Dictionary Pruning with Visual Word Significance for Medical Image Retrieval

Fan Zhang[a,b], Yang Song[a], Weidong Cai[a], Alexander G. Hauptmann[c], Sidong Liu[a], Sonia Pujol[b], Ron Kikinis[b], Michael J Fulham[d,e], David Dagan Feng[a,f], Mei Chen[g,h]

[a]*School of Information Technologies, University of Sydney, Australia*
[b]*Dept of Radiology, Brigham & Womens Hospital, Harvard Medical School, United States*
[c]*School of Computer Science, Carnegie Mellon University, United States*
[d]*Dept of PET and Nuclear Medicine, Royal Prince Alfred Hospital, Australia*
[e]*Sydney Medical School, University of Sydney, Australia*
[f]*Med-X Research Institute, Shanghai Jiaotong University, China*
[g]*Dept of Informatics, University of Albany State University of New York, United States*
[h]*Robotics Institute, Carnegie Mellon University, United States*

Abstract

Content-based medical image retrieval (CBMIR) is an active research area for disease diagnosis and treatment but it can be problematic given the small visual variations between anatomical structures. We propose a retrieval method based on a bag-of-visual-words (BoVW) to identify discriminative characteristics between different medical images with Pruned Dictionary based on Latent Semantic Topic description. We refer to this as the PD-LST retrieval. Our method has two main components. First, we calculate a topic-word significance value for each visual word given a certain latent topic to evaluate how the word is connected to this latent topic. The latent topics are learnt, based on the relationship between the images and words, and are employed to bridge the gap between low-level visual features and high-level semantics. These latent topics describe the images and words semantically and can thus facilitate more meaningful comparisons between the words. Second, we compute an overall-word significance value to evaluate the significance of a visual word within the entire dictionary. We designed an iterative ranking method to measure overall-word significance by considering the relationship between all latent topics and words. The words with higher values are considered meaningful with more significant discriminative power in differentiating medical images. We evaluated our method on two public medical imaging datasets and it showed improved retrieval accuracy and efficiency.

Keywords: Medical image retrieval, BoVW, Dictionary pruning

1. Introduction

Content-based medical image retrieval (CBMIR), which retrieves a subset of images that are visually similar to the query from a large image database, is the focus of intensive research (Müller et al., 2004; Cai et al., 2008; Akgül et al., 2011; Kumar et al., 2013). CBMIR provides the potential of having an efficient tool for disease diagnosis, by finding related pre-diagnosed cases and it can be used for disease treatment planning and management. In the past three decades, but in particular in the last decade, medical image data have expanded rapidly due to the pivotal role of imaging in patient management and the growing range of image modalities (Duncan and Ayache, 2000; Menze et al., 2014; Liu et al., 2015a,b). Traditional text-based retrieval, which manually indexes the images with alphanumerical keywords (Zhou et al., 2010), is unable to sufficiently meet the increased demand from this growth. At the same time, advances in computer-aided content-based medical image analysis systems mean that there are methods that can automatically extract the rich visual properties/features to characterize the images efficiently (El-Naqa et al., 2004; Lehmann et al., 2004; Napel et al., 2010; Avni et al., 2011; André et al., 2012a; Xu et al., 2012; Jiang et al., 2014; Zhang et al., 2014b, 2015d; Liu et al., 2015c; Song et al., 2015a,b).

In CBMIR research, the main challenge is to design an effective image representation so that images with visually similar anatomical structures are closely correlated. A number of research groups are working in this area (Müller et al., 2004; Zhang et al., 2010; Akgül et al., 2011; Cai et al., 2012; Hanbury et al., 2012; Kumar et al., 2013; Jiang et al., 2015a), and there is a trend to use a bag-of-visual-words (BoVW) for medical image representation (Castellani et al., 2010; Song et al., 2011c; Cruz-Roa et al., 2012; Kwitt et al., 2012; Foncubierta-Rodríguez et al., 2013; Liu et al., 2013a; Depeursinge et al., 2014; Zhang et al., 2015b). The BoVW model represents an image with a visual word frequency histogram that is obtained by assigning the local visual features to the closest visual words in the dictionary. Rather than matching the visual feature descriptors directly, BoVW retrieval approaches compare the images according to the visual words that are assumed to have higher discriminative power (Foncubierta-Rodríguez et al., 2012; Tamaki et al., 2013). The BoVW model was proposed by Sivic and Zisserman (Sivic and Zisserman, 2003) and has been adopted by many researchers in non-medical domains such as computer vision (Li and Pietro, 2005; Yang et al., 2007; Bosch et al., 2008), showing the advantages of describing local patterns over using global features only. This model has recently been applied to

tackle the large-scale medical image retrieval problem (Jiang et al., 2015b; Zhang et al., 2015e). In this study, we focus on a new BoVW-based retrieval for better retrieval accuracy and efficiency.

1.1. Related work

The aim of CBMIR is to extract visual characteristics of images to identify the level of similarity between two images. Feature extraction can be categorized into global-(GFM) and local-feature (LFM) models based on the scope of descriptors (Bannour et al., 2009). The GFM extracts a single feature vector from the whole image and the LFM partitions the image into a collection of smaller regions, namely patches, and considers that each patch has its own importance in describing the whole image (Avni et al., 2011). This patch-based model is particularly useful in medical image analysis since different image regions can represent the anatomical structures that play different and essential roles in medical imaging diagnosis (Tong et al., 2014; Zhang et al., 2013, 2014a).

The BoVW representation builds upon the LFM. Visually similar patches from different images are assigned to the same code in a codebook. Then, the patch-code co-occurrence assignment can be used to describe the image features and to compute the similarity between images. The workflow of BoVW-based image retrieval can be generalized into three steps (Caicedo et al., 2009): feature extraction, BoVW construction and similarity calculation. Specifically, the LFM is used to extract a collection of local patch features from each image. The entire patch feature set computed from all images in the database is then grouped into clusters, with each cluster regarded as a visual word and the whole cluster collection considered as the visual dictionary. Then, all patch features in one image are assigned to visual words, generating a visual word frequency histogram to represent this image. Finally, the similarity between images is computed based on these frequency histograms for retrieval.

In this workflow, an important issue is the dictionary construction. The visual word in the dictionary corresponds to a group of visually similar patches. Normally, these words are obtained within the local patch feature space using unsupervised clustering methods, e.g., *k*-means (André et al., 2011; Yang et al., 2012). These approaches often generate a redundant and noisy dictionary since they tend to accommodate all local patch feature patterns (Foncubierta-Rodríguez et al., 2013), thus reducing the effects of the most crucial words and increasing the computational cost. Hence, it is preferable to remove the visual words that are less essential for the BoVW representation.

To ensure that only the meaningful feature patterns are included, the supervised clustering method of Bilenko et al (Bilenko et al., 2004) can be used to regulate the construction of dictionary, but the method adaptability is limited because prior knowledge is required for the learning process. Another approach is to analyze the discriminative power of visual words (Caicedo et al., 2009), but the weighting scheme also requires supervised classifiers. Some researchers have suggested that the most frequent visual words in images are 'stop words', which occur widely but have little influence on differentiating images, and need to be removed from the dictionary (Sivic and Zisserman, 2003). Yang et al., however, showed that ranking the visual words based on their occurrences in the different images only was not sufficient to evaluate the importance of visual words (Yang et al., 2007). Term frequency-inverse document frequency (TF-IDF) (Jones, 1972) relies on the inverse frequency weighting and has demonstrated its benefits on visual word evaluation. Nevertheless, it merely utilizes the direct co-occurrence relationship between the images and visual words. Jiang et al. (Jiang et al., 2015b) proposed an unsupervised approach to refine the weights of visual words within the vocabulary tree and showed the advantages of using the correlations among the visual words. We suggest that this relationship can be further used to infer the semantic information and can provide a better description of the discriminative power of visual words.

The ultimate goal of CBMIR is to identify cases with similar clinical properties (Müller et al., 2004). Such similarity may not be accurately captured in the low-level visual features (Depeursinge et al., 2014). The BoVW model is also mostly restricted within the visual appearance scope since the images are represented by a collection of visual words (Yang et al., 2012). One approach to handle this limitation is to perform semantic feature extraction by inferring the high-level semantic information based on the low-level visual data. A number of researchers have shown the effectiveness of high-level feature description (Quellec et al., 2010; Song et al., 2011a,b; Batet et al., 2011; André et al., 2012b; Quddus and Basir, 2012; Kurtz et al., 2014a,b). It is important to emphasize that most of these approaches require additional information including manual annotation (André et al., 2012b), supervised learning (André et al., 2012b; Quddus and Basir, 2012) and biomedical ontological knowledge (Batet et al., 2011; Kurtz et al., 2014a,b).

The latent semantic topic model (LSTM) (Li and Pietro, 2005; Bosch et al., 2008) can be used to automatically extract semantic information and it has been recently introduced into medical image analysis (Castellani et al., 2010; Cruz-Roa et al., 2012; Kwitt et al., 2012; Foncubierta-Rodríguez et al., 2013). Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2001) is one of the more popular latent topic techniques. pLSA is a language modeling technique and it is widely used in document analysis. The underlying idea is that each document can be considered as a mixture of latent topics. The latent topic is a probability distribution of words, and can be inferred from the co-occurrence relationship between documents and words, i.e., the latent topics. It has been used to extract the semantic relationship of morphological abnormalities on the brain surface (Castellani et al., 2010) and model histological slides to construct similarities between images (Cruz-Roa et al., 2012). pLSA is also employed to identify the meaningful visual words for BoVW based on the latent topics (Foncubierta-Rodríguez et al., 2013). The words with conditional probabilities below a significance threshold are regarded meaningless and removed from the visual dictionary. Since the conditional probabilities only describe the individual words, this method does not consider the relationship among the words. It also assumes that all

latent topics can be treated equally in the evaluation of significant words but this is controversial, and so, this work reported by Foncubierta-Rodríguez et al (Foncubierta-Rodríguez et al., 2013) has not resulted in clear improvements in retrieval accuracy.

1.2. Contributions

We propose a BoVW-based medical image retrieval method with a Pruned Dictionary based on the Latent Semantic Topic description, which we refer to as PD-LST retrieval. Our goal is to measure the discriminative power of a visual word in the dictionary so that less meaningful words are removed to enable better similarity computation between images. This discriminative power is quantitatively measured by a ranking metric, which we define as the significance value. Our method has two main contributions: a topic-word significance computing with pLSA topic extraction and an overall-word significance computing with a ranking approach. For the topic-word significance, we compute a significance value for a word relative to a certain latent topic. A pLSA method is applied to extract the latent topics between images and words, and the learnt conditional probability of a word given a latent topic is then adopted to quantitatively measure the topic-word significance. For the overall-word significance, we calculate a final significance value for each word. We designed a ranking method to incorporate the overall relationship between all latent topics and words. While the topic-word significance is used to describe a word's individual significance, the overall-word significance is used to evaluate the word's discriminative power in the entire dictionary.

The benefits of this pruning are: a) The updated BoVW representation can better capture the similarity level between images so that it can obtain higher retrieval accuracy. b) Our PD-LST method can largely reduce the amount of required words, leading to higher retrieval efficiency. We evaluated our method on two publicly available datasets - the Early Lung Cancer Action Program (ELCAP) (ELCAP and VIA, 2003) and Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008). Our prior work (Zhang et al., 2015a) showed the effectiveness of the dictionary pruning-based analysis and reported some preliminary results. In this work, we elaborate the topic-word and overall-word significance computation process with further details. The ranking method is justified by a mathematical explanation. We extend the evaluation to the ADNI dataset for brain image retrieval task, in addition to the originally used ELCAP dataset, to demonstrate the general applicability of our method. More comprehensive performance comparison with various approaches are performed on the two datasets. We also compared the execution time for efficiency analysis.

The structure of this paper is as follows: in Section 2 we describe the two stages of the proposed PD-LST method; in Section 3 we introduce the experimental datasets and experimental design; in Section 4 we present the experimental results and discussion, and we provide a conclusion and an outline of future work in Section 5.

2. Methods

2.1. Overview of the PD-LST retrieval

The outline of our PD-LST method is shown in Fig.1(a). The left part shows the standard BoVW workflow (Section 1.1). A dictionary of size M is generated from the extracted low-level features using k-means. The word frequency histograms of images are then calculated and used to compare the image similarity with Euclidean distance for retrieval. In addition to the standard BoVW model, PD-LST incorporates a dictionary pruning stage to remove the less meaningful words, i.e., the ones with limited discriminative power, as illustrated in the right part of Fig.1(a). A pLSA method is employed to extract the latent semantic topics based on the image-word co-occurrence relationship, and the learnt conditional probability of a word given the latent topics is adopted to measure its individual significance (Section 2.2). A ranking algorithm is designed to update the significance value of the words by incorporating the overall relationship among all latent topics and words, and calculate the final significance of each word (Section 2.3). The words with lower overall-word significance are removed to prune the dictionary. The similarity between images is then calculated based on the new frequency histograms using the pruned dictionary, followed by a k-NN for retrieval (Section 2.4).

Fig. 1(b) gives the visual illustration of our PD-LST method. The underlying idea of our method is that the visual dictionary used for constructing the BoVW model can be very noisy and redundant, reducing the representative and discriminative power of the visual words in identifying similar images. For example, the patches from the pleural surface of lung nodule image in Fig. 1 (b) (left sample) are visually different, making the local features of these regions assigned to different visual words, i.e., w_2, w_3 and w_4, and causing confusions in finding similar images based on the visual word distributions. Building upon the aforementioned work of Foncubierta-Rodríguez et al (Foncubierta-Rodríguez et al., 2013), we propose a new way to prune the dictionary considering the overall relationship between latent topics and visual words. We hypothesize that such a design would perform well because with the help of latent topics, the relationship between images is captured in terms of semantic descriptions, instead of the visual appearance. In this example, w_2, w_3 and w_4 are connected to the first latent topic representing the pleural surface. Then w_2 and w_3 would be removed since they don't present the co-occurrence information between the two images, and the corresponding patches would be assigned to w_4. In this way, the Euclidean distance between the two images from the same category is smaller after the dictionary pruning process.

2.2. Topic-word significance

The topic-word significance describes the significance of a word based on the latent semantic topics inferred from the relationship between the images and words. These latent topics provide the semantic description to bridge the gap between low-level visual features and high-level semantics. While the words are considered as the visual content pattern obtained from the visually similar patches, the latent topics are regarded as the

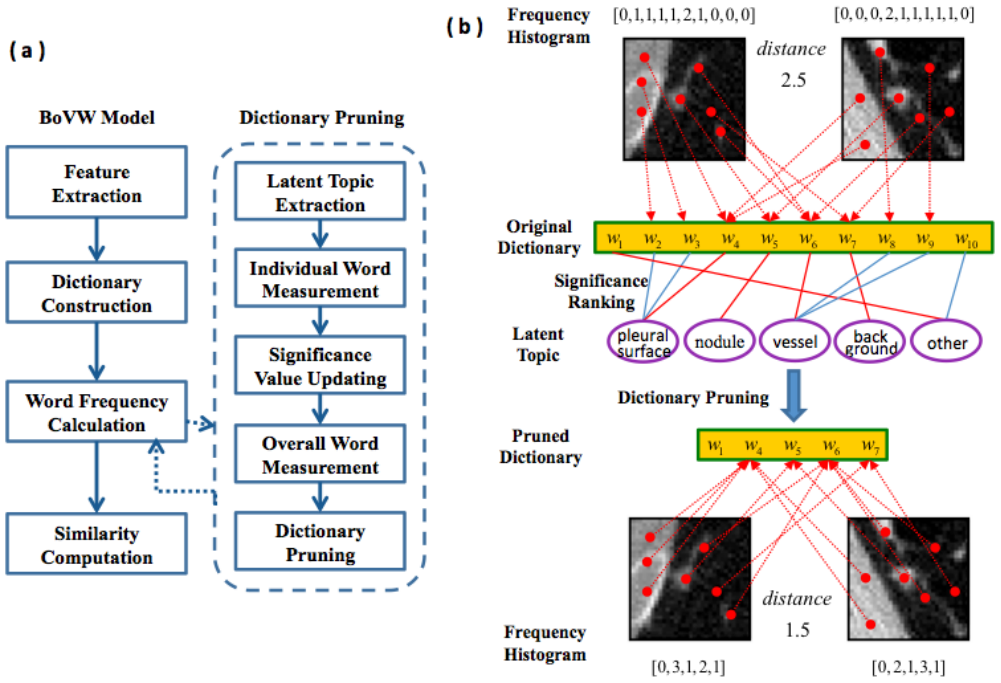

Figure 1: Outline (a) and visual illustration (b) of the proposed PD-LST retrieval framework. The samples are lung nodule images of type J (please see Section 3.1 for details).

pattern categories (Bosch et al., 2008). For example, in images to evaluate lung nodules, the latent topics can be used to describe the pulmonary structure that the patches belong to. An image that contains multiple instances of these patterns is modeled as a mixture of latent topics. A latent topic that describes the common characteristic of the content patterns is modeled as a mixture of words. The words are thus linked to the latent topics rather than directly to the images.

For this study, we used pLSA to extract the latent topics. In pLSA, the similar words tend to have high conditional probabilities given the same latent topic. The anatomical structures represented by the visual words are thus correlated indirectly with the latent topics. In an unsupevised manner, we do not need to explicitly specify these correlations, making our method more adaptive to different imaging problems, e.g., lung nodule and brain images. Fig.2 shows the flow of topic-word significance measurement using pLSA. A visual word is connected to a latent topic with a conditional probability that can quantitatively evaluate the significance of the word regarding this certain latent topic, i.e., the topic-word significance.

Formally, an image-word co-occurrence matrix $OCC_{M\times N}$ is computed by assigning all local patch features in an image to the visual words, where the element $occ(w_i, I_j)$ refers to the number of occurrences of word w_i with $i \in [1, M]$ in image I_j with $j \in [1, N]$, in which N is the number of images and M is the size of the dictionary. pLSA considers that the observed probability of a word w_i occurring in a given image I_j can be expressed with a latent or unobserved set of latent topics $Z = \{z_h : h \in [1, H]\}$ where H is a constant parameter as the

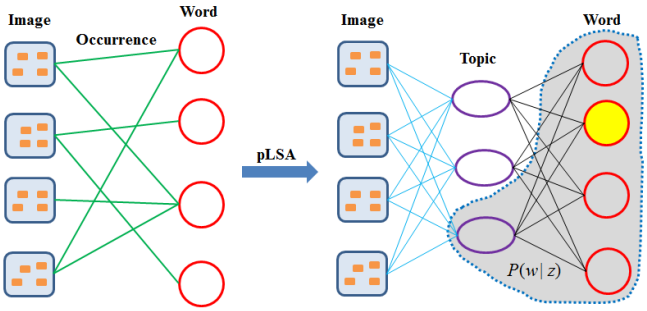

Figure 2: Outline of topic-word significance measurement with pLSA. The occurrence relationship (green lines) between the images (large blue rectangles) and visual words (small orange rectangles) is computed. pLSA is then used to extract the latent topics (purple ellipses). The conditional probability $P(w_i|z_h)$ (black line) of word w_i given the latent topic z_h is learnt and used to measure the significance of the word.The word with the higher probability (yellow) has the higher topic-word significance regarding this latent topic (shadowed).

number of latent topics, as:

$$P(w_i|I_j) = \sum_h P(w_i|z_h) \cdot P(z_h|I_j) \tag{1}$$

The probability $P(w_i|z_h)$ describes how word w_i is linked to latent topic z_h, with higher value of $P(w_i|z_h)$ indicating it is more connected. The latent topics Z can be learnt by fitting the model with Expectation-Maximization (EM) (Hofmann, 2001) algorithm that maximizes the likelihood function L:

$$L = \prod_i \prod_j P(w_i|I_j)^{occ(w_i,I_j)} \tag{2}$$

A total of H latent topics and the conditional probabilities of all words given these latent topics are learnt using pLSA. We consider that a visual word is more meaningful / discriminative if it is connected to the important latent topics (Section 2.3). The conditional probability is thus adopted as the topic-word significance to measure the closeness of a word relative to a certain latent topic for the overall-word significance computation.

2.3. Overall-word significance

With the obtained topic-word significance, the simplest dictionary pruning approach would be to keep the words with high conditional probabilities for all latent topics. Such an approach is based on the assumption that all latent topics can be treated equally, which is not appropriate for practical application. Taking lung nodule images as an example (Section 3.1), the latent topics are regarded as the local content pattern categories, which represent different types of anatomical structures such as the nodule, vessel, pleural surface or background. However, these structures do not have the same importance in determining the pathological categories of lung nodule images. A word might have high topic-word significance for certain latent topics, but it would be less significant compared to the other words if the connected latent topics are unimportant. For example, some stop-words that describe the background regions in lung nodule images tend to have high conditional probabilities with the unimportant latent topics that imply the 'background'. Thus, we wanted to compute the contribution of the latent topics for measuring the overall significance of words, i.e., the overall-word significance.

We designed a ranking-based method, based on the relationship between the latent topics and words, to derive their contributions and significances. Suppose we have some latent topics that make high contributions, then the word that is strongly connected with these latent topics will have a higher significance value. Similarly, if many high-significance words are strongly connected to a certain latent topic, it reflects that this latent topic will make a high contribution. The proposed ranking metric is based on this relationship to compute the significance of words and contribution of latent topics conditioned on each other. Fig.3 shows the flow of the overall-word significance measurement.

Firstly, a higher topic-word significance means the word is more closely linked to this latent topic, and thus is used to describe the 'strongly connected' relationship between the latent topic and word. Specifically, a topic-word threshold *twth* is used so that only the words with higher topic-word significance are regarded as strongly connected with the latent topic. Thus, the relationship between the latent topics and words is represented with a bipartite graph B, as:

$$\begin{cases} B(z_h, w_i) & = 1, \textbf{if } w_i \text{ has } P(w_i|z_h) \text{ ranked top } twth \\ B(z_h, w_i) & = 0, \textbf{otherwise} \end{cases} \tag{3}$$

where *twth* is a percentage such that the top *twth*, e.g., 10%, words with higher $P(w_i|z_h)$ are kept for the latent topic z_h. In this way, we can have the same number of words connected to each latent topic (we will discuss this after introducing the ranking method). With the connections defined in the bipartite graph B, the relationship between the latent topics and words for significance and contribution computation can be explained as follows: the significance value $s(w_i)$ of a word w_i is approximated from the contributions of the connected latent topics, and the contribution value $c(z_h)$ of a latent topic z_h is approximated based on the significance of the connected words. We define the values as:

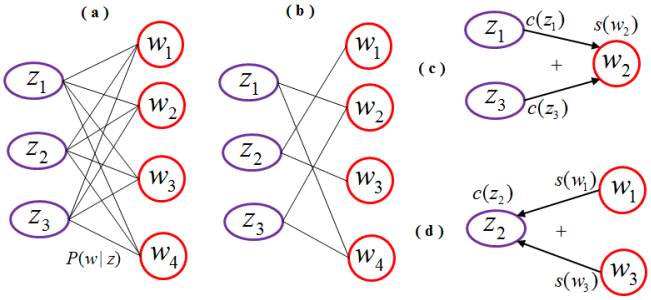

Figure 3: Outline of the overall-word significance measurement: (a) The relationship between the latent topics and words in terms of the conditional probabilities; (b) Bipartite relationship between the latent topics and words; (c) Computation of significance score of word w_2, i.e., $s(w_2) = c(z_1) + c(z_3)$; and (d) Computation of contribution score of latent topic z_2, i.e., $c(z_2) = s(w_1) + s(w_3)$.

$$s(w_i) = \sum_{z_h : B(z_h, w_i)=1} c(z_h) \tag{4}$$

$$c(z_h) = \sum_{w_i : B(z_h, w_i)=1} s(w_i) \tag{5}$$

Eqs.(4) and (5) can be alternatively solved iteratively to calculate the final overall-word significance, as shown in Algorithm 1. Supposing the significance value of all words is denoted with a vector $S \in \mathbf{R}^{M\times 1}$ and the contribution value of all latent topics is represented with a vector $C \in \mathbf{R}^{H\times 1}$, both the significance value vector $S = \{s(w_i) : i \in [1, M]\}$ and contribution value vector $C = \{c(z_h) : h \in [1, H]\}$ are initialized with 1, i.e., $s_0(w_i) = c_0(z_h) = 1$. At each iteration $t \in [1, T]$, the significance value $s_t(w_i)$ is updated with Eq.(4) and then the contribution value $c_t(z_h)$ is updated with Eq.(5). The two vectors are then $L2$ normalized so their squares sum to 1 at the end of each iteration, as:

$$s_t(w_i) = s'_t(w_i), \ \textbf{s.t.} \sum_{i\in[1,M]} s'_t(w_i)^2 = 1 \tag{6}$$

$$c_t(z_h) = c'_t(z_h), \ \textbf{s.t.} \sum_{h\in[1,H]} c'_t(z_h)^2 = 1 \tag{7}$$

The significance value of a word w_i at the final iteration T is thus the desired overall-word significance.

The ranking algorithm updates the significance and contribution values iteratively. At the beginning, we have the same number of words connected to each latent topic (same *twth* for all latent topics) and initialize the same contribution and significance values ($s_0 = 1$ and $c_0 = 1$) for all latent topics and words.

Algorithm 1 Pseudo code of the iterative ranking algorithm.

Input: Number of iterations T, bipartite graph B.
Output: Overall-word significance value vector S_T.

initialize $s_0(w_i) = 1$ and $c_0(z_h) = 1$.
for each t in $[1, T]$ **do**
  for each i in $[1, M]$ **do**
    Compute $s_t(w_i)$ based on $c_{t-1}(z_h)$ using Eq.(4);
  end for;
  for each h in $[1, H]$ **do**
    Compute $c_t(z_h)$ based on $s_t(w_i)$ using Eq.(5);
  end for;
  normalize $s_t(w_i)$ and $c_t(z_h)$ with Eqs.(6) and (7);
end for;
return S_T.

In this way, without the prior knowledge on the contribution of latent topics and discriminative power of words, we can treat all latent topics and words without any bias at the beginning of the ranking method. Then, within each iteration, the significance of a word is computed according to the most related latent topics and the shared knowledge between the latent topics and words is incorporated. Across the iterations, the significance of a certain word is diffused to the latent topics at the current iteration and gathered at the next iteration for updating the other words so that the relationship between the words is also used. Thus, the overall-word significance is derived based on the words and latent topics collectively.

The algorithm can be formulated alternatively as follows. With the bipartite graph $B \in \mathbf{R}^{M \times H}$ that indicates the adjacent matrix between all latent topics and words, Eq.(4) for word significance updating and Eq.(5) for latent topic contribution updating can be expressed as:

$$S = B \cdot C \tag{8}$$

$$C = B^{\top} \cdot S \tag{9}$$

Given the iteration $t \in [1, T]$, the sequence of the significance vectors $\{S_t\}$ can be expressed as:

$$\begin{aligned} S_t &= B \cdot C_t = B \cdot (B^{\top} \cdot S_{t-1}) \\ &= (B \cdot B^{\top})^2 \cdot S_{t-2} = ... \\ &= (B \cdot B^{\top})^t \cdot S_0 \end{aligned} \tag{10}$$

With the normalization in Eq.(6), S_t is the unit $L2$-norm vector in the direction of $(BB^{\top})^t S_0$ (similarly to C_t). As reported by Golub, Van Loan and Wilkinson, the unit $L2$-norm vector sequence of $\{S_1 ... S_t\}$ converges to a limit S^* as t increases arbitrarily, and so does the sequence of $\{C_1 ... C_t\}$ (Wilkinson, 1965; Golub and Van Loan, 2012).

The above explanation illustrates that our ranking method generates a convergent ranking result and the significance and contribution values can be estimated approximately with the principal eigenvectors of $BB^{\top}$ and $B^{\top}B$. This provides an alternative to compute the significance values. However, through the experiments, we observed that the retrieval performance tends to be stable with a relatively small number of iterations (Section 4.1). We can thus obtain the final ranking order for the retrieval without achieving the converged ranking values. This can also be helpful to improve the efficiency if there are a large amount of image data. In addition, the proposed iterative method represents that the significance of words and contribution of latent topics are computed based on each other and the final ranking is obtained from the overall perspective

2.4. PD-LST retrieval using the pruned dictionary

The dictionary is pruned according to the overall-word significance in the final step. All words within the dictionary are ranked, and the ones below a percentage point, namely the pruning percentage p, are considered meaningless and are removed, leading to a pruned dictionary with the size of $p \times M$. Then, the standard BoVW retrieval is conducted on the pruned dictionary. The co-occurrence matrix of the images is reconstructed by computing the new visual word frequency histograms on the pruned visual dictionary. Euclidean distance similarity is employed to calculate the similarity between images and k-NN method is used for retrieval.

3. Dataset and experimental design

We employed two publicly available medical imaging datasets, i.e., the ELCAP (ELCAP and VIA, 2003) and ADNI databases (Jack et al., 2008), for experimental evaluations.

3.1. Datasets

The ELCAP database contains 50 sets of low-dose computed tomography (LDCT) human lung scans, with the lung nodules annotated at the centroid. In our study, a set of 379 lung nodule images were used for evaluation. Lung nodules are small masses in the lung and can be divided into four different categories based on their location and connection with the surrounding structures such as vessels and the pleural surface (Diciotti et al., 2008), as follows: well-circumscribed (W), vascularized (V), juxta-pleural (J) and pleural-tail (P), as shown in Fig.4. The numbers of nodules for the four types are 57 (W), 60 (V), 114 (J), and 148 (P) respectively. The ADNI database comprises 331 subjects with magnetic resonance (MR) and positron emission tomography (PET) scans with a diagnosis of cognitively normal, mild cognitive impairment (MCI) and Alzheimer's Disease (AD). Examples are shown in Fig.5. We segmented each brain scan into 83 functional regions. The risk of progression to dementia is higher if more regions display glucose hypometabolism (Liu et al., 2013c). The numbers of subjects for the three stages are 77 (normal), 169 (MCI) and 85 (AD) respectively.

The literatures suggest that identifying the location information of lung nodules is essential for the early detection of lung cancer and determining the neurodegenerative progression stages is helpful for finding the patients at a high risk of dementia (Liu et al., 2014a, 2013b). Finding a list of related cases is of high clinical interest for the disease diagnosis and treatment. Therefore, in this study for the ELCAP dataset, we tried to retrieve the lung nodules at similar locations relative to the

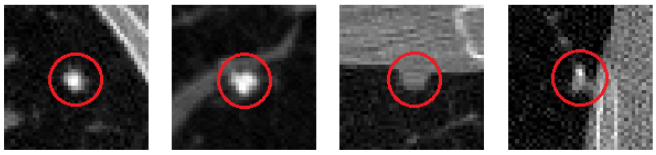

Figure 4: Transaxial CT images with typical nodules (from left to right) - well-circumscribed (W), vascularized (V), juxta-pleural (J) and pleural-tail (P). The nodules are circled in red.

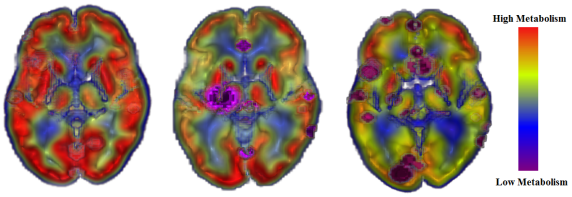

Figure 5: Lesion patterns for the three stages, shown from left to right as cognitively normal, MCI and AD. Red indicates high metabolism and blue color indicates low metabolism. Images were generated using 3D Slicer (Version 4.3) (Fedorov et al., 2012).

surrounding pulmonary structures as W, V, J and P, and for the ADNI dataset, we aimed to retrieve images with similar neurodegenerative progression patterns as AD, MCI and the cognitive normal.

3.2. Feature extraction and dictionary construction

In the ELCAP database the lung nodules are small and have an average size of 4×4 pixels (approximately from 3×3 to 7×7 pixels) across the centroid in the axial direction. Therefore, to restrict the problem scope to lung nodule analysis, an ROI of 33×33 pixels was cropped from each image slice with the annotated nodule centroid appearing in the center, similar to the processing in some related works for lung nodule analysis (Wu et al., 2010; Farag et al., 2010; Farag, 2013). We conducted a pixel-by-pixel patch feature extraction process to build the LF representation for the nodule and surrounding pulmonary structures. For each pixel around the annotated centroid (including the centroid pixel) as a keypoint, we computed a scale invariant feature transform (SIFT) (Lowe, 1999) descriptor using the VLfeat library[1], with the parameter $frames = [x, y, s = 4, o = 0]$, where x and y indicate the pixel position, s is the scale and o is the orientation. A 128-dimension vector was obtained for each frame and used as a local patch feature. Based on our previous work (Zhang et al., 2014a,b), incorporating too many surrounding pulmonary structures, e.g., including extra pleural surface, or too few, e.g., excluding the essential vessels, reduces the performance of recognizing the nodule type. Therefore, a total of 100 patch features were used by selecting the SIFT descriptors from the nearest 100 pixels around the nodule centroid.

For the ADNI dataset, the MR and PET data were preprocessed following the ADNI image correction protocols and nonlinearly registered to the ICBM_152 template to segment the entire brain into 83 functional regions (Liu et al., 2013c). Then, for each subject, we extracted 8 features. Each feature was an 83-dimension vector where each element described one of the 83 functional regions. The mean (Cai et al., 2010) and Fisher (Liu et al., 2011) indices, and difference-of-Gaussian-based (DoG area, DoG contrast, DoG mean) features (Toews et al., 2010; Cai et al., 2014a) were extracted from the PET data, and solidity, convexity (Batchelor et al., 2002; Liu et al., 2014b) and volume (Heckemann et al., 2011) were extracted from the MR data. Thus, we obtained an 8-dimension vector for each functional region as one local patch feature, and 83 feature vectors for each subject to construct the LFM. The overall statistics of the two datasets are shown in Table 1.

Table 1: Overall feature statistics of the two datasets.

Datasets	Patch feature length	Number of patches per case	Number of total cases
ELCAP	128	100	379
ADNI	8	83	331

3.3. Experimental design

In our study, leave-one-case-out cross-validation was conducted by using each case as query and the remaining cases in the dataset as the retrieval candidates. In this way, we can provide a comprehensive comparison by enabling the similarity computation between every two cases in the dataset. During the experiments, we had the same parameter setting for all testing queries. Therefore, the optimal values of the parameters did not result in biases with the leave-one-case-out cross-validation. All images in the dataset were included for, e.g., dictionary construction, latent topic extraction, word significance computing and dictionary pruning, due to the unsupervised nature of all comparison methods involved. With such experimental design, we could better utilize the image information including the testing images. It is worth noting that the class label information was not involved in these steps but only for the accuracy computation.

The most related items were retrieved for a given query as the retrieval results with an output number K. The performance was measured using the average retrieval accuracy (i.e., retrieval precision) of N queries, as,

$$Retrieval\ Accuracy = (\sum_{l \in [1,N]} (TP_{Q_l}/K))/N \tag{11}$$

where TP is the number of true positive items within the K retrieved results for the query image Q_l with l indicating the index of the query Q. The retrieved item is true positive if it is within the same class with the query image.

4. Experimental results and discussion

4.1. Parameter analysis

Our method has four major parameters: the number of latent topics H, topic-word threshold $twth$, number of iterations

[1]From VLfeat project, downloaded at: http://www.vlfeat.org/index.html

T, and the pruning percentage p. We have conducted the experiments given various dictionaries (M was from 100 to 2000) and outputs (K was from 1 to 10). In the following paragraphs, we provide the retrieval results from the 1-output for the 500-dictionary to show the effects of the four parameters.

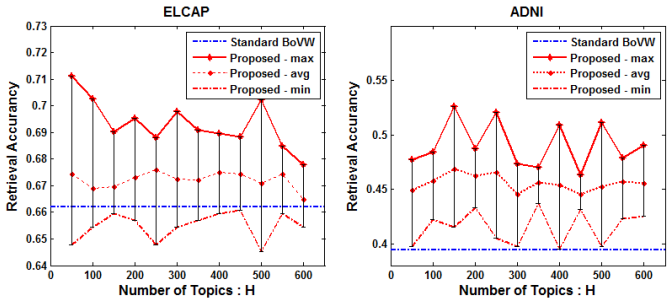

Figure 6: The retrieval accuracy curve given the number of latent topics.

Fig. 6 displays the accuracy curves given different numbers of latent topics on the two datasets. For each H (50 to 600), the maximum, minimum and average accuracies across different *twth*s (0.05 to 1) and *p*s (10 to 90) with T fixed at 20 are reported. While the curves on the ADNI dataset fluctuated more than with the ELCAP dataset, the accuracies of these two datasets were stable, in particular, for the average accuracy curves. This suggested that the number of latent topic had a limited impact on retrieval accuracy, due to the fact that only the latent topics making greater contributions affected the dictionary pruning.

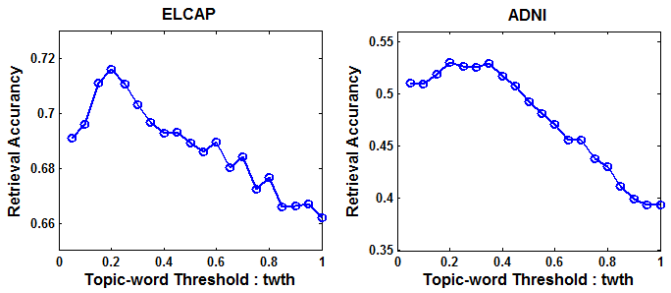

Figure 7: The retrieval accuracy curve given the topic-word thresholds.

Given different topic-word threshold *twth*s, Fig. 7 shows that if too few or too many words were kept to construct the bipartite graph between the latent topics and words there was lower performance. This finding was because too few words led to the loss of important knowledge and too many produced noise. In general, keeping relatively few words for each topic (10% and 30% for the ELCAP; 20% to 40% for the ADNI) generated better performance.

As shown in Fig. 8, the retrieval accuracy gradually increased with larger T values and then stayed constant after T reached a certain value. Less iterations were needed for stable retrieval results with a smaller dictionary and a larger number of outputs. The smaller dictionary led to a fewer connections between the latent topics and words and the larger number of outputs made it easier to include the most related items into the

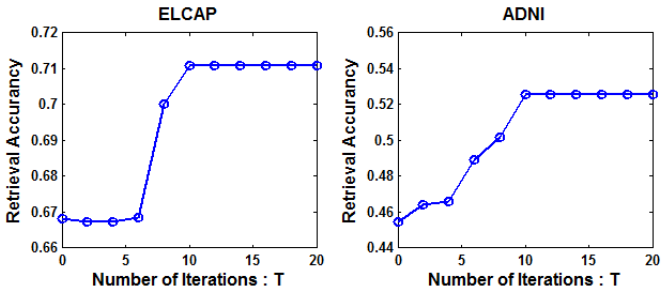

Figure 8: The retrieval accuracy curve given the number of iterations.

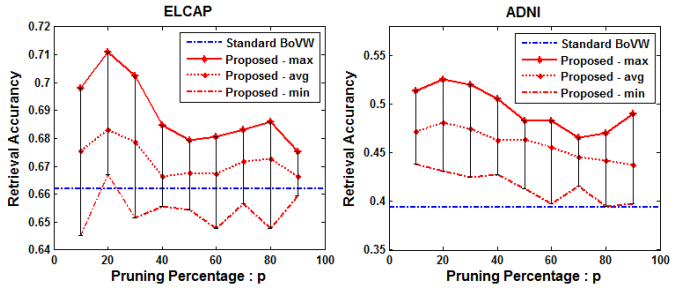

Figure 9: The retrieval accuracy curve given the pruning percentages.

results. Overall, fixing T at 20 was sufficient to obtain stable accuracies given different numbers of iterations.

Fig. 9 shows the effects of different pruning percentages on the two datasets. For each p (10 to 90), the maximum, minimum and average accuracies given different Hs (50 to 600) and *twth*s (0.05 to 1) with T fixed at 20 are reported. In general, the best accuracy results were obtained when p was between 20% and 40%. The similarity between images was better represented on the pruned dictionary, since the similar local patch features were more likely to be assigned to the same word. Based on our observations, incorporating more words for the dictionary construction helped on computing the similarities between the less related items, e.g., a pruning percentage of 40% normally performed better when $K = 9$ but 20% generate higher accuracy when $K = 1$. Hence, we suggested that a larger number of outputs needs more words kept and vice versa.

4.2. Retrieval performance evaluation

4.2.1. Visual retrieval results

Figs. 10 and 11 give the visual examples of the retrieval results from the standard BoVW method and our PD-LST method. The results were obtained on the original dictionary with $M = 100$. Our method conducted the dictionary pruning with $H = 50, twth = 0.2, T = 20$ and $p = 20$. It can be seen that our method can retrieve the cases with the same diagnosis, which are visually similar or different. For example of the lung nodule images, we retrieved #16(W) and #1(W) as the most desired cases. While the first result is visually similar to the query, the second one is with a larger lung nodule than that in the query image and has more noise in the background regions. In addition, the proposed method can find the differences between the visually similar images that present different diseases. For instance, given the query case #298(AD),

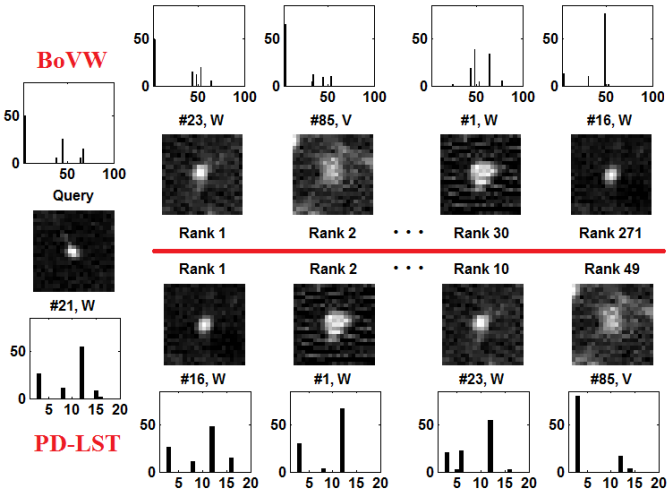

Figure 10: Visual retrieval results from the standard BoVW and PD-LST methods on the ELCAP dataset. The first two retrieved results are displayed for each method, followed by the cases retrieved by the other method. The case indices and categories are given below images. The corresponding word frequency histograms are showed with the x-coordinate as the index of the visual word and y-coordinate as the frequency.

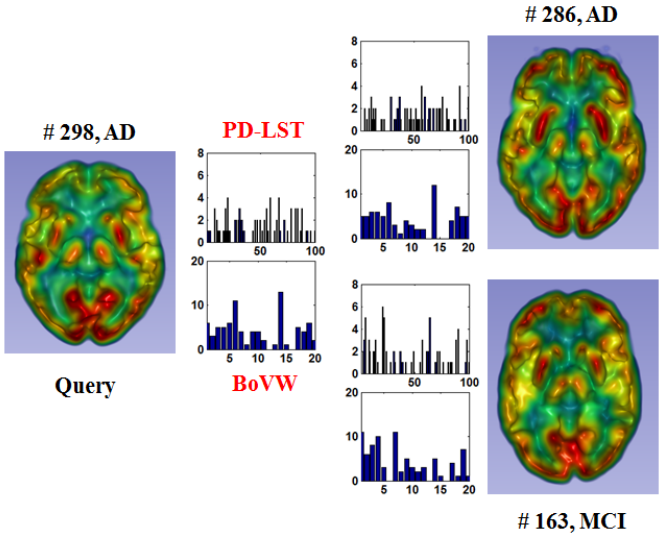

Figure 11: Visual retrieval results from the standard BoVW and PD-LST methods on the ADNI dataset. The case indices and categories are given around images. The corresponding word frequency histograms are showed with the x-coordinate as the index of the visual word and y-coordinate as the frequency.

our method retrieved #286(AD) as the first result and the standard BoVW found #163(MCI). While the two retrieved cases are very similar to the query case regarding the visual appearance, our method obtained the case with the same stage to the query. These observations can be explained by the fact that our method conducted the similarity computation between images through the latent topics, which provide high-level semantic descriptions, instead of merely using the visual content information. Given the pruned dictionary, our method generated a more compacted word frequency histogram that can better differentiate the images with the most discriminative words. We can observe that the frequency distributions between the query and retrieved results were more consistent with the pruned dictionary than the original one.

4.2.2. *Accuracy analysis*

We then quantitatively analysed the performance of the proposed method compared to related retrieval approaches regrading the retrieval accuracy. The experiments were conducted as follows: a) the comparison among the approaches that are based on the subsections of our method's pipeline (Table 2), b) the evaluation regarding the different dictionary pruning approaches in the literature (Fig. 12), and c) the investigation of performance improvement by integrating other retrieval methodologies (Fig. 13).

a) Table 2 shows the retrieval accuracy comparisons of the approaches that are based on parts of our PD-LST method on the ELCAP and ADNI datasets for the 1-NN retrieval. 1) For the GFM, we calculated a global feature vector[2] to represent an individual image and performed the k-NN retrieval with the Euclidean distance. 2) For the BoVW, it followed the standard BoVW model as introduced in Section 2.1. This method was adopted as baseline. 3) For the pLSA-F, we calculated the similarities between images based on the latent topic distribution $P(z_h|I_j)$ obtained during the pLSA parameter estimation (Bosch et al., 2008). 4) For the pLSA-P, we pruned the dictionary based on the conditional probability of a word given a certain latent topic (Foncubierta-Rodríguez et al., 2013), i.e., the topic-word significance. 5) For the VWW, we utilized the overall-word significance to perform visual word weighting, instead of pruning the visual dictionary. The images were represented as vectors with the element of a visual work's significance value other than its frequency. (6) For the TD, we truncated the word frequency histogram of the image based on the pruned dictionary instead of recomputing the histogram.

The GFM retrieval obtained the lower accuracies when compared to BoVW method that was based on the LFM feature extraction. By making use of more local content information, e.g., the surrounding pulmonary structures of a nodule and the spatial structure of different regions of the brain, the LFM is more effective in capturing the similarity between images. The pLSA-F retrieval had the lowest accuracies among all approaches over the two datasets. As we explained previously, latent topics can be used to categorize different anatomical structures. The pLSA-F method represents the images as the latent topic distributions, which can describe what structures are contained in the images but cannot differentiate the role of these structures for the diagnosis. The pLSA-P retrieval obtained higher accuracies regarding the pLSA-F method. This suggests that although the latent topics were not effective in measuring image similarity directly, they can be employed to evaluate the words' significance for the improved similarity computation. On the other hand, the pLSA-P method achieved the similar

[2]In the ELCAP dataset, we extracted one SIFT descriptor from the centroid of the nodule as the global feature; in the ADNI dataset, we used the combination of the eight features (introduced in Section 3.2) for global feature representation.

Table 2: Comparison among the approaches that are based on the subsections in our method's pipeline.

Dataset	Method	*M*	*H*	*twth*	*T*	*P*	*Accuracy*
ELCAP	GFM	-	-	-	-	-	0.6702
	BoVW	600	-	-	-	-	0.6805
	pLSA-F	1200	400	-	-	-	0.6251
	pLSA-P	600	150	-	-	10	0.6850
	VWW	1400	300	0.2	20	-	0.7190
	TD	900	750	0.4	20	10	0.6778
	PD-LST	1100	50	0.1	20	20	0.7414
ADNI	GFM	-	-	-	-	-	0.4060
	BoVW	1700	-	-	-	-	0.4505
	pLSA-F	100	500	-	-	-	0.4487
	pLSA-P	100	400	-	-	60	0.4758
	VWW	900	100	0.2	20	-	0.5106
	TD	1200	200	0.1	20	40	0.5111
	PD-LST	800	350	0.1	20	40	0.5432

accuracies with the BoVW approach, indicating that measuring the significance of the word individually would restrict the ability to identify the most meaningful words.

The VWW method showed retrieval improvement over the aforementioned methods. This was attributed to the overall-word significance that can emphasize the effects of the words with the most discriminative ability, considering the overall relationship between all latent topics and words. The TD method had an approximate 6% accuracy improvement on the ADNI dataset compared to the BoVW. However, TD did not perform well on the ELCAP dataset. This was due to the reason that some lung nodule images only contain a few anatomical structures, e.g., some type W nodule images only have nodules except for the background regions. The word frequency histograms can be very concentrated on a few words. Truncating the frequency histogram may remove these words resulting in an empty histogram. For the PD-LST method, recomputing the frequency histogram based on the pruned dictionary can relocate the local features to other words and thus can reserve the original feature information in the images.

b) In Fig. 12, we compared the following state-of-the-art dictionary pruning approaches on the ELCAP and ADNI datasets. 1) For the OCC, it ranks the words according to their occurrences on all images and prunes the ones with higher frequencies (Yang et al., 2007). 2) For the IDF, the method weights the visual words according to the inverse image frequency and keeps the ones with higher IDF values (Yang et al., 2007). 3) For the pLSA-P, it evaluates the words according to the conditional probabilities for each latent topic and prunes the ones with lower probabilities (Foncubierta-Rodríguez et al., 2013). 4) For the Fisher[3], the method aggregates local image descriptors in terms of the Fisher kernel representation and conducts the dimensionality reduction by principal component analysis (PCA) (Jégou et al., 2012). 5) For the VLAD[4], considered as a simplification of the Fisher kernel, it works on the visual dictionary obtained with k-means rather than with Gaussian mixture model (GMM) in the Fisher method, with the PCA employed for the dimensionality reduction (Jégou et al., 2012). 6) For our PD-LST, the method measures the words according to the overall-word significance and prunes the ones with lower values. During the experiments, we observed that the methods can obtain different retrieval accuracies given different parameter settings and datasets. Therefore, we reported the mean, maximum and minimum of the highest retrieval accuracies across the different dictionaries to compare the overall performances.

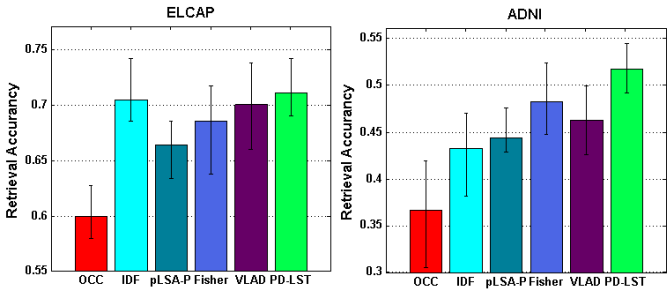

Figure 12: Comparisons among different dictionary pruning approaches on the two datasets. The statistics are from the highest retrieval accuracies across the dictionaries with the sizes from 100 to 2000. The mean of highest accuracies is shown with the bar, and the maximum and minimum are given as the upper and lower errors in the error-bar.

The OCC method generated the worst results and so had an unfavorable performance. These findings were in accordance with the work of Yang et al. who showed that the most frequent words are unlikely to be the stop words (Yang et al., 2007). Such comparison suggested that it was not sufficient to evaluate the significance of the words merely based on occurrence. The IDF method obtained retrieval improvement when compared to the OCC method, indicating that the inverse frequency weighting can assist on identifying discriminative power of the words. The IDF however utilized the direct image-word co-occurrence information to wight the words without further analysing the relationship among the words, which can lead to performance enhancement as used in our method. The pLSA-P method was also more accurate than the OCC approach but were similar to the baseline of BoVW as discussed above. Hence, the pLSA conditional probabilities can describe the significance of a word to a certain extent but it can be further improved upon. Although pLSA pruning did not achieve an observable improvement in retrieval accuracy, it did reduce the number of words required for feature representation and so would improve the efficiency of retrieval.

The Fisher and VLAD methods obtained better retrieval performance by incorporating the local feature encoding process.

[3]The Fisher package was downloaded from `http://lear.inrialpes.fr/src/inria_fisher/`

[4]From VLfeat project, downloaded at: `http://www.vlfeat.org/index.html`

One reason for the improvements is the feature dimensionality reduction with PCA as reported in the work by Jégou et al. In general, for the datasets under study, the highest accuracies were obtained given the reduction according to the first half components with PCA. These improvements showed the benefits of evaluating the discriminative ability of the visual words, though the two methods obtained the visual dictionaries in different ways (with GMM and k-means). Our method achieved the higher retrieval accuracies across all the approaches. The improvements were because that our method not only utilizes the semantic descriptions of the latent topics learnt from the word-image co-occurrence relationship but also measures the contributions of the topics based on the overall relationship between the words and latent topics. In addition, the differences between the maximum and minimum highest accuracies given different dictionary sizes of the proposed method were relatively smaller. These observations indicated the stability of our method with different dictionaries.

c) Fig.13 shows the retrieval performance of the approaches that utilize our PD-LST method. The large margin nearest neighbor (LMNN) retrieval[5] is a supervised method using distance metric learning to identify the most related neighbors (Weinberger and Saul, 2009). This method usually shows performance improvement over k-NN, therefore we employed it to exploit the accuracy enhancement with the pruned dictionary. During the experiments, half of the dataset was randomly selected for training the LMNN model. A leave-one-case-out cross-validation was then used to perform the retrieval given the trained model. The iterative ranking (ITRA) retrieval conducts the retrieval result refinement by calculating the ranking scores of the retrieved items and remaining candidates, which are the similarity measurement to depict their distances with the query, corresponding to the overall-word significance and contribution values of the words and latent topics (Cai et al., 2014b). The ranking score is computed based on the bipartite similarity relationship between the retrieved items and remaining candidates, in a way similar to the ranking method that works between the latent topics and words in this study. Therefore, we employed this method to investigate the performance of ranking method. The retrieval accuracies of the two approaches and k-NN method based on the original and pruned dictionaries are displayed.

Applying the LMNN method on the pruned dictionary gave retrieval accuracy improvement when compared to the original dictionary on the two datasets by incorporating the distance learning metric. The higher accuracies suggested our method's potential on retrieval improvement by using the prior knowledge, e.g., the labeling information. Although employing the ITRA method on the two dictionaries for the ELCAP dataset achieved similar accuracy, applying it on the pruned dictionary for the ADNI dataset gave the highest retrieval accuracy. The improvement showed the benefit of the combination of the retrieval result refinement and the pruned dictionary.

[5]The LMNN package was downloaded from http://www.cse.wustl.edu/~kiliancode/lmnn/lmnn.html

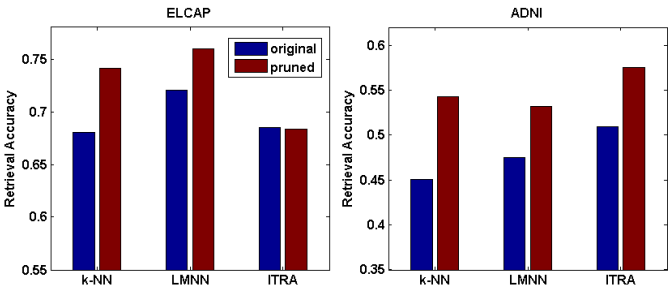

Figure 13: Retrieval accuracy comparisons among the k-NN, LMNN and ITRA approaches. For the LMNN method, the default parameter settings of the LMNN package were used with maximum number of iterations as 1000, suppress output as 0, output dimensionality as 3, tradeoff between loss and regularizer as 0.5. For the ITRA method, we fixed the numbers of initial retrieval results and neighbours for bipartite graph construction at 10 and the iteration number at 20. The parameter settings for obtaining the original and pruned dictionaries were the same as used for the BoVW and PD-LST in Table 2.

Table 3: Comparison of execution time between the offline and online stages.

		ELCAP	ADNI
Offline	ODC	237.713 s	154.541 s
	DP	8.883 s	9.910 s
Online	R	1.856 s	1.246 s

*ODC = original dictionary construction,
DP = dictionary pruning, R = retrieval, s = second.

4.2.3. *Efficiency analysis*

The process of PD-LST retrieval has two components: an offline dictionary construction stage and an online image retrieval stage. The first component contains the original dictionary construction and dictionary pruning and the second consists of the word frequency histogram calculation and similarity computation for the query image. Table 3 shows the comparisons of execution time between the two stages (with same parameter settings for the two datasets as $M = 1000$, $H = 300$, $twth = 0.3$, $T = 20$, $p = 0.2$). It can be observed the offline stage required more processing time (about 25 times for the ELCAP dataset and 15 times for the ADNI dataset) than online retrieval. In addition, the dictionary pruning only occupied a small portion of the whole offline processing.

Our method prunes the dictionary by keeping the most meaningful words and thus obtains a low-dimensional word frequency histogram vector for each image. Such dimensionality reduction can increase the speed of the retrieval process. Table 4 shows the total retrieval time of all query images (for leave-one-case-out validation) from the two datasets based on the original dictionary and the pruned dictionary. Our method had the shortest processing time with the pruned dictionary, suggesting an efficiency improvement.

5. Conclusions and future work

We have presented a PD-LST retrieval method for medical image analysis. Our method focused on dictionary pruning so that only the words with high discriminative power are kept.

Table 4: Comparison of execution time between the standard BOVW and our method.

	ELCAP	ADNI
BoVW	2.567 s	2.053 s
PD-LST	1.856 s	1.246 s

* s = second

The method has two main components: topic-word significance and overall-word significance computing. By pruning the trivial words, the updated BoVW representation better captures the similarity relationships between images and largely reduces the amount of required words.

In the study, we aimed at investigating the performance of dictionary pruning, hence we did not overemphasize on the analysis of the image data themselves. One aspect of the future work will be conducted by further exploring the image data. For example, we will use 3D raw data of the ELCAP dataset, which would provide a more comprehensive description of the lung nodule structures than the 2D images. For tackling a certain task in which more domain-specific knowledge can be incorporated, the analysis about the latent topic categories will be conducted to explore the correlations among different anatomical structures. In addition, the effectiveness of different low-level features, such as histogram of oriented gradients (HOG) (Dalal and Triggs, 2005), GIST (Oliva and Torralba, 2001) and Hashing features (Zhang et al., 2015d), for constructing the BoVW model will be further investigated. More experimental comparisons will be conducted to furhter validate the effectiveness of our methond. For instance, Zhang et al. (Zhang et al., 2015c) proposed an interesting re-ranking approach based on graph analysis, which is highly related to the currently used compared method of ITRA. We will also test our method on other medical or general imaging domains such as the lung tissue classification in high-resolution computed tomography (HRCT) images (Song et al., 2013a) and the thoracic tumor retrieval in positron emission tomography–computed tomography (PET-CT) images (Song et al., 2013b). As the scalability of image data has become an important issue in medical image retrieval, it would be also interesting to test if PD-LST can be integrated in a scalable CBIR approach, e.g., building on vocabulary tree (Jiang et al., 2015b) or hashing methods (Zhang et al., 2015e).

6. Acknowledgments

This work was supported in part by ARC, AADRF, NIH NAMIC (U54 EB005149) and NAC (P41 EB015902).

7. References

Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B., 2011. Content-based image retrieval in radiology: current status and future directions. Journal of Digital Imaging 24, 208–222.

André, B., Vercauteren, T., Ayache, N., 2012a. Content-based retrieval in endomicroscopy: toward an efficient smart atlas for clinical diagnosis, in: Medical Image Computing and Computer Assisted Intervention Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MICCAI MCBR-CDS), pp. 12–23.

André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2011. A smart atlas for endomicroscopy using automated video retrieval. Medical Image Analysis 15, 460–476.

André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2012b. Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Transactions on Medical Imaging 31, 1276–1288.

Avni, U., Greenspan, H., Konen, E., Sharon, M., Goldberger, J., 2011. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. IEEE Transactions on Medical Imaging 30, 733–746.

Bannour, H., Hlaoua, L., el Ayeb, B., 2009. Survey of the adequate descriptor for content-based image retrieval on the Web: global versus local features, in: CORIA, pp. 445–456.

Batchelor, P.G., Castellano Smith, A.D., Hill, D.L.G., Hawkes, D.J., Cox, T.C.S., Dean, A., 2002. Measures of folding applied to the development of the human fetal brain. IEEE Transactions on Medical Imaging 21, 953–965.

Batet, M., Sánchez, D., Valls, A., 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44, 118–125.

Bilenko, M., Basu, S., Mooney, R.J., 2004. Integrating constraints and metric learning in semi-supervised clustering, in: Proceedings of the Twenty-first International Conference on Machine Learning (ICML), p. 11.

Bosch, A., Zisserman, A., Muñoz, X., 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 712–727.

Cai, W., Kim, J., Feng, D., 2008. Content-based medical image retrieval. Elsevier. book section 4. pp. 83–113.

Cai, W., Liu, S., Song, Y., Pujol, S., Kikinis, R., Feng, D., 2014a. A 3D difference-of-Gaussian-based lesion detector for brain PET, in: International Symposium on Biomedical Imaging (ISBI), pp. 677–680.

Cai, W., Liu, S., Wen, L., Eberl, S., Fulham, M.J., Feng, D., 2010. 3D neurological image retrieval with localized pathology-centric CMRGlc patterns, in: International Conference on Image Processing (ICIP), pp. 3201–3204.

Cai, W., Song, Y., Feng, D.D., 2012. Regression and classification based distance metric learning for medical image retrieval, in: International Symposium on Biomedical Imaging (ISBI), pp. 1775–1778.

Cai, W., Zhang, F., Song, Y., Liu, S., Wen, L., Eberl, S., Fulham, M., Feng, D., 2014b. Automated feedback extraction for medical imaging retrieval, in: International Symposium on Biomedical Imaging (ISBI), pp. 907–910.

Caicedo, J., Cruz, A., Gonzalez, F., 2009. Histopathology image classification using bag of features and kernel functions. Artificial Intelligence in Medicine 5651, 126–135.

Castellani, U., Perina, A., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P., 2010. Brain morphometry by probabilistic latent semantic analysis, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 177–184.

Cruz-Roa, A., Gonzalez, F., Galaro, J., Judkins, A., Ellison, D., Baccon, J., Madabhushi, A., Romero, E., 2012. A visual latent semantic approach for automatic analysis and interpretation of anaplastic medulloblastoma virtual slides, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 157–164.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886–893.

Depeursinge, A., Kurtz, C., Beaulieu, C., Napel, S., Rubin, D., 2014. Predicting visual semantic descriptive terms from radiological image data: preliminary results with liver lesions in CT. IEEE Transactions on Medical Imaging , 1.

Diciotti, S., Picozzi, G., Falchini, M., Mascalchi, M., Villari, N., Valli, G., 2008. 3-D segmentation algorithm of small lung nodules in spiral CT images. IEEE Transactions on Information Technology in Biomedicine 12, 7–19.

Duncan, J.S., Ayache, N., 2000. Medical image analysis: progress over two decades and the challenges ahead. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 85–106.

El-Naqa, I., Yang, Y., Galatsanos, N.P., Nishikawa, R.M., Wernick, M.N., 2004. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Transactions on Medical Imaging 23, 1233–1244.

ELCAP, VIA, 2003. ELCAP public lung image database [online database].

URL: http://www.via.cornell.edu/databases/lungdb.html.

Farag, A., Elhabian, S., Graham, J., Farag, A., Falk, R., 2010. Toward precise pulmonary nodule descriptors for nodule type classification, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 626–633.

Farag, A.A., 2013. A variational approach for small-size lung nodule segmentation, in: International Symposium on Biomedical Imaging (ISBI), pp. 81–84.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R., 2012. 3D Slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging 30, 1323–1341.

Foncubierta-Rodríguez, A., Depeursinge, A., Müller, H., 2012. Using multiscale visual words for lung texture classification and retrieval, in: Medical Image Computing and Computer Assisted Intervention Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MICCAI MCBR-CDS), pp. 69–79.

Foncubierta-Rodríguez, A., Herrera, A.G.S.d., Müller, H., 2013. Medical image retrieval using bag of meaningful visual words: unsupervised visual vocabulary pruning with pLSA, in: Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, 2505336. pp. 75–82.

Golub, G.H., Van Loan, C.F., 2012. Matrix computations. volume 3. The Johns Hopkins University Press.

Hanbury, A., Müller, H., Langs, G., Weber, M.A., Menze, B.H., Fernandez, T.S., 2012. Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis, in: Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Springer, pp. 24–29.

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Gray, K.R., Nielsen, C., Rueckert, D., Hajnal, J.V., Hammers, A., 2011. Automatic morphometry in Alzheimer's se and mild cognitive impairment. NeuroImage 56, 2024–2037.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27, 685–691.

Jiang, M., Zhang, S., Fang, R., Metaxas, D.N., 2015a. Leveraging coupled multi-index for scalable retrieval of mammographic masses, in: International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 276–280.

Jiang, M., Zhang, S., Li, H., Metaxas, D., 2015b. Computer-aided diagnosis of mammographic masses using scalable image retrieval. IEEE Transactions on Biomedical Engineering 62, 783–792.

Jiang, M., Zhang, S., Metaxas, D.N., 2014. Detection of mammographic masses by content-based image retrieval, in: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention Workshop on Machine Learning in Medical Imaging (MICCAI MLMI). Springer, pp. 33–41.

Jones, K.S., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1704–1716.

Kumar, A., Kim, J., Cai, W., Fulham, M., Feng, D., 2013. Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data. Journal of Digital Imaging 26, 1025–1039.

Kurtz, C., Beaulieu, C.F., Napel, S., Rubin, D.L., 2014a. A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations. Journal of Biomedical Informatics 49, 227–244.

Kurtz, C., Depeursinge, A., Napel, S., Beaulieu, C.F., Rubin, D.L., 2014b. On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. Medical Image Analysis 18, 1082–1100.

Kwitt, R., Vasconcelos, N., Rasiwasia, N., Uhl, A., Davis, B., Hafner, M., Wrba, F., 2012. Endoscopic image analysis in semantic space. Medical Image Analysis 16, 1415–1422.

Lehmann, T.M., Gold, M., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B., 2004. Content-based image retrieval in medical applications. Methods of Information in Medicine 43, 354–361.

Li, F.F., Pietro, P., 2005. A bayesian hierarchical model for learning natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 524–531.

Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., Pujol, S., Kikinis, R., 2015a. Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. Brain Informatics 2, 167–180.

Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., Pujol, S., Kikinis, R., 2015b. Multimodal neuroimaging computing: the workflows, methods, and platforms. Brain Informatics 2, 181–195.

Liu, S., Cai, W., Song, Y., Pujol, S., Kikinis, R., Feng, D., 2013a. A bag of semantic words model for medical content-based retrieval, in: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MICCAI MCBR-CDS), pp. 125 –131.

Liu, S., Cai, W., Wen, L., Eberl, S., Fulham, M.J., Feng, D.D., 2011. Generalized regional disorder-sensitive-weighting scheme for 3D neuroimaging retrieval, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 7009–7012.

Liu, S., Cai, W., Wen, L., Feng, D., 2013b. Multi-channel brain atrophy pattern analysis in neuroimaging retrieval, in: International Symposium on Biomedical Imaging (ISBI), pp. 202–205.

Liu, S., Cai, W., Wen, L., Feng, D.D., Pujol, S., Kikinis, R., Fulham, M.J., Eberl, S., et al., 2014a. Multi-channel neurodegenerative pattern analysis and its application in Alzheimer's disease characterization. Computerized Medical Imaging and Graphics 38, 436–444.

Liu, S., Liu, S., Pujol, S., Kikinis, R., Feng, D., Cai, W., 2014b. Propagation graph fusion for multi-modal medical content-based retrieval, in: International Conference on Control Automation Robotics and Vision (ICARCV), pp. 849–854.

Liu, S., Song, Y., Cai, W., Pujol, S., Kikinis, R., Wang, X., Feng, D., 2013c. Multifold Bayesian kernelization in Alzheimer's diagnosis, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 303–310.

Liu, S.Q., Liu, S., Zhang, F., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2015c. Longitudinal brain MR retrieval with diffeomorphic demons registration: What happened to those patients with similar changes?, in: International Symposium on Biomedical Imaging (ISBI), pp. 588–591.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: IEEE International Conference on Computer Vsion (ICCV), pp. 1150–1157.

Menze, B., Langs, G., Montillo, A., Kelm, M., Müller, H., Zhang, S., Cai, W., Metaxas, D., 2014. Medical computer vision: Algorithms for big data, in: Lecture Notes in Computer Science 8848.

Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. International Journal of Medical Informatics 73, 1–23.

Napel, S.A., Beaulieu, C.F., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., Rubin, D.L., 2010. Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. Radiology 256, 243–252.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42, 145–175.

Quddus, A., Basir, O., 2012. Semantic image retrieval in magnetic resonance brain volumes. IEEE Transactions on Information Technology in Biomedicine 16, 348–355.

Quellec, G., Lamard, M., Cazuguel, G., Cochener, B., Roux, C., 2010. Wavelet optimization for content-based image retrieval in medical databases. Medical Image Analysis 14, 227–241.

Sivic, J., Zisserman, A., 2003. Video Google: a text retrieval approach to object matching in videos, in: International Conference on Computer Vision (ICCV), pp. 1470–1477.

Song, Y., Cai, W., Eberl, S., Fulham, M.J., Feng, D., 2011a. Discriminative pathological context detection in thoracic images based on multi-level inference, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 191–198.

Song, Y., Cai, W., Eberl, S., Fulham, M.J., Feng, D., 2011b. Thoracic image case retrieval with spatial and contextual information, in: IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1885–1888.

Song, Y., Cai, W., Feng, D., 2011c. Hierarchical spatial matching for medical image retrieval, in: The Annual ACM International Conference on Mul-

timedia Workshop on Medical Multimedia Analysis and Retrieval (ACM MMAR), pp. 1–6.

Song, Y., Cai, W., Huang, H., Zhou, Y., Wang, Y., Feng, D.D., 2015a. Locality-constrained subcluster representation ensemble for lung image classification. Medical image analysis 22, 102–113.

Song, Y., Cai, W., Zhang, F., Huang, H., Zhou, Y., Feng, D., 2015b. Bone texture characterization with fisher encoding of local descriptors, in: International Symposium on Biomedical Imaging (ISBI), pp. 5–8.

Song, Y., Cai, W., Zhou, Y., Feng, D.D., 2013a. Feature-based image patch approximation for lung tissue classification. IEEE Transactions on Medical Imaging 32, 797–808.

Song, Y., Cai, W., Zhou, Y., Wen, L., Feng, D.D., 2013b. Pathology-centric medical image retrieval with hierarchical contextual spatial descriptor, in: International Symposium on Biomedical Imaging (ISBI), pp. 198–201.

Tamaki, T., Yoshimuta, J., Kawakami, M., Raytchev, B., Kaneda, K., Yoshida, S., Takemura, Y., Onji, K., Miyaki, R., Tanaka, S., 2013. Computer-aided colorectal tumor classification in NBI endoscopy using local features. Medical Image Analysis 17, 78–100.

Toews, M., Wells III, W., Collins, D.L., Arbel, T., 2010. Feature-based morphometry: discovering group-related anatomical patterns. NeuroImage 49, 2318–2327.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., 2014. Multiple instance learning for classification of dementia in brain MRI. Medical Image Analysis 18, 808–818.

Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research 10, 207–244.

Wilkinson, J., 1965. Convergence of the LR, QR, and related algorithms. The Computer Journal 8, 77–84.

Wu, D., Lu, L., Bi, J., Shinagawa, Y., Boyer, K., Krishnan, A., Salganicoff, M., 2010. Stratified learning of local anatomical context for lung nodules in CT images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2791–2798.

Xu, J., Faruque, J., Beaulieu, C.F., Rubin, D., Napel, S., 2012. A comprehensive descriptor of shape: method and application to content-based retrieval of similar appearing lesions in medical images. Journal of Digital Imaging 25, 121–128.

Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W., 2007. Evaluating bag-of-visual-words representations in scene classification, in: Proceedings of the International Workshop on Multimedia Information Retrieval, pp. 197–206.

Yang, W., Lu, Z., Yu, M., Huang, M., Feng, Q., Chen, W., 2012. Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single- and multiphase contrast-enhanced CT images. Journal of Digital Imaging 25, 708–719.

Zhang, F., Song, Y., Cai, W., Hauptmann, A.G., Liu, S., Liu, S., Feng, D.D., Chen, M., 2015a. Ranking-based vocabulary pruning in bag-of-features for image retrieval, in: Artificial Life and Computational Intelligence, Lecture Notes in Artificial Intelligence 8955. Springer, pp. 436–445.

Zhang, F., Song, Y., Cai, W., Lee, M.Z., Zhou, Y., Huang, H., Shan, S., Fulham, M., Feng, D., 2014a. Lung nodule classification with multi-level patch-based context analysis. IEEE Transactions on Biomedical Engineering 61, 1155–1166.

Zhang, F., Song, Y., Cai, W., Liu, S., Pujol, S., Kikinis, R., Xia, Y., Fulham, M., Feng, D., 2015b. Pairwise latent semantic association for similarity computation in medical imaging. IEEE transactions on Biomedical Engineering .

Zhang, F., Song, Y., Cai, W., Zhou, Y., Fulham, M., Eberl, S., Shan, S., Feng, D., 2014b. A ranking-based lung nodule image classification method using unlabeled image knowledge, in: International Symposium on Biomedical Imaging (ISBI), pp. 1356–1359.

Zhang, F., Song, Y., Cai, W., Zhou, Y., Shan, S., Feng, D., 2013. Context curves for classification of lung nodule images, in: International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7.

Zhang, J., Zhou, S.K., Brunke, S., Lowery, C., Comaniciu, D., 2010. Detection and retrieval of cysts in joint ultrasound B-mode and elasticity breast images, in: International Symposium on Biomedical Imaging (ISBI), pp. 173–176.

Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N., 2015c. Query specific rank fusion for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 803–815.

Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2015d. Towards large-scale histopathological image analysis: Hashing-based image retrieval. IEEE Transactions on Medical Imaging 34, 496–506.

Zhang, X., Su, H., Yang, L., Zhang, S., 2015e. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5361–5368.

Zhou, X., Depeursinge, A., Müller, H., 2010. Information fusion for combining visual and textual image retrieval, in: International Conference on Pattern Recognition (ICPR), IEEE. pp. 1590–1593.