONS IN MYOCARDIUM OR OTHER AREAS. (B) THE REFINEMENTS OF THE HIGH-UPTAKE MYOCARDIUM.

	Myocardium	Others
T	9	4
N	0	5

(a)

	Unary	Pairwise
T	5	2
M	4	7

(b)

pairwise term improved the refinements by incorporating the volume-level spatial features. However, when a large portion of the volume was mislabeled, the volume would still be detected as T, and hence the remaining 2 false positive tumors. After all the three stages of processing, the final detection results of tumors and abnormal lymph nodes are shown in Table VIII. The improvements resulted from stage-3 is also shown in Fig. 13a, with AUC value of 0.9705.

If only HOG features were used without the pleural-distance feature, the feature set became less discriminative as the probabilities of a region being M or T would be quite similar,

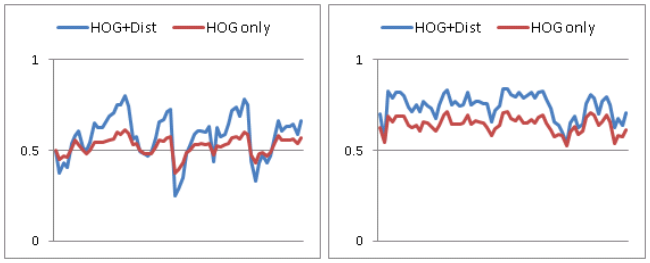
TABLE VIII
THE FINAL RESULTS OF TUMOR (T) AND LYMPH NODE (N) DETECTION. (A) THE NUMBERS OF TRUE POSITIVES (TP), FALSE NEGATIVES (FN) AND FALSE POSITIVES (FP). (B) THE RECALL, PRECISION AND F-SCORE (%).

	T	N
TP	91	56
FN	2	9
FP	12	7

(a)

	T	N
Recall	97.9	86.2
Precision	88.4	88.9
F-score	92.4	87.5

(b)



(a) (b)

Fig. 16. Distributions of the probability estimates of tumors and myocardium on key slices, where x axis is the slice index and y axis is the probability value. A probability larger than 0.5 indicates a correct labeling. (a) Probability of M for the myocardium regions. (b) Probability of T for the tumor regions.

as shown in Fig. 16. The number of unidentified myocardium also increased by one without the pleural distances. All true positive tumors were correctly classified in both cases. We also tested the classification with only pleural distances but not the HOG features; however, such a feature set did not produce an acceptable performance, with about 23% accuracy only. Therefore, the HOG features appeared to be the most representative features for differentiating between the myocardium and tumor, while the pleural distances were helpful in enhancing the separations.

D. Performance Comparison

Besides evaluating the individual components of our proposed method, we also performed comparisons with other existing approaches. A summary of the various datasets used and the detection recalls and precisions are listed in Table IX.

We first compared the results to the previous method [4], which seems to be the only work that addresses the detection of both tumor and abnormal lymph nodes (Table IX row 2). Since the original work [4] was evaluated on a different dataset, we repeated the test in our current dataset (Table IX row 3). Our new method exhibited clear improvements, which were mainly attributed to three factors: (i) fewer abnormal lymph nodes mislabeled as tumors, especially for those nodes lying close to the lung field, with the new spatial and contextual features (unary term); (ii) fewer tumors mislabeled as abnormal lymph nodes, especially those previously caused by inconsistent labeling of regions in one tumor volume, with the spatially-smoothed 3D volume labeling (pairwise term); and (iii) fewer high-uptake myocardium areas mislabeled as abnormalities.

For further evaluation, and due to few works reported in our problem domain, we also present comparisons with other less related works, as shown in Table IX (rows 4 to 5). We chose

TABLE IX
PERFORMANCE ON OTHER DATASETS AND COMPARISON WITH OTHER METHODS. '–' MEANS NA. *: RERUN RESULTS ON CURRENT DATASET. ^: NOT SHOWN IN THE REFERENCED PAPER BUT COMPUTED BASED ON THE REPORTED RESULTS.

Method	Test size	# tumors	# nodes	T-Recall (%)	T-Prec (%)	N-Recall (%)	N-Prec (%)
Proposed method	85 cases	93	65	97.9	88.4	86.2	88.9
Song et.al. [4]	50 cases	53	36	84.4	83.8^	77.8	76.9^
Song et.al. [4]*	85 cases	93	65	89.3	79.1	72.3	82.5
Jafar et.al. [8]	3 cases	3	–	75	100	–	–
Gubbi et.al. [11]	7 cases	7	–	80	100	–	–

TABLE X
AVERAGE COMPUTATION TIME IN SECONDS FOR A 3D IMAGE SET.

Stage-1			Stage-2		Stage-3	
Cluster	Low	High	Unary	Pairwise	Unary	Pairwise
36.8	12.5	3.1	21.9	6.2	5.5	3.8

to list the recent works that proposed automatic detection methods for lung tumors on PET-CT images, and reported good precision and recall. While both the tumor detection methods [8], [11] (Table IX rows 4 to 5) achieved perfect precisions, they were only validated on three and seven image studies. Furthermore, both approaches did not consider cases with abnormal lymph nodes, which could otherwise affect the performance of tumor detections, e.g. a lower precision if misclassifying lymph nodes as tumors. The method [8] also relied on segmentation accuracy of lung fields, which might not be robust if a lung tumor was adjacent to the mediastinum; and the method [11] only discussed about high-uptake areas in the heart that could cause confusions.

E. Computational Efficiency

Table X shows the average computational time for a 3D image set (about 30 transaxial slice pairs). Our method was implemented in Matlab v2009b, running on a standard PC with a 2.66 GHz dual core CPU. In total, an average of 89.8 s was required, with about 41% of the time spent on region clustering. The extraction of unary features for stage-2 was the second most time consuming component taking 21.9 s. The low-level labeling at stage-1 took about 12.5 s, which was mainly incurred by the SUV normalization step. The other processings were much faster with lower feature dimensions.

VIII. CONCLUSION AND FUTURE WORK

We have presented a fully automatic detection method for lung tumor and disease in regional lymph node from PET-CT thoracic images. Abnormalities are first detected based on the low-level intensity and neighborhood features and high-level contrast-type features, with a two-level SVM classification. The detected abnormalities are then differentiated into tumors or abnormal lymph nodes with a CRF model, based on the unary-level contextual and spatial features and pairwise-level spatial features. Another CRF model is then employed to relabel the detected tumors as either true tumor or mediastinum by filtering the high-uptake myocardium areas. The detection recall and precision were measured for each stage, and the discriminative power of each feature set was also evaluated. On a clinical dataset of 93 tumor and 65 abnormal lymph nodes from 85 3D image sets, we found the proposed method showed high detection performance and capability in handling a wide variety of abnormal patterns.

We are working on further reducing the false negatives of abnormal lymph nodes, which were produced due to undetected abnormalities or mislabeled as tumors, accounting for the 13.8% less recall from a total recall level. Our current investigation is on improving shape analysis by coupling with spatial priors for better lymph node detection, while avoiding any impact on tumor detection.

ACKNOWLEDGMENT

This work was supported in part by ARC and PolyU grants. The authors would like to thank Professor Michael Fulham from the PET and Nuclear Medicine Department in Royal Prince Alfred Hospital in Sydney, Australia, for providing the clinical data sets.

REFERENCES

[1] World Health Organization, "Cancer, fact sheet no. 297," 2011, http://www.who.int/mediacentre/factsheets/fs297/.
[2] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, and A. Trotti (Eds.), *AJCC cancer staging handbook, 7th ed*. Springer, 2010.
[3] W. Wever, S. Stroobants, J. Coolen, and J. Verschakelen, "Integrated PET/CT in the staging of nonsmall cell lung cancer: technical aspects and clinical integration," *Eur. Respir. J.*, vol. 33, pp. 201–212, 2009.
[4] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng, "Discriminative pathological context detection in thoracic images based on multi-level inference," *in MICCAI 2011, LNCS*, vol. 6893, pp. 185–192, 2011.
[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *in Proc. ICML*, pp. 282–289, 2001.
[7] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng, "Thoracic image case retrieval with spatial and contextual information," *in Proc. ISBI*, pp. 1885–1888, 2011.
[8] I. Jafar, H. Ying, A. Shields, and O. Muzik, "Computerized detection of lung tumors in PET/CT images," *in Proc. EMBC*, pp. 2320–2323, 2006.
[9] Y. Cui, B. Zhao, T. Akhurst, J. Yan, and L. Schwartz, "CT-guided automated detection of lung tumors on PET images," *in SPIE Med. Imaging*, vol. 6915, p. 69152N, 2008.
[10] C. Ballangan, X. Wang, S. Eberl, M. Fulham, and D. Feng, "Automated lung tumor segmentation for whole body PET volume based on novel downhill region growing," *in SPIE Med. Imaging*, vol. 7623, p. 76233O, 2010.
[11] J. Gubbi, A. Kanakatte, T. Kron, D. Binns, B. Srinivasan, N. Mani, and M. Palaniswami, "Automatic tumour volume delineation in respiratory-gated PET images," *J. Med. Imag. Radia. Oncol.*, vol. 55, pp. 65–76, 2011.
[12] G. Saradhi, G. Gopalakrishnan, A. Roy, R. Mullick, R. Manjeshwar, K. Thielemans, and U. Patil, "A framework for automated tumor detection in thoracic FDG PET images using texture-based features," *in Proc. ISBI*, pp. 97–100, 2009.

[13] S. Renisch, R. Opfer, and R. Wiemker, "Towards automatic determination of total tumor burden from PET images," *in SPIE Med. Imaging*, vol. 7624, p. 76241T, 2010.

[14] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng, "Automatic detection of lung tumor and abnormal regional lymph nodes in PET-CT images," *J. Nucl. Med. 52*, vol. 52, no. Supplement 1, p. 211, 2011.

[15] H. Gutte, D. Jakobsson, F. Olofsson, M. Ohlsson, S. Valind, A. Loft, L. Edenbrandt, and A. Kjaer, "Automated interpretation of PET/CT images in patients with lung cancer," *Nucl. Med. Commun.*, vol. 28, no. 2, pp. 79–84, 2007.

[16] A. Kiraly, L. Zhang, C. Novak, D. Naidich, and L. Guendel, "Novel method and applications for labeling and identifying lymph nodes," *in SPIE Med. Imaging*, vol. 6911, p. 691111, 2007.

[17] K. Lu, S. Merritt, and W. Higgins, "Extraction and visualization of the central chest lymph-node stations," *in SPIE Med. Imaging*, vol. 6915, p. 69151B, 2008.

[18] M. Feuerstein, D. Deguchi, T. Kitaska, S. Iwano, K. Imaizumi, Y. Hasegawa, Y. Suenaga, and K. Mori, "Automatic mediastinal lymph node detection in chest CT," *in SPIE Med. Imaging*, vol. 7260, p. 72600, 2009.

[19] J. Feulner, S. K. Zhou, M. Huber, J. Hornegger, D. Comaniciu, and A. Cavallaro, "Lymph nodes detection in 3-D chest CT using a spatial prior probability," *in Proc. CVPR*, pp. 2926–2932, 2010.

[20] M. Feuerstein, B. Glocker, T. Kitasaka, Y. Nakamura, S. Iwano, and K. Mori, "Mediastinal atlas creation from 3-D chest computed tomography images: application to automated detection and station mapping of lymph nodes," *Med. Image Anal.*, vol. 16, no. 1, pp. 63–74, 2011.

[21] T. Kitasaka, Y. Tsujimura, Y. Nakamura, K. Mori, Y. Suenaga, M. Ito, and S. Nawano, "Automated extraction of lymph nodes from 3-D abdominal CT images using 3-D minimum directional difference filter," *in MICCAI 2007, LNCS*, vol. 4792, pp. 336–343, 2007.

[22] A. Bardu, M. Suehling, X. Xu, D. Liu, S. Zhou, and D. Comaniciu, "Automatic detection and segmentation of axillary lymph nodes," *in MICCAI 2010, LNCS*, vol. 6361, pp. 28–36, 2010.

[23] H. Zaidi and I. E. Naqa, "PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques," *Eur J. Nucl Med Mol Imaging*, vol. 37, no. 11, pp. 2165–2187, 2010.

[24] J. Kuhnigk, V. Dicken, L. Bornemann, A. Bakai, D. Wormanns, S. Krass, and H. Peitgen, "Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans," *IEEE Trans. Medical Imaging*, vol. 25, no. 4, pp. 417–434, 2006.

[25] M. Kakar and D. R. Olsen, "Automatic segmentation and recognition of lungs and lesions from CT scans of thorax," *Comput. Med. Imaging Graph.*, vol. 33, no. 1, pp. 72–82, 2009.

[26] Q. Song, M. Chen, J. Bai, M. Sonka, and X. Wu, "Surface-region context in optimal multi-object graph-based segmentation: robust delineation of pulmonary tumors," *in IPMI 2011, LNCS*, vol. 6801, pp. 61–72, 2011.

[27] V. Potesil, X. Huang, and X. Zhou, "Automated tumour delineation using joint PET/CT information," *in SPIE Med. Imaging*, vol. 6514, p. 65142Y, 2007.

[28] H. Gribben, P. Miller, G. Hanna, K. Carson, and A. Hounsell, "MAP-MRF segmentation of lung tumours in PET/CT images," *in Proc. ISBI*, pp. 290–293, 2009.

[29] J. Wojak, E. Angelini, and I. Bloch, "Joint variational segmentation of CT-PET data for tumoral lesions," *in Proc. ISBI*, pp. 217–220, 2010.

[30] J. Yan, T. Zhuang, B. Zhao, and L. Schwartz, "Lymph node segmentation from CT images using fast marching method," *Comput. Med. Imaging Graph.*, vol. 28, no. 1-2, pp. 33–38, 2004.

[31] J. Yan, B. Zhao, L. Wang, A. Zelenetz, and L. Schwartz, "Marker-controlled watershed for lymphoma segmentation in sequential CT images," *Med. Phys.*, vol. 33, no. 7, pp. 2452–2460, 2006.

[32] D. Maleike, M. Fabel, R. Tetzlaff, H. Tengg-kogligk, T. Heimann, H. Meinzer, and I. Wolf, "Lymph node segementation on CT images by a shape model guided deformable surface method," *in SPIE Med. Imaging*, vol. 6914, p. 69141S, 2008.

[33] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng, "A content-based image retrieval framework for multi-modality lung images," *in Proc. CBMS*, pp. 285–290, 2010.

[34] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," *in ECCV 2008, LNCS*, vol. 5305, pp. 705–718, 2008.

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[36] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation," *in ECCV 2006, LNCS*, vol. 3951, pp. 1–15, 2006.

[37] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1222–1239, 2001.

[38] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, 2004.

[39] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, 2004.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in Proc. CVPR*, pp. 886–893, 2005.

[41] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.