Disease-Specific Context Modeling and Retrieval with Fast Structure Localization

Yang Song[1], Weidong Cai[1], David Dagan Feng[1,2,3]
[1]Biomedical and Multimedia Information Technology (BMIT) Research Group,
School of Information Technologies, University of Sydney, Australia
[2]Center for Multimedia Signal Processing (CMSP), Department of Electronic &
Information Engineering, Hong Kong Polytechnic University, Hong Kong
[3]Med-X Research Institute, Shanghai Jiao Tong University, China

Abstract

Medical images showing the same type of disease can be visually very different, but they often exhibit common structural and semantic patterns in localized areas. The retrieved images for a given one are thus expected to be similar in disease-specific contexts other than basic visual similarities; and we propose to incorporate automatic context detection into medical image retrieval in this paper. We design a fast multi-class discriminative model to first localize the structures of interests. The contextual information is then inferred from the localized bounding boxes to rank the image similarities based on a learned distance function. Positron emission tomography – computed tomography (PET-CT) thoracic imaging data from patients with non-small cell lung cancer (NSCLC) are used in this study, and our approach shows high retrieval performance.

1. Introduction

Digital medical images are produced in ever increasing quantities and play an important role in modern health care. There is great interest for physicians to gather valuable information from the large collection of images for decision support, by retrieving images similar to a given one. The early medical image retrieval systems mainly measure the image similarities based on low-level visual features, such as texture and shape [9]. However, medical images, especially the ones with pathology, may not be well compared using only these low-level visual features. For example, two images showing different stages of lung cancer may contain tumors of similar appearances, but one with tumor invasion into the chest wall and metastasis in regional lymph nodes. Therefore, the disease-specific contextual features are important for measuring the image similarities.

Such contexts are incorporated in medical image retrieval in predominately two ways: (i) the low-level features are classified into a number of concept categories and represented as bag-of-features for the whole image [1, 10]; and (ii) the low-level features are first classified to detect or align regions of interest, and local features extracted for the regions are used to compute image dissimilarity [13, 15, 11]. The region-based approach has an advantage over the bag-of-features that the background regions do not affect the image similarity measure; but by focusing on comparing only the regions of interest, it disregards the contexts of the regions within the anatomical structures.

A recent work on thoracic image retrieval [12] attempts to mitigate this problem by delineating not only the pathological suspected regions of interest, but also the related anatomical structures. However, the approach relies on clustering to form irregularly-shaped regions, which is computational inefficient; and the necessity of performing such boundary delineation is unsure – it might be possible to achieve effective retrieval by locating the bounding boxes of the structures. Furthermore, in the work [12], each region is classified based on its local feature and adjacent regions only, without considering the structural interactions, e.g. object-based relationships between the tumor and lung field. Such higher-level interactions have been recently explored for multi-class object localization in general computer vision problems, and demonstrated better localization performance [4, 6, 2]. For medical images, however, the features used for object localization should be specifically designed to represent the pathological and anatomical characteristics; and the inference of object locations can also benefit from prior knowledge of the anatomy and diseases.

In this work, we propose a new medical image retrieval framework based on disease-specific contexts in thoracic imaging studies. Our main contributions are threefold. First, we design a discriminative model with three levels of features to localize the major structures from the thoracic

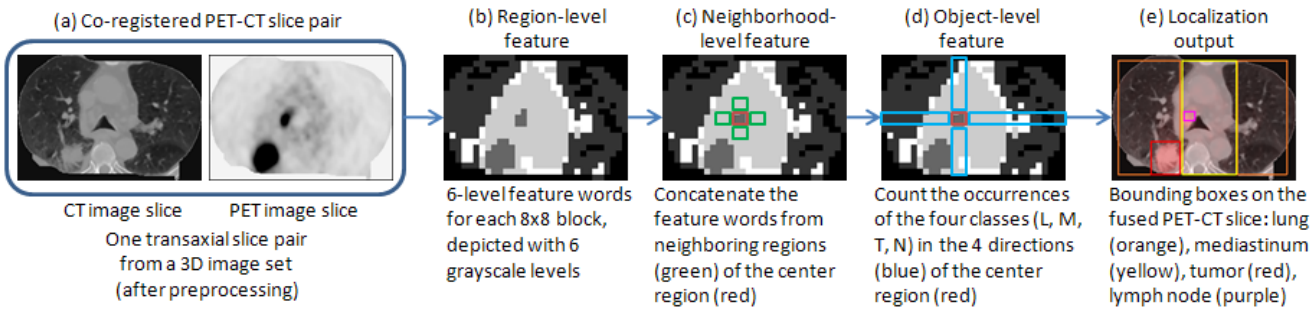

Figure 1. Illustration of the features and structure localization.

images to detect the overall contexts. Second, the localization is modeled as a fast bounding box detection incorporating anatomical and pathological knowledge to avoid time-consuming region boundary delineations. Third, the ranking of image dissimilarity is based on discriminative learning with integrated precision measure.

In this paper, we present our work using PET-CT thoracic images with NSCLC. The PET-CT scanner produces a 3D image set comprising co-registered transaxial slice pairs. Although CT images depict the anatomical structures well, they have poor soft tissue contrast resolution and difficulty in separating tumors from normal tissues. PET images have high contrast and highlight the abnormal areas well, but have lower resolution than the CT data and so do not delineate the precise location of an abnormality [14]. The integrated PET-CT imaging thus provides advantages in depicting the complementary anatomical (CT) and functional (PET) patient information, and is now widely accepted as the best imaging technique for cancer staging.

Staging of lung cancer is mainly based on the characteristics of the primary lung tumor, especially its spatial context (e.g. adjacent to the lung wall, invasion into the mediastinum, etc); and any detection of disease in regional lymph nodes and their locations. Figure 1a shows an example transaxial slice pair containing a primary lung tumor at the posterior lung field near the mediastinum, and an abnormal lymph node next to the right lung. The retrieved images (3D image sets) are thus expected to exhibit similar pathological contexts representing similar stage of lung cancer, which is our aim of this work.

2. Contextual Image Retrieval

The proposed contextual image retrieval method consists of the preprocessing, structure localization and image ranking components. The preprocessing removed the patient bed and soft tissues outside of the lung and mediastinum from the PET-CT thoracic images automatically based on simple thresholding, morphological operations and connected component analysis [12]. The rest of processing steps were then performed on the preprocessed images.

2.1. Structure Localization

We firstly modeled the PET-CT slice image pair $\mathbf{I}$ as a collection of regions representing the major structures in the thorax – left (L1) and right (L2) lung fields, mediastinum (M), tumor (T) and abnormal lymph nodes (N):

$$\begin{gathered}\mathbf{I} = \bigcup\{R_{L1}, R_{L2}, R_M, \bigcup_0^{X_T} R_T, \bigcup_0^{X_N} R_N\} \\ R = \{x_1, y_1, x_2, y_2\}\end{gathered} \tag{1}$$

where a region R was depicted by its x and y coordinates of the top left and bottom right corners. The number of T or N objects could range from 0 to $X_{T/N}$. The bounding boxes could overlap with each other, and the union of all regions should cover the entire slice (example in Figure 1e).

Denoting the set of regions in $\mathbf{I}$ as $\mathbf{R} = \{R_i, i = 1, ..., 3 + X_T + X_N\}$, we then defined the score of delineating $\mathbf{I}$ into regions $\mathbf{R}$ with class labeling $\mathbf{Y}$ as:

$$S(\mathbf{I}, \mathbf{R}, \mathbf{Y}) = \sum_i \alpha'_{y_i} f_i + \sum_i \beta'_{y_i} g_i + \sum_i \gamma'_{y_i} h_i \tag{2}$$

where y_i was the class labeling of region R_i, with the possible value of 1 to K ($K = 4$: L, M, T and N classes); f_i, g_i and h_i were the three levels of features (region, neighborhood, and object levels) of R_i; and α, β and γ were the respective feature weights associated with y_i. The goal was then to find the set of regions $\mathbf{R}$ with the labeling $\mathbf{Y}$ that maximized the score S for the slice pair $\mathbf{I}$.

Region-level features. For a PET-CT slice pair, the co-registered PET and CT image slices were each divided into 8x8-pixel blocks. Texture features were extracted from each block: the mean, standard deviation, skewness and kurtosis of the Gabor filtered slices [3]. The features from both modalities for the corresponding block were combined, and a k-means clustering was then applied to form six visual words (Figure 1b). Six was chosen because the clustering output closely resembled the original images visually. The bag-of-feature representation of the image block was

then f_i, which was 6-dimensional as the occurrence frequencies of the visual words in the image block. Since a candidate region R_i could (and indeed normally) contain multiple blocks, the feature f_i was then consolidated from all comprising blocks in R_i.

Neighborhood-level features. Since the surrounding contexts were informative for determining region types, especially for differentiating between T and N, the spatial features beyond the region level were introduced. The neighborhood-level feature g_i was computed from the four neighboring areas of R_i with the same size as R_i (Figure 1c). The neighbors' region-level features (each of 6-dimensional) were concatenated to form g_i, which was thus 24-dimensional. The concatenation was ordered by placing the neighbors nearer to the image center first, so that the feature was independent of the absolute location of R_i. For example, if R_i was left (or right) to the image center, its neighboring regions were concatenated in a counter-clockwise (or clockwise) order starting from the neighboring region to its right (or left).

Object-level features. Although the neighborhood-level features incorporated the surrounding contexts, the rather rigid spatial locations of the regions limited the descriptiveness of the spatial features. They could not describe the object-level features, i.e. relative spatial locations between regions of the various types (L, M, T and N). Such spatial information, however, was important to effectively differentiated between T and N. For example, T should be entirely or partially within L, while N should reside in M. Therefore, the object-level feature was incorporated. Specifically, the object-level feature vector h_i was 16-dimensional ($K \times O$) encoding the spatial relationships between $R_j \in \{\mathbf{R} - Ri\}$ and R_i. O was four for the *above*, *below*, *left to*, and *right to* relationships (Figure 1d). Since each region R_j could be of multiple directions relative to R_i (e.g. both above and right to), it would contribute to 0 to O numbers of elements of h_i:

$$h_i(k,o) = \textstyle\sum_j f_j^o / a_i, \quad s.t.\ y_j = k,\ sr_{j,i} = o \qquad (3)$$

where sr stood for the spatial relationship between R_j and R_i; f_j^o was the portion of f_j that met the spatial relationship criteria between R_j and R_i; and a_i was the size of R_i. Note that the main distinction of h_i compared to g_i was that, the extent of the surrounding contexts was dependent on the other candidate regions R_j, rather than the neighboring areas that were determined by R_i only. This object-level feature also captured the interaction between all candidate regions in the image $\mathbf{I}$, so that detection of one region would affect the localization of other regions.

Learning of feature weights. The score S needed to be directly comparable between classes in order to solve $\text{argmax}_{\mathbf{R},\mathbf{Y}}\, S$ for $\mathbf{I}$, thus the feature weights α, β and γ needed to be globally consistent. We denote by F, G and H the feature vectors of the region set $\mathbf{R}$. Considering f_i a $K \times 6$ vector with 6 non-zero entries, which were the placeholders for y_i, then $F = \sum_i f_i$, and similarly $G = \sum_i g_i$ and $H = \sum_i h_i$, where i indexed the regions of $\mathbf{R}$. The feature weights α, β and γ thus corresponded to F, G and H, incorporating all K region classes into one vector. The score function was then rewritten as:

$$S(\mathbf{I},\mathbf{R},\mathbf{Y}) = \langle \alpha \cdot F \rangle + \langle \beta \cdot G \rangle + \langle \gamma \cdot H \rangle = \langle \omega \cdot V \rangle \quad (4)$$

where the value of V depended on both $\mathbf{R}$ and $\mathbf{Y}$. We then learned ω with a large-margin optimization method, similar to the triplet learning framework [5].

Detection of bounding boxes. Exhaustive search for $\mathbf{R}$ and $\mathbf{Y}$ was not feasible due to the large number of possible combinations. We thus designed a fast inference algorithm to reduce the search space by pruning away impossible bounding box arrangements based on the anatomical and pathological characteristics of the thorax, as listed in Algorithm 1 (example in Figure 1e). Such an inference procedure was fast, because the scanning computation during each iteration was linear to the image width (divided by 8), and the total number of iterations was limited due to the small number of T or N candidates per image and each with only four possible class assignments.

2.2. Image Ranking

The features extracted for R_T and R_N of a PET-CT slice pair were denoted as E_T and E_N, and comprised three types: (1) texture: the mean, standard deviation, skewness and kurtosis of R_T and R_N on the Gabor filtered PET and CT slices; (2) spatial: the distances to the four sides of R_{L1} or R_{L2} for R_T depending on whether the tumor was in the right or left lung lobe, and the distances to the four sides of the R_M for R_N; and (3) shape: the size of R_T and R_N, and the length of the long and short axis.

The contextual features for a 3D image set, C^T and C^N, were then calculated by weighted combination of the slice-level features E_T and E_N:

$$C^{T/N} = \frac{\sum_p E_p^{T/N} S_p^{T/N}}{\sum_p S_p^{T/N}} \qquad (6)$$

where p was the index of the slice pair within the 3D image set. E_p were from regions that were spatially (z direction) connected on adjacent image slices. S_p^T and S_p^N were the scores for the detected T and N regions indicating the level of confidence of the detections:

$$S_p^{T/N} = \langle \omega_{T/N} \cdot V_{T/N} \rangle \qquad (7)$$

So, the slice pairs with better delineated R_T or R_N would contribute more to the 3D image-level feature.

Similarity measure. Given the query 3D image set I and the reference image J, their distance was defined as:

$$D_{I,J} = \langle W \cdot \frac{|C_I - C_J|}{C_I + C_J} \rangle = \langle W \cdot C_{I,J} \rangle \qquad (8)$$

Algorithm 1: Inference of bounding boxes

Data: the PET-CT slice image pair $\mathbf{I}$.
Result: the bounding box set $\mathbf{R}$ and associated class labels $\mathbf{Y}$ of $\mathbf{I}$.
forall the 1×2 *block (*8×16 *pixels)* r_i *in* $\mathbf{I}$ *then* **do**
  Classify L/M/T/N based on its region- and neighborhood-level features y_i:

$$\underset{k}{\operatorname{argmax}} \sum_i \alpha'_k f_i + \sum_i \beta'_k g_i, \ \forall k = 1...K \quad (5)$$

end
repeat
  Fill T/N areas into bounding boxes, R_T and R_N;
  forall the $1\times$*[the height of* R_T *or* R_N*] block to the right and left of* R_T *and* R_N **do**
    Classifying as L or M using Eq. 5;
  end
  R_M = {area continuously classified as M}, R_{L1} = {area left to R_M}, R_{L2} = {areas right to R_M}, and the height of R_{L1}, R_{L2} and R_M were the same as the image height;
  Compute total score based on the current R_T, R_N, R_{L1}, R_{L2} and R_M using Eq. 2;
  Change the labeling of R_T or R_N to another class;
until *all four possible classes have been tested for the T/N areas*;
Choose $\mathbf{R}$ and $\mathbf{Y}$ as the set with the highest score;

where C was the feature vector concatenating C^T and C^N. Not all features were equally important for differentiating two feature vectors, thus the weight W was incorporated.

Weight learning. For P 3D image sets, and using each set as the query image to retrieve X most similar image sets from the $(P-1)$ training sets, we would like the average precision of the top-ranked $X \times P$ retrievals for the P queries to be maximized:

$$\frac{1}{P} \sum_p \frac{\sum_x v(x,p)}{X} \quad (9)$$

where $v(x,p)$ was 1 if the xth retrieved image set was truly similar to the pth query image, and 0 otherwise.

To perform the optimization, we first labeled a $P \times P$ matrix $\mathbf{U}$: $U_{i,j} = 1$ if image sets i and j were similar, and 0 otherwise. Assuming a total of Q entries in $\mathbf{U}$ were 1, we then consolidated Q training samples: $T_q : \langle I, J, K \rangle$, with $U_{I,J} = 1$ and $U_{I,K} = 0$; and the large-margin optimization method was applied to solve for W. Then, based on the initially trained W, $X \times P$ retrievals were performed. Assuming image set J was similar to query image I but ranked lower than the dissimilar set K, the training sample $\langle I, J, K \rangle$ was then added, and another optimization for W was computed. Such iterative precision verification and sample addition procedure was repeated until the average precision started to drop for three consecutive loops.

3. Experimental Results

3.1. Experimental Setup

In this study, a total of 1134 transaxial PET-CT thoracic image pairs were selected from 40 patients with NSCLC, which were acquired using a Siemens TrueV 64-slice PET-CT scanner. The locations of tumors and disease in regional lymph nodes were annotated manually as the ground truth. All 40 cases contained primary lung tumors, and 22 of them contained abnormal lymph nodes. Tumor and disease in lymph node were visible in 615 and 303 images, and 138 exhibited both types of abnormalities in one image. For each patient study, the other 39 patient studies were marked similar or dissimilar as a benchmark for retrieval performance by an expert reader. The similarity of cases was determined based on the location and appearance of the tumor and abnormal lymph nodes, which would be helpful for lung cancer staging. The number of similar cases for each case ranged from 1 to 11, with an average of 4.75.

3.2. Results on Structure Localization

We first evaluated the localization performance of the four types of structures: L, M, T and N; and compared our 3-level feature design with using only region-level features. As shown in Figure 2a, our method was particularly more effective in detecting the abnormal lymph nodes, since they could only be reliably differentiated from tumors based on their spatial features. The inclusion of neighborhood and object levels of features also helped to enhance the detection of the other structure types.

We then assessed the accuracy of the tumor and lymph nodes localization relative to the lung and mediastinum (Figure 2b). We were mainly interested in evaluating whether the tumor was correctly detected as closer to the lung wall (T-Wall) or the mediastinum (T-Med), and if the lymph node closer to the left (N-Left) or right (N-Right) lung. Comparing our method with only region-level features used, the accuracy of lymph nodes localization was much higher, because a much smaller number of lymph nodes were misclassified as tumors. And because the localization of lymph nodes also affected the bounding boxes of the lung and mediastinum, the localization for tumor near the mediastinum was also more accurate with our method.

Examples of the region detection results are shown in Figure 3. With the bag-of-words model, the high FDG uptake areas representing tumor and abnormal lymph nodes were detected. However, the tumor and abnormal lymph node could not be differentiated (shown as the same gray-scale value indicating the same cluster), and the tumors

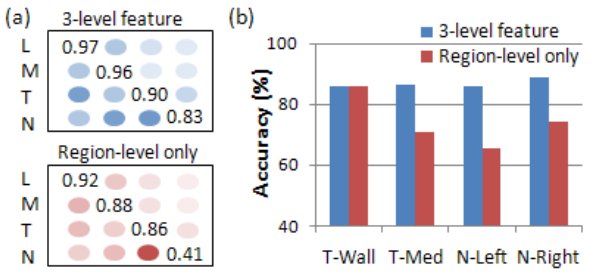

Figure 2. (a) L/M/T/N confusion matrices. (b) Accuracy of structure localization.

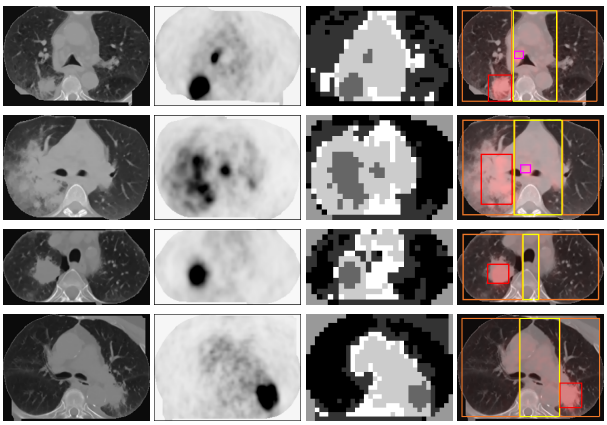

Figure 3. Each row shows CT and PET images from a single slice. Far left is the CT image; next is the PET image; next is the block-based bag-of-words representation; far right is our proposed region detection output, with a red box outlining the lung tumor, pink the abnormal lymph nodes, yellow depicting the mediastinum and orange outlining the lung fields, which are overlaid on the fused PET-CT image. Top two rows show a primary tumor and disease in lymph nodes and bottom two rows show just lung tumor.

were located in an area incorrectly clustered as mediastinum. With our proposed method, the tumor and lymph nodes were correctly distinguished, and the bounding box representations facilitated easy interpretation of the relative locations of the tumors and disease in lymph nodes.

3.3. Results on Image Retrieval

The image retrieval performance was evaluated by using each 3D image set as a query image, and the other 39 sets were ranked according to their similarity level with the query image. As shown in Figure 4, our method achieved higher retrieval precision comparing to the work [12] for recall levels up to about 70%. This validated our hypothesis that good retrieval performance could be achieved based on roughly located structures, without the need of boundary delineations that would be time consuming.

We also compared our method with three standard approaches: (i) bag of features (BOF), classifying 8 by 8 image patches into 6 clusters (identical to using only region-level features); (ii) bag of SIFT [8] features (BoSF), classifying the SIFT features into 32 clusters (the size empirically chosen); and (iii) spatial pyramid matching (SPM), with 3 levels (as suggested in the original paper [7]) and 6 clusters. The BOF method performed better than BoSF, which suggested that a dense feature grid was more suitable than keypoint representations for this dataset. As expected, SPM improved considerably over BoSF and BOF, since it modeled the spatial information. However, SPM descriptor was not translation invariant, which could cause large difference between images due to translation only. Our method resolved the translation problem with the structure localization step, thus resulted in higher retrieval precisions.

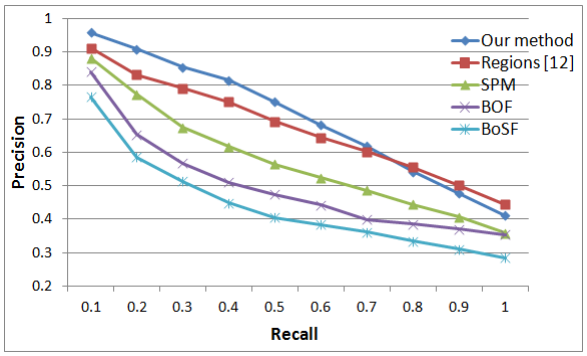

Figure 4. The retrieval precision and recall.

Figure 5 shows four retrieval examples to demonstrate the effectiveness of our retrieval method. Take the first example (shown in the first row) for a brief description: the query case contained a primary lung tumor and disease in lymph nodes; the first retrieved case presented a similar tumor and similar abnormal lymph nodes (not shown in the image slice); the second retrieved case exhibited less similar tumor and lymph node characteristics; and the third retrieved case depicted a tumor at the similar location but without nodal disease.

3.4. Computational Efficiency

The system was implemented in Matlab v2009b on a 2.66 GHz PC. For the structure localization, the total average time taken for one case (about 28 thoracic images) was 14 seconds, among which, about 13 seconds were needed for region-level feature calculation, and 1 second for region detection and case feature extraction. The retrieval stage took about 1 second for all 40 retrieval tests (with features computed previously).

4. Conclusions

We develop a new method for retrieving medical images that are similar in disease-specific contexts. PET-CT tho-

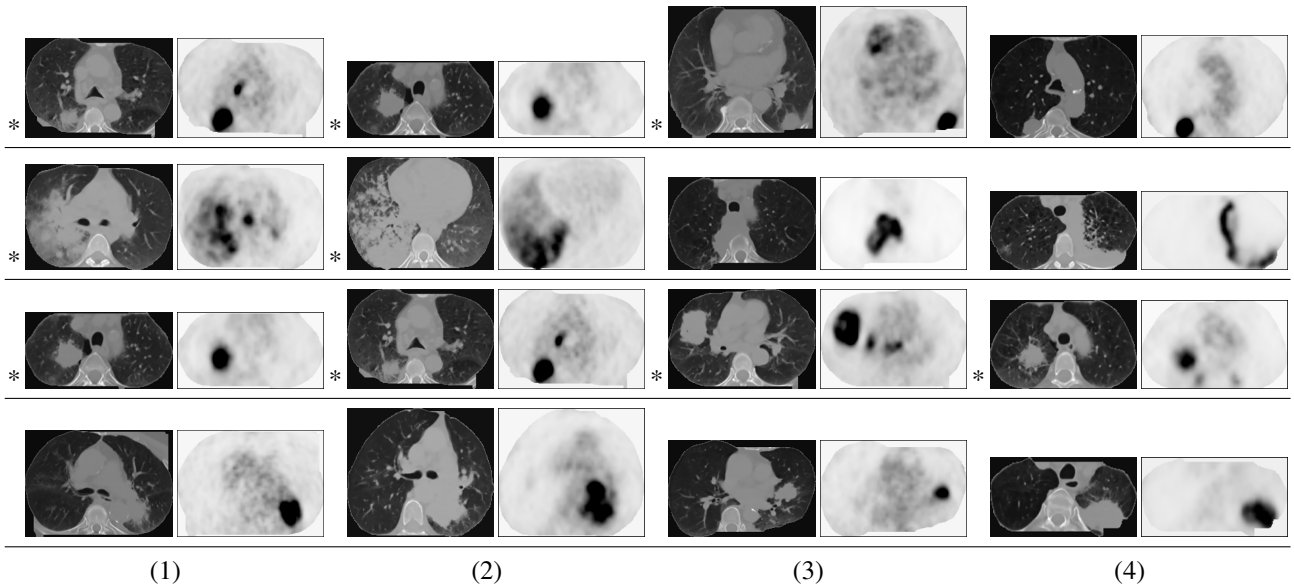

Figure 5. Each row shows one retrieval example. Column (1): the case in query. Column (2) to (4): retrieval results of the three most matching cases. For easy visualization, the PET-CT image depicting the center of tumor in the case is shown (showing the thorax only). The "*" leading a PET-CT image pair indicates the case has disease in lymph nodes, which may only be visible on a different image plane.

racic images from patients with NSCLC are used in this study. The contexts are detected by first localizing the primary lung tumor, abnormal lymph nodes, lung fields and mediastinum using a fast discriminative learning approach. Similar 3D image sets are then retrieved based on the contextual features with an optimized image similarity measure. Our evaluation on clinical data shows both highly effective structure localization, and higher retrieval precision. Our retrieval framework can also be adapted to other types of medical images with associated disease-specific contexts.

References

[1] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *TMI*, 30(3):733–746, 2011. 89

[2] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. *CVPR*, pages 129–136, 2010. 89

[3] R. Datteri, D. Raicu, and J. Furst. Local versus global texture analysis for lung nodule image retrieval. *SPIE*, page 691908, 2008. 90

[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *ICCV*, pages 229–236, 2009. 89

[5] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. *ICCV*, pages 1–8, 2007. 91

[6] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. *CVPR*, pages 113–120, 2010. 89

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *CVPR*, pages 2169–2178, 2006. 93

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 93

[9] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Int J. Medical Informatics*, 73:1–23, 2004. 89

[10] M. Rahman, S. Antani, and G. Thoma. A medical image retrieval framework in correlation enhanced visual concept feature space. *CBMS*, pages 1–4, 2009. 89

[11] K. Simonyan, A. Zisserman, and A. Criminisi. Immediate structured visual search for medical images. *MICCAI LNCS 2011*, 6893:288–296, 2011. 89

[12] Y. Song, W. Cai, S. Eberl, M. Fulham, and D. Feng. Thoracic image case retrieval with spatial and contextual information. *ISBI*, pages 1885–1888, 2011. 89, 90, 93

[13] L. Sorensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. Duin, and M. Bruijne. Image dissimilarity-based quantification of lung disease from CT. *MICCAI LNCS 2010*, 6361:37–44, 2010. 89

[14] 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*, 37(11):2165–2187, 2010. 90

[15] J. Zhang, S. Zhou, S. Brunke, C. Lowery, and D. Comaniciu. Detection and retrieval of cysts in joint ultrasound b-mode and elasticity breast images. *ISBI*, pages 173–176, 2010. 89