Hierarchical Spatial Matching for Medical Image Retrieval

Yang Song
BMIT Research Group
School of IT
University of Sydney, Australia
ysong@it.usyd.edu.au

Weidong Cai
BMIT Research Group
School of IT
University of Sydney, Australia
tomc@it.usyd.edu.au

David Dagan Feng
BMIT Research Group
School of IT
University of Sydney, Australia
feng@it.usyd.edu.au

ABSTRACT

Content-based medical image retrieval is likely becoming an important tool to provide valuable information to assist physician to make critical diagnosis decisions. While most existing works perform the retrieval based on low-level visual features, the pathological spatial context, which is critical for analysis of the disease characteristics, has been less studied. We thus aim to effectively extract and represent the spatial context of pathological tissues, and design a novel hierarchical spatial matching (HSM) method based on the spatial pyramid matching. Our method is able to (1) handle the translation variations of the main pathological object; (2) describe the spatial information surrounding the pathological object in an adaptive scale; and (3) compute image similarities with an optimally weighted distance function. The proposed method shows better retrieval performance comparing to the other widely used techniques.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation

Keywords

Hierarchical, spatial, discriminative matching

1. INTRODUCTION

Digital medical images are produced in ever increasing quantities and used for diagnosis and therapy. There is great interest for physicians to use imaging studies from previous patients as a reference guide to diagnose the current patient. A system to automatically retrieve similar images for a given one can thus be useful for assisting the diagnosis process, and recent works have shown some promising results [12, 4, 2]. Medical image similarities are normally computed by comparing their low-level visual features, such as texture and shape [9]. These low-level visual features can be effective in differentiating pathological tissues from normal structures, but only representing local regional features; while it is believed that disease characteristics can be better evaluated based on the spatial context that the pathological tissue is in. For example, to stage lung cancer, whether the tumor invades into its adjacent structures (e.g. mediastinum) is a critical factor.

Some recent works have thus incorporated spatial information in various ways, such as using the pixel coordinates as a feature vector [1], combining the feature words into a sliding window representation [10] or a circular histogram [13], and applying the spatial pyramid matching (SPM) [7] in medical images [3]. These approaches are normally built on the bag-of-words model. The spatial information is encoded by first dividing the image into a grid structure of image blocks, in coordinate, sliding window, circular space, or a multi-scale pyramid; and then the feature-word histograms of all blocks are concatenated into a large histogram as the feature descriptor of the image. A histogram intersection or Euclidean type of vector-based distance function is usually used to compute the differences between images.

These spatial descriptors, however, are often too rigid to handle spatial variations. For example, given two images containing two similar tumors, e.g. both tumors are of round shapes and near to the mediastinum, except one being in the left lung and the other in the right lung, the derived distance between the two images would be quite large. However, the two images should be considered similar, since the absolute locations of the pathological objects are of less diagnostic interests than their relative spatial relationships with adjacent anatomical structures. Furthermore, a commonly overlooked issue is that in the distance function, uniform or predefined feature weights are normally employed. However, it is often acknowledged that different features should carry different weights, and image blocks at more salient locations should be more important for measuring the image similarities than those at the background.

In light of the above, in this paper, we propose a hierarchical spatial matching (HSM) method for medical image retrieval, based on the concept of SPM, but with the following major differences: (i) a multi-level partition is originated from the most salient area (e.g. center of a tumor), to handle spatial translation variances between images respective to the main pathological objects; (ii) at one level, the partition blocks are of different sizes and concentric around the

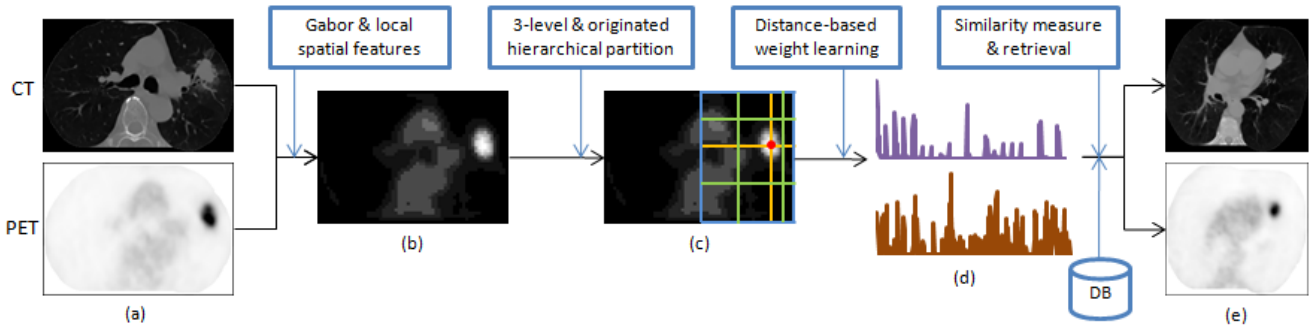

Figure 1: Illustration of our HSM method. (a) A PET-CT image transaxial slice pair (after preprocessing), with a primary lung tumor shown as high uptake values on PET (large black region). (b) Bag-of-words indicating feature words with grayscale values. (c) The hierarchical partition – blue lines depicting level-0 border, orange dividing into level-1 blocks, green further partitioning into level-2 blocks, and red dot showing the division origin. (d) The feature word distribution (upper) and the associated weights (lower), with feature index as the x axis. (e) A similar image pair retrieved by the query image pair in (a).

origin, to represent the spatial features at and surrounding the pathological area in an adaptive scale; and (iii) a discriminative learning-based approach is designed to assign a weight of each feature word for each block at each level, to compute a weighted similarity measure.

In this paper, we present our work using positron emission tomography – computed tomography (PET-CT) thoracic images from patients with non-small cell lung cancer (NSCLC). PET-CT imaging, which is widely used in lung cancer staging, produces co-registered anatomical (CT) and functional (PET) patient information from a single scanning session, creating a 3D image set comprising co-registered transaxial slice pairs. Since the characteristics of the primary lung tumor, especially its spatial context (e.g. adjacent to the lung wall, invasion into the mediastinum, etc), are important to stage NSCLC, the retrieved images are expected to exhibit such similar pathological patterns to the query image. To the best of our knowledge, the only existing work in this scope is our previous work [11], in which pathological regions are first detected to extract features of the detected regions, and hence the retrieval performance heavily depends on the detection accuracy. In this paper, we thus propose the HSM approach to avoid the stringent accuracy requirement on the pathological region detection.

2. HIERARCHICAL SPATIAL MATCHING

Figure 1 shows an overview of our proposed method. We measure the similarity between two images by (i) generating the bag-of-words representation of each image based on Gabor and local spatial features, (ii) computing a feature vector for each image with a 3-level and originated hierarchical partition, and (iii) calculating a learning-based weighted distance between the two feature vectors. Images having high degrees of matching with the query image are then retrieved.

It is worth mentioning that our method is designed to be easily generalizable to other anatomical domains and imaging modalities, especially that the local features used for bag-of-words generation (section 2.1) and the discriminative learning-based similarity measure (section 2.3) can be directly applied. Although the detailed structure of the hierarchical partition model (section 2.2) is optimized for representing the thoracic imaging characteristics, the overall concept is still applicable to other imaging studies where the relative spatial relationships between the pathological object and its surrounding structures are important, by first locating the pathological centroid, and then constructing a multi-level hierarchy around the centroid.

2.1 Bag-of-words Generation

In recent years, the bag-of-words model has become a very popular method for image classification and retrieval. It works by quantizing visual features extracted over image patches or keypoints into distinct categories of visual words, and using an orderless histogram to indicate the number of occurrences of each visual word in an image. In our method, we focused on designing suitable features to effectively differentiate various types of pathological and anatomical structures, such as tumors, lung fields and mediastinum.

To do this, we first preprocessed each PET-CT slice pair for every 3D image set to roughly extract the lung and mediastinum areas from the whole image slice, by removing the patient bed, other soft tissues and background using the Otsu thresholding and connected component analysis [11].

Then, 24 Gabor filters (with scale 6, orientation 4 and mask size 13×13) were applied on each PET-CT slice pair (24 Gabor filters processed for PET and 24 for CT), to extract the texture features. The PET-CT slice pair was then divided into a grid of 4×4 pixel patches, forming aligned grid pair of PET and CT. Each PET-CT patch could then be represented by 48 Gabor filtered images. To reduce the feature dimension, the 24 Gabor filtered PET (and then CT) images $\mathbf{G}$ were weighted combined into one image $\overline{\mathbf{G}}$:

$$\overline{\mathbf{G}} = \frac{\sum_{k=1}^{24} \sigma_k \mathbf{G}_k}{\sum_{k=1}^{24} \sigma_k} \quad (1)$$

where σ_k was the standard deviation of the kth filtered image. Based on such a combined PET (and then CT) image, the mean, minimum and maximum values were calculated. Subsequently, each PET-CT patch was represented by a 6-dimensional feature vector $\mathbf{f}$.

Next, to make the features more discriminative, we incorporated local spatial information into the feature vector, based on the local auto-correlation of feature vectors [6].

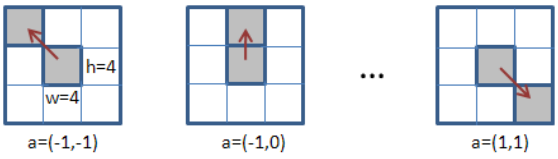

Figure 2: Illustration of the auto-correlation feature.

The first-order auto-correlation $\mathbf{g}(i, \mathbf{a})$ was obtained by:

$$\mathbf{g}(i, \mathbf{a}) = \mathbf{f}(i)\mathbf{f}(i + \mathbf{a})^T \quad (2)$$

where $\mathbf{f}(i)$ was the 6-dimensional feature vector of PET-CT patch i, $\mathbf{a}$ was the displacement vector to index the 8 immediate neighboring patches of patch i (Figure 2), and $\mathbf{g}(i, \mathbf{a})$ thus encoded the local spatial information around patch i. Each patch i was then represented by $\{\mathbf{f}, \mathbf{g}\}$, with a total dimension of 294 (i.e. $6 + 8 \times 6 \times 6$). We chose to use only the first-order auto-correlation because the feature dimension would increase exponentially with the increase in the order.

A k-means clustering algorithm was then performed on the grid of feature vectors to generate 16 categories of visual words. The number of clusters was chosen as 16 empirically. Each PET-CT patch was then assigned to the nearest visual word. The bag-of-words representation of the PET-CT image slice pair was thus a histogram counting the occurrence frequencies of all visual words.

2.2 Hierarchical Partition Model

Since the bag-of-words representation discards the spatial order of the local feature vectors, two images with similar visual word histograms but different structural information can not be differentiated. To describe the spatial features for thoracic images, here we designed a hierarchical partition model.

First, we introduced a division origin, which was the patch perceived as the most probably pathological. For this, we simply chose the patch having the highest mean standard uptake value (SUV) [14] in the PET image, as high uptake values normally indicated area of the primary tumor or metastases. Note that since we only needed to detect one single patch as the origin and not to detect the entire pathological area, this detection was essentially different from the previous tumor detection approach [11]. For normal images without pathology, the division origins were still detected. This was not a problem for matching because the distributions of visual words around the origins would be different between normal images and the ones with pathology.

We then created a 3-level hierarchical partition based on the division origin (Figure 3):

1) Level-0: A single large block represented level-0. Depending on whether the origin was in the left or right half of the image, the block would cover the left or right half of the image patches. We used only half of the image because if a tumor was present in the left lung, we wanted to retrieve images with similar patterns in either the left or right lung (together with the left or right side of the mediastinum), not to match the entire thoracic area. And rather than simply combining all visual words into one histogram, each visual word was weighted according to its distance to the origin to

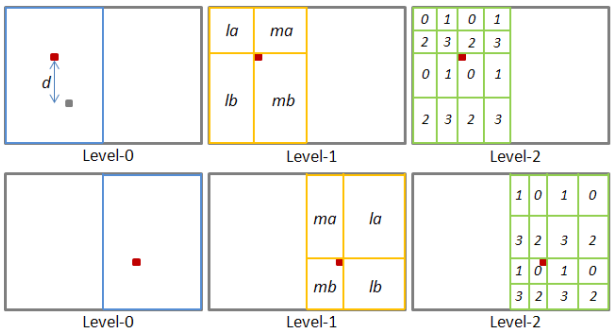

Figure 3: Illustration of the partition hierarchy. The upper row shows the case when the origin (i.e. the red block) is on the left, and the lower row for origin on the right. The upper level-0 image also shows an image patch (in gray) with its distance d to the origin.

generate the level-0 feature vector $\mathbf{H}_0$:

$$\mathbf{H}_0 = \sum_i exp(-d_i/\alpha)\mathbf{v}_i \quad (3)$$

where $\mathbf{v}_i$ was a vector of size M (16 in our case) with one non-zero element indicating the visual word of patch i, and d_i was the Chebyshev distance between patch i and the origin, and α was a constant factor (chosen as 10 empirically). In this way, the patches were considered as arranged in concentric circles around the origin, and the patches nearer to the origin would be more important for image matching.

2) Level-1: Four blocks were created by extending vertical and horizontal lines from the division origin. Such partitions were to capture the spatial information near to the mediastinum or the lung wall, and to differentiate the features at the anterior or posterior of the lung field. The feature vector of each block was constructed in the same way as the level-0 features, where α was chosen as 20 so that the weight differences between patches within one block were smaller. Indexing the four blocks as *la* (the block near to the lung wall and above the origin), *ma* (the block near to the mediastinum and above the origin), *lb* (the block near to the lung wall and below the origin), and *mb* (the block near to the mediastinuma and below the origin), the level-1 feature $\mathbf{H}_1$ was then concatenated by:

$$\mathbf{H}_1 = \{\mathbf{h}_{la}, \mathbf{h}_{ma}, \mathbf{h}_{lb}, \mathbf{h}_{mb}\} \quad (4)$$

Note that if the origin was in the left half of the image, *la* and *lb* would be to the left of the origin, and *ma* and *mb* to the right of the origin. The positions of the two sets of blocks would be swapped if the origin was in the right half of the image. The four blocks could also be of different sizes, depending on the location of the division origin; this made the level-1 features more discriminative between images. The feature vector also limited the rotation invariance within each block not across blocks, since each block represented roughly different anatomical areas of the thorax.

3) Level-2: Each level-1 block was divided evenly into four blocks, hence totally 16 blocks were formed at level-2. The inner most four blocks adjacent to the origin represented the patches that were nearer to the pathological area,

while the outer blocks would contain more normal anatomical structures. The feature vector of each block was also computed in the same way as the level-0 features, where α was further increased to 30 so that the further blocks would contribute considerably to the similarity computation at level-2. The feature vectors of all blocks were then concatenated as the level-2 feature $\mathbf{H}_2$:

$$\mathbf{H}_2 = \{\mathbf{h}^0_{la}, \mathbf{h}^1_{la}, \mathbf{h}^2_{la}, \mathbf{h}^3_{la}, \mathbf{h}^0_{ma}, ..., \mathbf{h}^3_{mb}\} \quad (5)$$

The hierarchical feature vector $\mathbf{H}$ of a PET-CT image slice pair was thus a concatenation from all of three levels:

$$\mathbf{H} = \{\mathbf{H}_0, \mathbf{H}_1, \mathbf{H}_2\} \quad (6)$$

with a total dimension of $M+4\times M+16\times M$, with $M = 16$. Although we could extend the hierarchy with more levels, the feature dimension would increase exponentially. We also used only three levels to follow the convention used by SPM.

On a side note, since a patient's PET-CT image is actually a 3D image set, retrieval of 3D image sets based on their overall features is of practical interests as well. While the above hierarchical partition was performed for 2D PET-CT image slice pairs, the feature of a 3D image set $\mathbf{H}^{3D}$ was obtained as a weighted combination of all the comprising 2D images:

$$\mathbf{H}^{3D} = \frac{\sum_n p_n \mathbf{H}^n}{\sum_n p_n} \quad (7)$$

where n indexed the slice pairs of the 3D image set, and p_n was the mean SUV of the origin patch, so that image slices containing more prominent pathological areas would contribute more to the overall 3D features; and the feature dimension of $\mathbf{H}^{3D}$ was the same as the $\mathbf{H}$.

2.3 Discriminative Matching

Since the image feature $\mathbf{H}$ was concatenated from all block features without any weights, all blocks would be equally important for differentiating between images. However, intuitively, blocks nearer to the origin at the finer levels should carry higher weights; and within one block, some feature words would be more discriminative than the others. Therefore in our method, we designed a learning-based method based on the triplet-learning model [5] to compute optimal weights for image matching, and the weights were not only level-specific, but also specific to each feature word of each block in the partition hierarchy.

To do this, we first defined the distance between two image features $\mathbf{H}_I$ and $\mathbf{H}_J$ as $D_{I,J}$:

$$D_{I,J} = \langle \Omega \cdot \frac{|\mathbf{H}_I - \mathbf{H}_J|}{\mathbf{H}_I + \mathbf{H}_J} \rangle = \langle \Omega \cdot \mathbf{H}_{I,J} \rangle \quad (8)$$

where Ω was the weight vector of the same size as $\mathbf{H}$.

To compute Ω, the triplet learning method was adapted. Assuming three images I, J and K were given, where I and J were similar and I and K were dissimilar, then the distance between I and J would be less than between I and K: $D_{I,J} < D_{I,K}$. With Q such training triplets $T_q = \{I, J, K\}$, a large-margin optimization was then used to solve for Ω:

$$\begin{aligned} \text{argmin}_{\Omega,\xi\geq 0} \quad & \tfrac{1}{2}\|\Omega\|^2 + c\textstyle\sum_q \xi_q \\ s.t.\ \forall q: \quad & \langle \Omega \cdot (\mathbf{H}_{I,K} - \mathbf{H}_{I,J}) \rangle \geq 1 - \xi_q \end{aligned} \quad (9)$$

To construct the Q training samples T, we first created a $P \times P$ matrix $\mathbf{U}$: $U_{i,j} = 1$ if the images i and j were similar, and 0 otherwise. Because of the large possible combinations of triplets and the possibility of over-fitting, we chose only a subset of triplets based on the label $\mathbf{U}$, using a two-step process. First, we selected N pairs of similar images ($U_{i,j} = 1$) that were viewed as *very* similar. The reason of such selection was that, we noticed that among all pairs of similar images, some pairs exhibited less obvious similarities, and choosing such pairs could introduce more bias in the training. Second, for each I, J pair selected, if X_I images were indicated as dissimilar to I in $\mathbf{U}$, X_I triplets were then formed: $T(i) = \{I, J, dissimilar(I)\}, i = 1, ..., X_I$. This training sample size was significantly lower than the maximum possible size, and the selection process ensured that the training would focus on highly discriminative images.

To simplify the training process, the labeling $\mathbf{U}$ was constructed for 3D image sets. To compute Ω for 2D image slice pairs, we chose one image slice pair that exhibited obvious tumor area from each 3D image set to compute the feature vector $\mathbf{H}$. Then by applying the training sample selection and optimization, the weight Ω was computed for 2D distance measure. We also performed another training process by replacing the 2D feature vector with $\mathbf{H}^{3D}$, and the derived Ω thus represented the weights corresponding to the feature vectors of 3D image sets.

3. EXPERIMENTS AND RESULTS

3.1 Image Datasets

In this study, we collected 40 PET-CT patient studies (3D image sets) with a total of 1134 thoracic slice pairs from the Royal Prince Alfred Hospital, Sydney. Primary lung tumors were visible in all the 40 image sets, with a variety of patterns. For each patient study, the other 39 image sets were marked similar or dissimilar as the ground truth for retrieval, based on the appearance and spatial characteristics of tumors. The number of similar image sets for each set ranged from 1 to 11, with an average of 4.75.

3.2 Experimental Methods

We performed two types of image retrieval experiments:

1) Retrieval of 3D image sets: By using each 3D image set as a query image, the other 39 sets were ranked according to their similarity levels with the query image. The retrieval results were then compared with the ground truth to compute the average precision and recall.

2) Retrieval of 2D image slices: We selected 5 key images (i.e. 2D image slice pairs exhibiting tumors) from each 3D image set as the query test set. Then the ground truth was propagated from the 3D image set level to all 2D images – if two images belonged to two 3D sets that were marked similar, the two 2D images were considered similar as well. Each query image was then compared to all images of the other 39 3D image sets. The retrieval results were then measured against the propagated ground truth, and the average precisions were computed for the Top-3, -5 and -7 retrievals.

Besides evaluating the retrieval performance of our proposed HSM method, we also assessed the effects of its major components by: (i) omitting the local spatial features (auto-correlation) from the feature extraction for bag-of-words representation; (ii) using subsets of the three levels of spatial features other than all three levels; and (iii) using pre-defined level weights for the distance function.

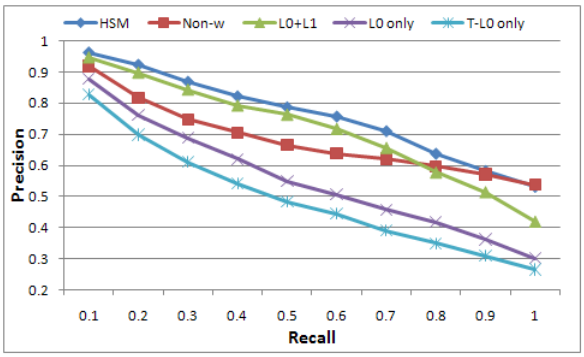

Figure 4: The retrieval precision-recall curves, evaluating various components – 3D image sets.

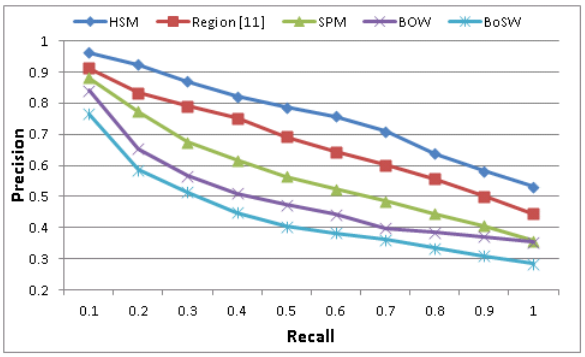

Figure 6: The retrieval precision-recall curves, comparing with the other methods – 3D image sets.

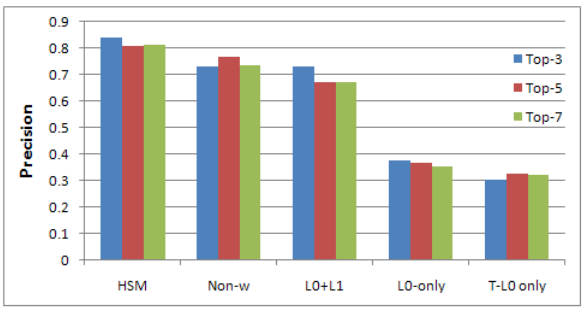

Figure 5: The average precisions of most similar retrievals, evaluating various components – 2D image slices.

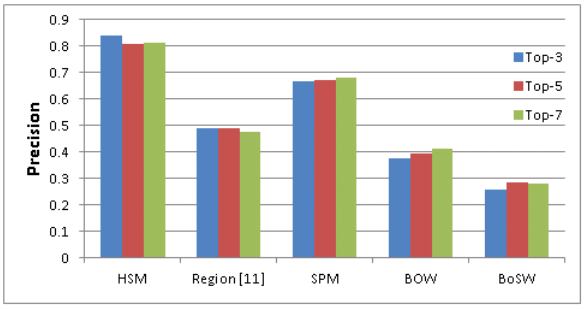

Figure 7: The average precisions of most similar retrievals, comparing with the other methods – 2D image slices.

We then compared our HSM method with the other approaches that were often used in image retrievals, including (i) the bag-of-words model with patch-based features – BOW, which did not include any spatial information; (ii) the bag-of-words model with the SIFT keypoints [8] – BoSW, which encoded the local spatial features; and (iii) the SPM method (built on BOW), which described the spatial information with a certain balance between subdividing and disordering. We also compared the proposed method with our previous work (region-based) [11] that reported good results on retrievals of PET-CT thoracic images.

3.3 Evaluation of Components

As shown in Figure 4, if the partition hierarchy was only of two levels – level-0 and level-1 only (*L0+L1*), the retrieval performance was slightly lower than HSM, suggesting the benefit of including the level-2 features. By comparing *L0+L1* with *L0-only* (level-0 features only), the advantage of introducing level-1 features can be clearly seen. We tested all the other combinations of the three levels of features, and using all three levels (*HSM*) resulted in the highest average precisions. We also tested further increasing to four levels, but the improvements were negligible.

Not surprisingly, with optimal feature weights, *HSM* achieved much higher precisions than without weighting for all feature words (*Non-w*). If the weights were assigned according to the hierarchy levels in the same way as SPM, the precisions were on average 1% higher than *Non-w* for recall levels from 0.1 to 0.4, but about 1% lower than *Non-w* for the rest of recall levels, which were thus all lower than *HSM*. It was also observed that when the recall level reached to 0.9 or 1, the effect of feature weights decreased. This was because our training procedure was tuned to match images that were very similar, not all degrees of similarity.

We also tested generating the bag-of-words representation based on the Gabor features only without the local auto-correlation. Using level-0 features computed from such bag-of-words outputs (*T-L0 only*), and comparing its precision readings with *L0-only*, the advantage of incorporating the local spatial features was prominent.

Similarly consistent results were also obtained for the 2D image slice tests, as shown in Figure 5. Notice that sometimes when the number of retrievals increased (e.g. for Top-3 to Top-5), the average precisions tended to increase, as more similar images were actually ranked lower than dissimilar images during the retrieval.

3.4 Comparison with The Other Methods

As shown in Figure 6, our HSM method outperformed all the other approaches. The BOW method performed better than BoSW, which suggested that a dense feature grid was more suitable than keypoint representations for this dataset. As expected, SPM improved considerably over BoSW and BOW, since it modeled the spatial information. The improvement of our method over SPM indicated the benefit of our partition hierarchy and the optimized feature weights.

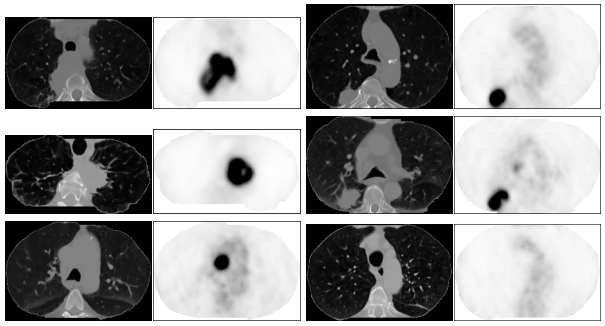

Figure 8: Example retrievals. Left two columns: Example 1. Right two columns: Example 2. Top row: the query image slice pairs (with primary lung tumors). Middle row: the most similar images retrieved using our HSM (exhibiting similar tumor characteristics to corresponding query images). Bottom row: the most similar images retrieved using SPM (not containing tumors, but with an abnormal lymph node in Example 1 possibly mistaken as a tumor).

HSM also achieved on average 8% higher precision than our previous work [11], which was encouraging that the new type of approach could extract the discriminative features well without accurately detecting the pathological regions.

Since our learning process for the feature weights can be easily incorporated into the BOW, BoSW and SPM methods, we also tested these three methods with optimized weights. The results showed that weighted SPM exhibited the largest improvement (7-16% for all recall levels), but still lower than our HSM method (4-8%). This test was mainly used to show the advantage of our hierarchical partition.

For 2D tests, SPM actually performed better than our previous work (Figure 7). This was because our previous method was designed for 3D features, and when retesting it for 2D image slices, not all techniques were actually utilized.

We also show two examples comparing the Top-1 retrieval outputs by our HSM method and the standard SPM approach, as in Figure 8. While our method was able to retrieve images that depicted similar tumor appearances and spatial contexts (both near to the mediastinum for Example 1, and near to the posterior lung field for Example 2), SPM retrieved images that were more similar in shape of the lung fields only.

Since our proposed HSM method was based on the bag-of-words model over 4×4 image patches, the computation was low-cost. On a 2.66 GHz PC with Matlab v2009b, processing of one PET-CT image slice pair took about 2 seconds.

4. CONCLUSIONS

We developed a novel matching method incorporating spatial contexts for medical image retrieval. Inspired by the popular SPM technique, and with the considerations of the special features presented in medical images, we designed the HSM method that could well retrieve images with similar pathological characteristics, especially the spatial features surrounding the pathological areas. Our evaluation on clinical data showed high retrieval precisions, and advantages over the other approaches. Our proposed method can also be easily adapted to other imaging domains, which we will investigate in further studies.

5. ACKNOWLEDGMENT

This work was supported in part by ARC and PolyU grants.

6. REFERENCES

[1] U. Avni, H. Greenspan, M. Sharon, E. Konen, and J. Goldberger. X-ray image categorization and retrieval using patch-based visual words representation. *ISBI*, pages 350–353, 2009.

[2] P. Bugatti, M. Silva, A. Traina, C. Traina, and P. Marques. Content-based retrieval of medical images: from context to perception. *CBMS*, pages 1–8, 2009.

[3] J. Feulner, S. Zhou, E. Angelopoulou, S. Seifert, A. Cavallaro, J. Hornegger, and D. Comaniciu. Comparing axial CT slices in quantized N-dimensional SURF descriptor space to estimate the visible body region. *Comput Med Imaging Graph*, 35(3):227–236, 2011.

[4] B. Fischer, A. Brosig, P. Welter, C. Grouls, R. Gunther, and T. Deserno. Content-based image retrieval applied to bone age assessment. *SPIE*, page 762412, 2010.

[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.

[6] T. Harada, H. Nakayama, and Y. Kuniyoshi. Improving local descriptors by embedding global and local spatial information. *ECCV*, pages 736–749, 2010.

[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.

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

[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 Med Inform*, 73:1–23, 2004.

[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.

[11] 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.

[12] 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*, pages 37–44, 2010.

[13] D. Unay, A. Ekin, and R. Jasinschi. Local structure-based region-of-interest retrieval in brain MR images. *IEEE Trans. Inf. Technol. Biomed.*, 14(4):897–903, 2010.

[14] H. Zaidi and I. Naqa. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *Eur J Nucl Med Mol Imaging*, 37:2165–2187, 2010.