Pairwise Latent Semantic Association for Similarity Computation in Medical Imaging

Fan Zhang*, *Student Member, IEEE,* Yang Song, *Member, IEEE,* Weidong Cai*, *Member, IEEE,*
Sidong Liu, *Student Member, IEEE,* Siqi Liu, *Student Member, IEEE,* Sonia Pujol, Ron Kikinis, Yong Xia,
Michael J Fulham, David Dagan Feng, *Fellow, IEEE*, and Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Retrieving medical images that present similar diseases is an active research area for diagnostics and therapy. However, it can be problematic given the visual variations between anatomical structures. In this paper, we propose a new feature extraction method for similarity computation in medical imaging. Instead of the low-level visual appearance, we design a CCA-PairLDA feature representation method to capture the similarity between images with high-level semantics. First, we extract the PairLDA topics to represent an image as a mixture of latent semantic topics in an image pair context. Second, we generate a CCA-correlation model to represent the semantic association between an image pair for similarity computation. While PairLDA adjusts the latent topics for all image pairs, CCA-correlation helps to associate an individual image pair. In this way, the semantic descriptions of an image pair are closely correlated, and naturally correspond to similarity computation between images. We evaluated our method on two public medical imaging datasets for image retrieval and showed improved performance.**

***Index Terms*—Medical image retrieval, latent topic, semantic association.**

I. Introduction

OVER the past decade, there has been intensive research in retrieving medical images of the same category, e.g., categories of healthy or abnormal organs, for disease diagnosis and treatment [1]. Computer-based image analysis systems enable automated and efficient search of similar cases in large-scale databases. In these systems, images are represented based on their visual content characteristics [2]–[4]. Similarity between images is then obtained by comparing the visual features. The retrieval performance is however often hindered by visual variations between images of similar categories and visual similarities between images of different categories. In other words, images with similar diagnosis may show different patterns of anatomical structures; on the other hand, the irrelevant cases may show visually similar structures. Thus, it is important to design a descriptive and discriminative feature descriptor so that only images with similar diagnosis will be retrieved.

A. Related Works

Feature extraction is essential for computer-aided diagnosis (CAD) applications, such as medical image retrieval and classification [5], segmentation [6], and lesion detection [7]. The feature descriptor translates an image into a set of numeric vectors and is used to quantitatively characterize the image content. The effectiveness of image feature description depends on distinction and invariance, which means that the descriptor needs to capture the distinctive characteristics and be robust to the various imaging conditions [8]. For this aim, various features have been proposed: the grey-level distribution feature to describe the intensity variations [9]; filter-based feature to identify the edges and shapes [10]; geometric feature to depict the spatial and gradient information [11], etc.

The aforementioned low-level visual features can be directly applied or easily adjusted for different medical imaging systems. However, images with the same disease may present dissimilarities in the usual visual sense [12], [13]. Low-level features are also not descriptive enough to capture the semantic concept that the users are interested in. The semantic gap between the low-level features and users' high-level expectations can thus impair the retrieval performance [14]–[16]. Incorporating semantic descriptions has recently been advocated to deal with the limitations of low-level visual features [17]–[21].

There are studies that make use of the ontological knowledge to infer the semantic concepts [17], [18]. These methods

This work was supported in part by ARC, AADRF, NIH NA-MIC (U54 EB005149), NAC (P41 EB015902), the National Natural Science Foundation of China under Grants 61471297, the Natural Science Foundation of Shaanxi Province, China, under Grant 2015JM6287, and the Fundamental Research Funds for the Central Universities under Grants 3102014JSJ0006. The support of the VIA and ELCAP groups is gratefully acknowledged. The ADNI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from many other sources. Detailed ADNI Acknowledgements information is available in http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

F. Zhang*(fzha8048@uni.sydney.edu.au), W. Cai*(tom.cai@sydney.edu.au) and SD. Liu are with Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia, and Surgical Planning Lab, Brigham & Women's Hospital, Harvard Medical School, MA 02115, USA.

Y. Song, SQ. Liu are with Biomedical and BMIT Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia.

S. Pujol and R. Kikinis are with Surgical Planning Lab, Brigham & Women's Hospital, Harvard Medical School, MA 02115, USA.

Y. Xia is with Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science and Technology Northwestern Polytechnical University, Xi'an 710072, China.

M. J. Fulham is with the Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, NSW 2050, Australia, and Sydney Medical School, University of Sydney, NSW 2006, Australia.

D. D. Feng is with BMIT Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia, and with Med-X Research Institute, Shanghai Jiaotong University, Shanghai 200030, China.

however highly rely on the ontology structure and involve many human interactions, e.g., manual ontology matching. It is preferable to infer the semantics based on the images themselves without external information. The bag-of-visual-words (BoVW) approach is a possible solution by using the image local content information only [22]. The visual words are generated by clustering local features from the image collection. They abstract the similar local content patterns from different images and can reduce the gap between the low-level features and high-level image understanding [15]. Currently, k-means clustering is the most popular method for dictionary construction and has been effectively used for a variety of medical image applications [23], [24]. However it often generates a redundant and noisy dictionary by trying to accommodate all local feature patterns [19].

Instead of directly using the visual words, the latent topic model (LTM) represents the images as a mixture of latent topics, and provides a higher level of semantic description compared to the standard BoVW model [25], [26]. The latent topic is a probability distribution of words, and can be inferred from the co-occurrence relationship between images and words. While the visual words represent the local visual patterns, the topics are regarded as the pattern categories [26]. Accordingly, an image that contains multiple instances of these patterns is interpreted in terms of the pattern category rather than the individual patterns.

LTM has recently been incorporated into medical image analysis. As one of the most representative LTM techniques, probabilistic Latent Semantic Analysis (pLSA) [27] was adopted to extract the semantic relationship between morphological abnormalities on the brain surfaces [20] and to model the histological slides to construct the similarities between the medulloblastoma images [21]. These studies focused on images of the same organ, indicating that LTM can recognize images that are visually similar. pLSA was also used to describe the images with different modalities and various organs [19], suggesting its ability to capture the similarity between images that have large visual appearance variations. Despite the popularity of pLSA, the Latent Dirichlet Allocation (LDA) model [28] is considered more advanced than pLSA by defining a complete generative process [29]. LDA and its variants have been widely investigated for natural language processing problems [30], [31]. They were also adopted in the imaging domain, e.g., natural scene image classification [25], and showed its advantage in image feature description. We expect that LDA-based approaches can provide a more powerful semantic description for similarity computation in medical imaging.

For image retrieval, the similarity computation is conducted in a pairwise context between images. An association can be built to model the similarity relationship between two images. A limitation of the existing LTM techniques is that they typically extract the topics for each image independently. Consequently, the topics are not generated based on image pairs, while the pairwise context is important in similarity computation. In addition, similarity between images is normally measured by direct distance computation between the topic distributions of the two images. This however, does not incorporate the semantic association between the specific image pairs, and might not represent the actual diagnosis-related similarity.

B. Our Contributions

In this work, we propose a LTM-based CCA-PairLDA feature extraction method to retrieve images of similar disease characteristics. Our CCA-PairLDA method has two main components: latent topic extraction and semantic association generation. For the latent topic extraction, we designed a PairLDA-topic generation process by inferring the latent topics in the contexts of image pairs. For the semantic association generation, we designed a CCA-correlation extraction process by learning an association coefficient between images of the same diagnosis with canonical correlation analysis (CCA) [32]. In our method, the PairLDA adjusts the topic distributions for image pairs rather than individual images, and the CCA-correlation helps to make the distributions correlated closely between images of similar semantics. The images are then represented as the PairLDA topic distribution conditioned on the CCA-correlation model, which is our CCA-PairLDA feature. Similar images are retrieved based on the distances between the CCA-PairLDA feature vectors.

We evaluated our method on two publicly available datasets - the Early Lung Cancer Action Program (ELCAP) [33] and Alzheimer's disease Neuroimaging Initiative (ADNI) [34]. Our prior work [35] showed the effectiveness of the semantic association-based analysis and reported some preliminary results. In this work, we enhance the PairLDA topic extraction based on the local features for better image-word co-occurrence exploration, instead of the global features. We also elaborate the CCA-correlation process with further association coefficient generation and parameter estimation details. In addition, the formulation of CCA-PairLDA is enhanced to provide a general image representation, so that the similarity computation can be conducted across the training and testing images. We extend the evaluation to the ELCAP dataset for lung nodule image retrieval task, in addition to the originally used ADNI dataset. The more comprehensive performance evaluations are performed on the two datasets.

The structure of this paper is as follows. In Section II, we introduce the details of our CCA-PairLDA method. In Section III we describe the experimental datasets and experimental design. In Section IV we present the experimental results and discussion. We provide a conclusion and an outline of future work in Section V.

II. Methods

A. Outline of CCA-PairLDA

The goal of our CCA-PairLDA method is to find an optimal feature representation of medical images in the semantic association space, which can be used to construct the similarity relationships between different groups of images. The method flow contains four stages that correspond to image representation at four cascading granularity levels: local feature level, visual word level, latent topic level and semantic association level, as shown in Fig. 1. Accordingly, the similarity between

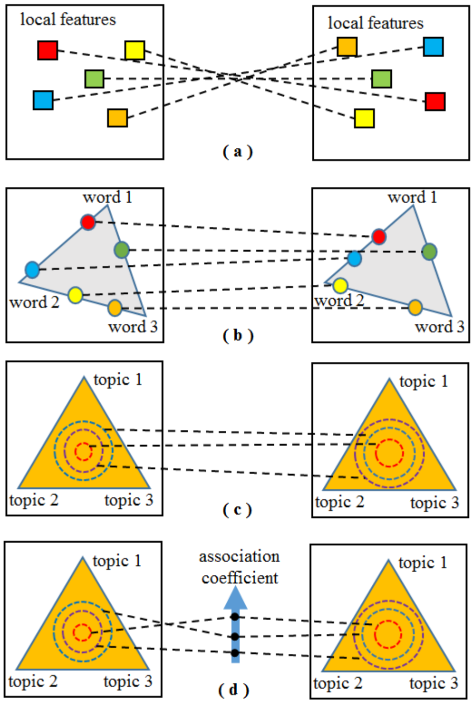

Fig. 1. Similarity computation between images in different granularity spaces: (a) local feature space, where the image is represented as an orderless collection of local features (multiple color rectangles), (b) visual word space, where the image is modeled as a word frequency histogram (multiple colored circles) derived by assigning local features over the word simplex (grey triangle) where each corner corresponds to a word, (c) latent topic space, where the image is described by a latent topic distribution (concentric circles) over the topic simplex (orange triangle) where each corner is a latent topic, and (d) semantic association space, where the images are associated with the association coefficient (blue arrow) between the latent topic distributions. The latent topics are extracted based on the visual words (local features) across the images.

images can be calculated based on the local feature sets, word frequency histograms, latent topic distributions and semantic association coefficients. Our CCA-PairLDA method focuses on the third and fourth levels, with 1) PairLDA topic extraction, which generates latent topics based on the image-word co-occurrence relationship in image pairs, and 2) CCA-correlation generation, which learns association coefficient between the PairLDA topic distributions of images.

Outline of the CCA-PairLDA feature extraction method is shown in Fig. 2. The first two stages of our method follow the standard BoVW construction, including local feature extraction, visual dictionary generation, and word frequency histogram calculation [22]. Then, we divide the entire image set randomly into two subsets as source and target sets. Images from the source set are paired with all of those from the target set, as shown in Fig. 2(a). PairLDA topics are extracted based on all image pairs without involving the label information. In the next step, we select a group of training images with category labels to learn the association coefficient between the PairLDA topic distributions of each individual image pair. The training set contains the same number of source and target images, and one-to-one pairing of training images of the same category is randomly constructed across the source and target sets, as shown in Fig. 2(b). After training, the test images (as well as the training images) are represented as the PairLDA topic distribution conditioned on the CCA-correlation model to measure the similarity between images for retrieval, as shown in Fig. 2(c).

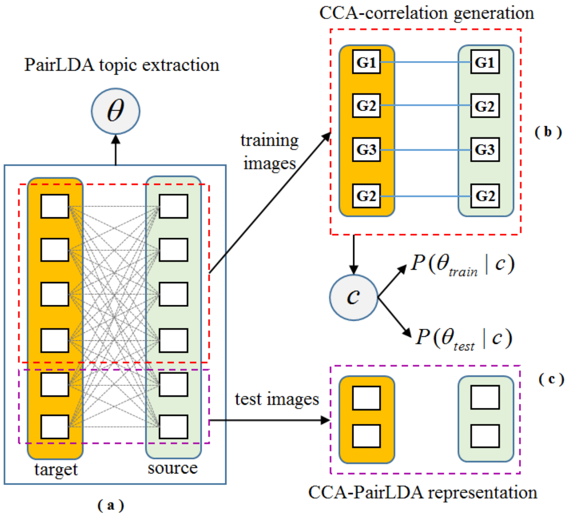

Fig. 2. Outline of our CCA-PairLDA feature representation: (a) PairLDA topics are extracted by pairing all images from target and source sets, resulting in a topic distribution θ_l for each image I_l, (b) association coefficient c_{st} is learnt to capture the semantic association between the training image pair (I_s, I_t) with the same category label (indicated with 'G'), and (c) the test images (similarly for training images) are represented as the CCA-PairLDA feature, which is the probability of its PairLDA topic distribution given the CCA-correlation model.

B. PairLDA Topic Extraction

PairLDA assumes that an image is represented by a set of hidden variables, i.e., the latent topics, to describe the image semantics. It is a generative model that generates the observable visual words from a convex combination of the latent topics as introduced in LDA. However, unlike LDA that assigns a different subset of topics to each individual image [29], our method constructs a shared topic distribution for a pair of images from the source and target sets respectively to represent the relationship between the two images. As a result, the extracted topics can fit for image pairs instead of single images. This pairwise relationship naturally corresponds to similarity measure between images.

Assume we have an image set $EI = \{I_l | l \in 1...N\}$, which is divided into the source image set $SI = \{I_s | s \in 1...N_{SI}\}$ and target image set $TI = \{I_t | t \in 1...N_{TI}\}$ with $SI \cup TI = EI$ and $SI \cap TI = \emptyset$. A total of D image pairs (I_s, I_t) are formed from the N_{SI} sources and N_{TI} targets with $D = N_{SI} \times N_{TI}$. Denote the dictionary as $DY = \{w_v | v \in 1...W\}$ where w is the word and W is the dictionary size. In our PairLDA method, each image is represented as a random mixture over K latent topics: for the source set, we have a source topic collection $ST = \{T_k^{SI} | k \in 1...K\}$, and for the target set, we have $TT = \{T_k^{TI} | k \in 1...K\}$. Fig. 3 shows the dependencies among all variables and depicts the choices of the word w_i^s and word w_j^t from their topics z_i^s and z_j^t for the

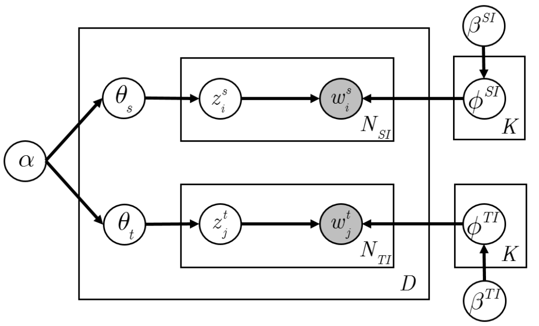

Fig. 3. Graphical model of PairLDA generation. α and β are the priors of Dirichlet distributions; θ is the $N \times K$ matrix indicating the image-topic distribution; ϕ is the $K \times W$ matrix indicating the topic-word distribution; z and w are the instances of variables for the topic and word.

image pair. We use w_i^s to denote the word in source image I_s with index i corresponding to the word w_v in DY, and w_j^t to denote the word in target image I_t with index j corresponding to the word $w_{v'}$ in DY. The generative process contains the following steps:

1) For each image I_l, choose a topic distribution θ_l of size K from a symmetric Dirichlet prior with concentration parameter α, i.e., $\theta_l \sim \boldsymbol{Dir}(\alpha)$, where θ_l represents the probability of topic occurrences in this image;
2) For each topic T_k^{SI} of the source set, choose a word distribution ϕ_k^{SI} of size W from a symmetric Dirichlet prior with concentration parameter β^{SI}, i.e., $\phi_k^{SI} \sim \boldsymbol{Dir}(\beta^{SI})$, where ϕ_k^{SI} represents the probability of word occurrences given the topic T_k^{SI} in any source image I_s. Similarly, choose a word distribution ϕ_k^{TI} with the parameter β^{TI}, i.e., $\phi_k^{TI} \sim \boldsymbol{Dir}(\beta^{TI})$;
3) For each image pair (I_s, I_t),
    a) Choose a topic $z_i^s \in ST$ from a Multinomial prior with the topic distribution θ_s for image I_s, i.e., $z_i^s \sim \boldsymbol{Mult}(\theta_s)$. Similarly, choose a topic z_j^t from θ_t, i.e., $z_j^t \sim \boldsymbol{Mult}(\theta_t)$;
    b) Choose a word $w_i^s \in DY$ from a Multinomial prior with the word distribution ϕ^{SI} conditioned on the topic z_i^s for image I_s, i.e., $w_i^s \sim \boldsymbol{Mult}(\phi_{z_i^s}^{SI})$. Similarly, choose a word w_j^t from ϕ^{TI}, i.e., $w_j^t \sim \boldsymbol{Mult}(\phi_{z_j^t}^{TI})$;

The original LDA does not consider the image pairing information and generates one collection of topics. In our PairLDA, however, the words are generated from two separate topic collections and thus the images from the source and target sets become independent at the word level. On the other hand, the topic distribution θ is chosen from the Dirichlet distribution of α for both of the images. This adjusts the topic distributions of the image pair collectively and hence makes the image pair correlated at the topic level.

We extended the Gibbs sampling algorithm to learn the parameters in PairLDA, i.e., θ, ϕ^{SI} and ϕ^{TI}. The conditional posterior for choosing the topics of an image pair for the words w_v and $w_{v'}$, i.e., the update equation used in Gibbs sampling, is:

$$\begin{aligned} &P(z_i^s = T_k^{SI}, z_j^t = T_{k'}^{TI} | w_i^s = w_v, w_j^t = w_{v'}, \\ &\quad \vec{z}_{\neg i}^s, \vec{z}_{\neg j}^t, \vec{w}_{\neg i}^s, \vec{w}_{\neg j}^t, \alpha, \beta^{SI}, \beta^{TI}) \\ &\propto \frac{n_{k,\neg i}^{(w_v)} + \beta^{SI}}{\sum\limits_{w_v \in I_s} n_{k,\neg i}^{(w_v)} + W\beta^{SI}} (n_{s,\neg i}^{(k)} + \alpha) \cdot \\ &\quad \frac{n_{k',\neg j}^{(w_{v'})} + \beta^{TI}}{\sum\limits_{w_{v'} \in I_t} n_{k',\neg j}^{(w_{v'})} + W\beta^{TI}} (n_{t,\neg j}^{(k')} + \alpha) \end{aligned} \tag{1}$$

where $\vec{w}^s = \{w_i^s = w_v, \vec{w}_{\neg i}^s\}$, $\vec{z}^s = \{z_i^s = T_k^{SI}, \vec{z}_{\neg i}\}$, $n_{k,\neg i}^{(w_v)}$ indicates the number of occurrences that a word w_v (excluding the word w_i^s in image I_s) has been observed with topic T_k^{SI}, and $n_{s,\neg i}^{(k)}$ indicates the number of occurrences that a topic T_k^{SI} has been observed with a word (excluding the word w_i^s) of image I_s. The notations $\vec{w}^t$, $w_{v'}$, $\vec{z}^t$, $T_{k'}^{TI}$, $n_{k',\neg j}^{(w_{v'})}$ and $n_{t,\neg j}^{(k')}$ are defined similarly for image I_t. Subsequently, the parameters introduced in Pair-LDA can be estimated with the following equations:

$$E(\phi_{k,v}^{SI}) = \frac{n_k^{(w_v)} + \beta^{SI}}{\sum\limits_{w_v \in I_s} n_k^{(w_v)} + W\beta^{SI}} \tag{2}$$

$$E(\theta_{s,k}) = \frac{n_s^{(k)} + \alpha}{\sum\limits_{k \in [1,K]} n_s^{(k)} + K\alpha} \tag{3}$$

Eq. (2) gives the independent topic collection for the source set, and Eq. (3) is the topic distribution of the source image. The parameters for the target set are estimated similarly. During the experiments, we evaluated the parameters (α from $0.1/K$ to $100/K$ and β from 10^{-4} to 10^{-1}) and found that these parameters had insignificant influence which is similar to the findings by Lu and Ramage et al. [36], [37]. The more widely used settings of $\alpha = 50/K$, $\beta^{SI} = 0.01$ and $\beta^{TI} = 0.01$ were thus fixed for all experiments. The overall time complexity of PairLDA is $O(N_{it}KN_{SI}N_{TI})$. N_{it} is the number of iterations of Gibbs sampling and was set at 30 throughout the experiments, which was sufficient to generate stable sampling results. Considering that including a few new images would have insignificant influence on the whole topic distributions, we can sample the topics for an individual new image without changing the existing topic collections. On the other hand, if a large number of new images are introduced, we suggest a new PairLDA topic extraction is necessary since the topic collections could largely change. The pseudo code of PairLDA extraction is displayed in Algorithm 1.

C. CCA Correlation Generation

PairLDA topics can be directly used to measure the similarity between images by calculating their topic distribution distance in latent topic space (Fig. 1(c)). However, PairLDA generates the topics in the context of all image pairs, adjusting the topics to fit for each image pair. This would reduce

Algorithm 1 PairLDA Extraction

```
Input: word vector matrices $\boldsymbol{w}^{SI}$ and $\boldsymbol{w}^{TI}$, hyperparameters $\alpha$, $\beta^{SI}$ and $\beta^{TI}$, topic number $K$, iteration number $N_{it}$
Output: word-topic and topic-image distributions $\theta$ and $\phi$.
Set all occurrence variables $n_*^* = 0$.
// Initialization of word-topic and topic-image distributions
for all source images $I_s \in SI$ do
    for all words $w_v$ in $I_s$ do
        Randomly sample topic $T_k^{SI} \sim \boldsymbol{Mult}(1/K)$;
        Increase word-topic occurrence $n_k^{(w_v)}$ by 1;
        Increase topic-image occurrence $n_s^{(k)}$ by 1;
Similarly, initialize the occurrences for the target set;
// Gibbs sampling
for $it \in [1, N_{it}]$ do
    for all image pairs $(I_s, I_t) \in \{I_s \in SI, I_t \in TI\}$ do
        for $k \in [1, K]$ do
            Decrease $n_k^{(w_i^s)}$, $n_k^{(w_j^t)}$, $n_s^{(k)}$ and $n_t^{(k)}$ by 1;
            Sample $T_k^{SI} \sim P(z_i^s = T_k^{SI}, z_j^t = T_{k'}^{TI})$ for source;
            Increase $n_k^{(w_i^s)}$, $n_k^{(w_j^t)}$, $n_s^{(k)}$ and $n_t^{(k)}$ by 1;
            Similarly, sample $T_{k'}^{TI}$ and update occurrences for target;
Parameter estimation according to Eqs. (2) - (3);
return $\theta$ and $\phi$;
```

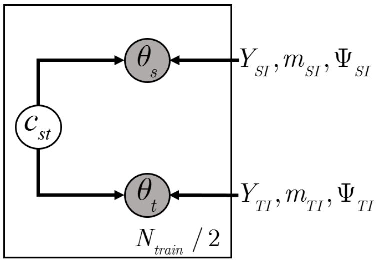

Fig. 4. Graphical model of CCA-correlation generation. c_{st} is calculated for each image pair by constructing the one-to-one pairing between the images of the same category, with a total of $N_{train}/2$ coefficients learnt given the training size N_{train}.

the difference between the topic distributions of two images and hence their discriminative ability in the topic space. To overcome this issue, we propose to extract latent semantic description of an image differently when coupled with others, i.e., making the topic distribution interpreted differently in different pair contexts (Fig. 1(d)).

At this stage, we would like to capture the semantic association of an image pair based on the extracted Pair-LDA topics. Rather than directly using the topic distributions that are obtained in the context of all image pairs, an association coefficient is defined to connect the images of the same category in an individual image pair context. In other words, while PairLDA adapts the latent topics for all image pairs, the semantic association works on an individual image pair from the same category. In this way, the topic distribution for one image can be flexibly assigned if it is paired with different images, enhancing the correlation between the two images. We adopt the CCA model for this purpose.

Given two sets of random variables, CCA finds a pair of linear transformations, making the transformed variables of these two sets correlated to the largest extent. Fig. 4 gives the probabilistic interpretation of CCA, which depicts the generation process of the latent topic distribution from the association coefficient, which is a latent variable following a normal distribution [32]. The method involves the parameter set $PS = \{Y_{SI}, Y_{TI}, m_{SI}, m_{TI}, \Psi_{SI}, \Psi_{TI}\}$, where Y is a $K \times d$ transformation matrix that relates to the two sets of variables (θ and c) with the length of canonical correlations d, m is a vector of size K that makes the transformed variables to non-zero mean and Ψ is an error covariance matrix of size $K \times K$. The generative process is described as follows:

1) For each pair of image (I_s, I_t), choose an association coefficient c_{st} of size d from a Normal distribution with parameters $\mathbf{0}$ and $\mathbf{I}_d$, i.e., $c_{st} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}_d)$, where $\mathbf{0}$ is the mean vector of size d and $\mathbf{I}_d$ is the unit variance of size d with $1 \leq d \leq K$ denoting the length of the coefficient.
2) For the topic distributions of the two images, choose θ_s from a Normal distribution based on the association coefficient c_{st}, i.e., $\theta_s \sim \boldsymbol{N}(Y_{SI}c_{st} + m_{SI}, \Psi_{SI})$. Similarly, choose θ_t from c_{st}, i.e., $\theta_t \sim \boldsymbol{N}(Y_{TI}c_{st} + m_{TI}, \Psi_{TI})$.

While θ_s and θ_t represent the images in terms of PairLDA topics, the association coefficient c_{st} indicates how these two images are correlated at the semantic association level. The coefficient is adapted between different topic distributions, making the semantic descriptions of images interpreted differently for different image pairs. During the training process, each training source image is paired with one training target image of the same category. With this one-to-one mapping manner, we associate the individual image pairs, instead of all image pairs as in the PairLDA topic extraction. Given the CCA transformation, the transformed topic distributions of the two mapped images are correlated to the largest extent. Thus, while PairLDA generates the latent topics for all image pairs simultaneously, CCA-correlation helps to associate the individual image pairs.

The parameter set $PS = \{Y_{SI}, Y_{TI}, m_{SI}, m_{TI}, \Psi_{SI}, \Psi_{TI}\}$ can be estimated using maximum likelihood estimation [32], as:

$$\begin{aligned}
\overline{Y}_{SI} &= \widetilde{\Sigma}_{SI,SI} U_{SI,d} M_{SI}, \\
\overline{Y}_{TI} &= \widetilde{\Sigma}_{TI,TI} U_{TI,d} M_{TI}, \\
\overline{m}_{SI} &= \widetilde{m}_{SI}, \\
\overline{m}_{TI} &= \widetilde{m}_{TI}, \\
\overline{\Psi}_{SI} &= \widetilde{\Sigma}_{SI,SI} - \overline{Y}_{SI} \overline{Y}_{SI}^{\top}, \\
\overline{\Psi}_{TI} &= \widetilde{\Sigma}_{TI,TI} - \overline{Y}_{TI} \overline{Y}_{TI}^{\top}
\end{aligned} \tag{4}$$

where $\widetilde{\Sigma} = \left\{ \begin{array}{cc} \widetilde{\Sigma}_{SI,SI} & \widetilde{\Sigma}_{SI,TI} \\ \widetilde{\Sigma}_{TI,SI} & \widetilde{\Sigma}_{TI,TI} \end{array} \right\}$ is the sample covariance matrix of θ, $U_{SI,d}$ and $U_{TI,d}$ are the matrices containing the first d canonical directions, M_{SI} and M_{TI} are arbitrary matrices such that $M_{SI} M_{TI}^T = C_d$ that is the diagonal matrix with the first d canonical correlations and $\widetilde{m}_{SI}$ and $\widetilde{m}_{TI}$ are the sample means. During the experiments, d was set as the minimum of the ranks of topic distribution matrices of the training source and target sets, and $M_{SI} = M_{TI} = C_d^{1/2} R$ where R is a rotation matrix of size d [1]. In this way, we have the collection of Y of size $K \times d$, m of size $K \times 1$ and Ψ of size $K \times K$, for the source and target sets. The posterior expectations and variances of the association coefficient c_{st} given θ_s and θ_t are:

$$\begin{aligned} E(c_{st}|\theta_s) &= M_{SI}^\top U_{SI,d}^\top(\theta_s - \overline{m}_{SI}), \\ E(c_{st}|\theta_t) &= M_{TI}^\top U_{TI,d}^\top(\theta_t - \overline{m}_{TI}), \\ var(c_{st}|\theta_s) &= \mathbf{I}_d - M_{SI} M_{SI}^\top, \\ var(c_{st}|\theta_t) &= \mathbf{I}_d - M_{TI} M_{TI}^\top \end{aligned} \tag{5}$$

The pseudo code of CCA correlation generation is displayed in Algorithm 2.

Algorithm 2 CCA Correlation Generation

```
Input: training topic distribution matrices $\boldsymbol{\theta}^{SI}_{train}$ and $\boldsymbol{\theta}^{TI}_{train}$,
Output: model parameter set $PS$
Compute covariance matrices $\widetilde{\Sigma} = cov(\boldsymbol{\theta}^{SI}_{train}, \boldsymbol{\theta}^{TI}_{train})$;
Compute canonical directions $(U_{SI}, U_{TI}) = svd(\widetilde{\Sigma})$;
Set $d = min(rank(\boldsymbol{\theta}^{SI}_{train}), rank(\boldsymbol{\theta}^{TI}_{train}))$;
Compute correlation matrix $C_d = (U_{SI}^\top \widetilde{\Sigma}_{SI,TI} U_{TI})_d$;
Set $M_{SI} = M_{TI} = C_d^{1/2} R$;
Compute sample means $\widetilde{m}_{SI}$ and $\widetilde{m}_{TI}$;
Parameter estimation according to Eq. (4);
return $PS = \{\overline{Y}_{SI}, \overline{Y}_{TI}, \overline{m}_{SI}, \overline{m}_{TI}, \overline{\Psi}_{SI}, \overline{\Psi}_{TI}\}$;
```

D. CCA-PairLDA Feature Representation

In this study, we compute the similarity between two images in the semantic association space. An image I_l is represented as the PairLDA topic distribution θ_l with the association coefficient c learnt using the CCA-correlation model with $PS = \{\overline{Y}_{SI}, \overline{Y}_{TI}, \overline{m}_{SI}, \overline{m}_{TI}, \overline{\Psi}_{SI}, \overline{\Psi}_{TI}\}$ in Eq. (4). Therefore, we have our CCA-PairLDA feature representation as:

$$P(\theta_l|c) \propto P(\theta_l|PS) \tag{6}$$

The similarity between a test image I_{test} and a training image I_{train} is thus formulated as:

$$\begin{aligned} Sim(I_{test}, I_{train}) &\propto P(\theta_{test}, \theta_{train}|PS) \\ &= P(\theta_{test}|PS) P(\theta_{train}|PS) \end{aligned} \tag{7}$$

The CCA-PairLDA feature of the images I_{test} and I_{train} can be estimated according to Eq. (4), as

$$P(\theta_{test}|PS) \sim \begin{cases} \boldsymbol{N}(\overline{Y}_{SI} c + \overline{m}_{SI}, \overline{\Psi}_{SI}), if I_{test} \in SI \\ \boldsymbol{N}(\overline{Y}_{TI} c + \overline{m}_{TI}, \overline{\Psi}_{TI}), if I_{test} \in TI \end{cases} \tag{8}$$

$$P(\theta_{train}|PS) \sim \begin{cases} \boldsymbol{N}(\overline{Y}_{SI} c + \overline{m}_{SI}, \overline{\Psi}_{SI}), if I_{train} \in SI \\ \boldsymbol{N}(\overline{Y}_{TI} c + \overline{m}_{TI}, \overline{\Psi}_{TI}), if I_{train} \in TI \end{cases} \tag{9}$$

where the correlation coefficient c follows a normal distribution according to Eq. (5), as:

$$c \sim \begin{cases} \boldsymbol{N}(M_{SI}^\top U_{SI,d}^\top(\theta_{train} - \overline{m}_{SI}), \mathbf{I}_d - M_{SI} M_{SI}^\top), \\ \quad if I_{train} \in SI \\ \boldsymbol{N}(M_{TI}^\top U_{TI,d}^\top(\theta_{train} - \overline{m}_{TI}), \mathbf{I}_d - M_{TI} M_{TI}^\top), \\ \quad if I_{train} \in TI \end{cases} \tag{10}$$

The pseudo code of CCA-PairLDA similarity computation is displayed in Algorithm 3.

Algorithm 3 CCA-PairLDA Similarity Computation

```
Input: training and test topic distribution matrices $\boldsymbol{\theta}_{train}$ and
    $\boldsymbol{\theta}_{test}$, CCA Correlation model $PS$
Output: similarity matrix Sim
for all test images $I_{test}$ do
    Obtain type of $I_{test}$ as source or target;
    Compute $c$ based on the type using Eq. (10);
    Estimate feature of $I_{test}$ using Eq. (8);
    for all training images $I_{train}$ do
        Obtain type of $I_{train}$ as source or target;
        Compute $c$ based on the type using Eq. (10);
        Estimate feature of $I_{train}$ using Eq. (9);
        Compute similarity $Sim(I_{test}, I_{train})$ using Eq. (7);
return Sim;
```

III. Datasets and Experimental Design

We employed two publicly available medical imaging datasets – the ELCAP [33] and ADNI [34] – to evaluate our CCA-PairLDA feature representation for retrieving images of similar disease and symptom.

A. Datasets and Implementation

For the ELCAP dataset, our aim is to retrieve the images of lung nodules of the same category. Lung nodules are small masses in the lung. Intra-parenchymal nodules are more likely to be malignant than those connected with the surrounding structures. Hence, the lung nodules are normally divided into four different categories according to their location and connection with surrounding structures, as: well-circumscribed (W), vascularized (V), juxta-pleural (J) and pleural-tail (P), as shown in Fig. 5. The ELCAP database contains 50 sets of low-dose computed tomography (LDCT) human lung scans with 379 unduplicated lung nodules annotated at the centroid, where 57 are type W, 60 are type V, 114 are type J and 148 are type P.

In the ELCAP database the lung nodules are small and have an average size of 4 × 4 pixels across the centroid in the axial direction. Therefore, for nodule analysis, a sub-window of 33 × 33 pixels was cropped from each image slice with the annotated nodule centroid at the center. With each pixel around the annotated centroid (including the centroid pixel) as a keypoint, we computed a scale invariant feature transform (SIFT) [38] descriptor using the VLfeat[2] library [39], with the parameter $frames = [px; py; sc = 4; or = 0]$, where px and py indicate the pixel position, sc is the scale and or is

[1] In the experiments, R was computed arbitrarily based on the nested dimensions method as introduced in https://en.wikipedia.org/wiki/Rotation_matrix.

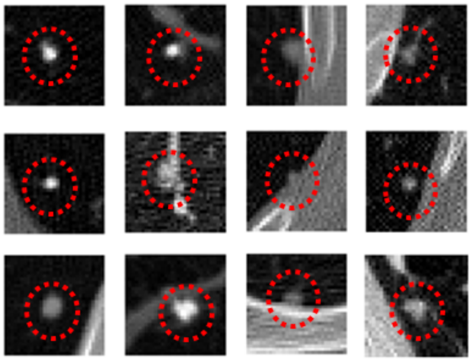

Fig. 5. Transaxial CT images with typical nodules (from left to right) - well-circumscribed (W), vascularized (V), juxta-pleural (J) and pleural-tail (P). The nodules are circled in red.

the orientation. A 128-dimension vector was obtained for each frame and used as a local feature. Based on our previous work [40], [41], incorporating too many or too few surrounding structures would reduce the performance of recognizing the nodule type. Therefore, a total of 100 local features were used by selecting the pixels near the nodule centroid.

For the ADNI dataset, our goal is to retrieve the brain images that show the same progression stage to dementia. Alzheimer's disease (AD) is the most common neurodegenerative disorder and its symptoms of cognitive impairment develop gradually over years. Mild cognitive impairment (MCI) represents the transitional state between AD and cognitively normal with a high conversion rate to AD. The risk of progression to dementia is higher if more regions display glucose hypometabolism [42], as displayed in Fig. 6. The ADNI database comprises 331 subjects with magnetic resonance (MR) and positron emission tomography (PET) scans, which provide important structural and functional information of the brain [43], [44]. The diagnoses of these subjects include three stages, where 77 are cognitively normal (CN), 169 are MCI and 85 are AD.

In the ADNI database, we pre-processed the MR and PET data following the ADNI image correction protocols and non-linearly registered to segment the entire brain into 83 functional regions [42]. We firstly used FSL FLIRT [45] to align the PET images to the corresponding MR images. The selected MR data in ADNI database have been labeled with 83 brain regions of interest (ROI) using the multi-atlas propagation with enhanced registration (MAPER) approach [46]–[48]. The MAPER-generated labelmaps were then applied to segment the brain PET data. A complete list of the 83 ROIs can be found in previous papers [46], [49]. After the segmentation, for each ROI, we extracted eight features. The mean [50] and Fisher [51] indices, and difference-of-Gaussian-based features (DoG area, DoG contrast, DoG mean) features [52], [53] were extracted from the PET data, and solidity, convexity [54] and gray matter volume [46] were extracted from the MR data. The gray matter volume features were calculated as the summation of the gray matter voxels captured by voxel-based morphometry (VBM) [55]. Thus, we obtained an 8-dimension vector for each ROI as one local feature, and 83 local feature vectors for each subject.

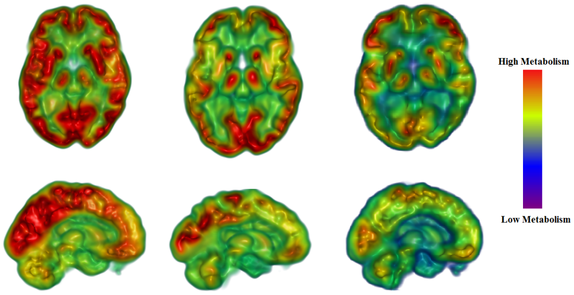

Fig. 6. Lesion patterns for the three stages, shown from left to right as cognitively normal, MCI and AD. Red indicates high metabolism and blue color indicates low metabolism. The images were generated using 3D Slicer V4.3.1 [56].

For each dataset, with the local features extracted from all images, we applied the k-means method to generate the dictionary with the Euclidean distance. Then visual word frequency histograms were generated to represent the images as BoVW models. The co-occurrence relationship between the images and words was obtained for PairLDA topic extraction.

B. Experimental Design and Evaluation Metrics

While PairLDA topics were extracted in an unsupervised manner within the entire image collection, CCA-correlation was learnt during the supervised training stage. We conducted 5-fold cross-validation. The parameters of dictionary size W and topic number K were optimized on the training set by maximizing the mean accuracy. The mean and standard deviation of the accuracies across the five folds were reported for experimental comparisons. The training set was divided into targets and sources evenly to build the one-to-one mapping for CCA-correlation generation. The testing images were used as queries to conduct the retrieval of top tk related results following Eq. (7).

The retrieval performance was quantitatively measured using the average accuracy of N_{test} queries with the top tk retrieval results, as,

$$Accuracy = (\sum_{q \in [1, N_{test}]} (TP_{I_q}/tk))/N_{test} \tag{11}$$

where TP is the number of true positive items within the tk retrieved results for the query image I_q with the index of q. To assess the performance of different categories, we also analyzed the recall and precision:

$$Recall = TP_{I_q}/(TP_{I_q} + FN_{I_q}) \tag{12}$$

$$Precision = TP_{I_q}/(TP_{I_q} + FP_{I_q}) \tag{13}$$

where FN and FP are the numbers of false negative and false positive items within the tk retrieved results for the query image I_q.

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

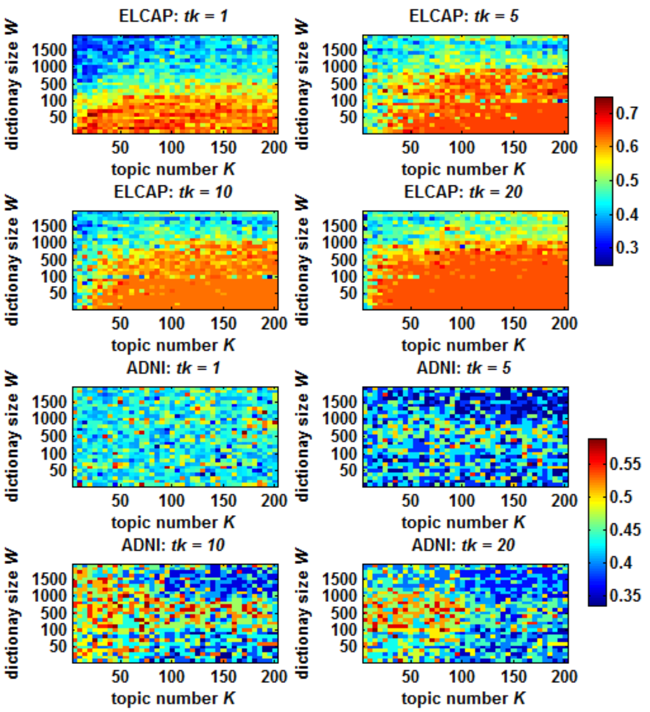

Fig. 7. The retrieval accuracy matrices given different topic numbers and dictionaries sizes of the different numbers of outputs, i.e., tk = 1, 5, 10 and 20. K ranges from 5 to 200 with interval 5, and W is from 10 to 2000 with interval 10 for 10 to 100 and interval 100 for 100 to 2000. The accuracies with pure guessing were 0.25 and 0.33 for the ELCAP and ADNI datasets.

IV. Experimental Results and Discussion

A. Visual Words v.s. Topics

Our PairLDA extracts the latent topics from the co-occurrence relationship between the images and visual words. An appropriate size of dictionary (W) is important for constructing the co-occurrence relationship. In addition, similarity between images is measured in the semantic association space of the PairLDA topics. The proper number of latent topics (K) is essential to capture the similarity relationship. Fig. 7 shows the effects of these two parameters on the two datasets. Here, for the different Ws (from 10 to 2000, with interval 10 from 10 to 100 and interval 100 from 100 to 2000) and Ks from 5 to 200 (with interval 5), we displayed the average accuracy of top tk retrievals given $tk = 1, 5, 10$ and 20, as the colour value of the table. The results represented in the following sections were obtained with the same ranges of the two parameters as aforementioned.

For the ELCAP dataset, the accuracy reduced when the dictionary size was large. For the ADNI dataset, the highest accuracies were obtained given a medium range of dictionary size (W was from 100 to 1000). The different accuracy variations on these two datasets were attributed to the characteristics of the imaging data. As we introduced previously, the dictionary generated by k-means is often redundant and noisy when its size is large. For lung nodule images, the nodules are small and nodules of different categories exhibit similar visual patterns. A larger dictionary would identify more unnecessary visual details and thus influenced more mismatching between images of the same category. To the contrary, the visual details uncovered by a larger dictionary could indeed be useful in obtaining a more descriptiveness representation of the brain images that present the very complicated anatomical structures.

The results from the ELCAP dataset present a relatively clear accuracy pattern varying the dictionary sizes and topic numbers when compared to the ADNI dataset. This is because the visual features of brain images can have large intra-class variation and small inter-class difference. For example, given a late stage MCI query subject that has presented higher transition risk to AD, there could be some late stage MCI subjects or early stage AD subjects, all of which are very similar to the query. Given the stochastic nature of the algorithm and different parameter settings, the topmost ranked results could be obtained from either of the two categories across different validation runs. Thus, the accuracy matrix presents a noisy appearance when the output number was small (e.g., tk = 1 or 5). Increasing the output numbers could result in a more stable set of the most similar cases, though the ranking orders of them may be different across the different runs. We hence can observe better retrieval performance from a smaller topic set and a larger dictionary. We did not observe improved performance for lower values of $K (K < 5)$ and $W (W < 10)$. For the extreme values, e.g., $W = 1$ or $K = 1$, the method will fail since the same feature vectors will be obtained for each case.

B. LTM-based Representation

Our CCA-PairLDA is a LTM-based approach that extracts the Pair-LDA topics and then applies CCA to learn the correlations between these topics. We conducted comparisons among six LTM-based methods on the two datasets. The first three methods, i.e., pLSA, LDA and PairLDA, were used to show the effects of different latent topic extraction methods. The other three, i.e., CCA-pLSA, CCA-LDA and CCA-PairLDA, were employed to show the performance of CCA-correlation learnt upon these topics. Fig. 8 shows the statistics of 1-NN retrieval results, with varying settings of dictionary sizes (from 10 to 2000) and topic numbers (from 5 to 200).

Among the first three approaches that calculated the similarity in the latent topic space, pLSA generated the worst retrieval performance. One aspect was that pLSA had lower overall retrieval accuracy in terms of median, minimum and upper extreme values. In addition, although the maximum accuracy of pLSA was close to LDA and better than PairLDA, it resulted in many outliers, which suggested its unstable performance. LDA obtained higher retrieval accuracy and better stability than pLSA, indicating its advantages over pLSA with a complete generative process. Our PairLDA delivered the most stable performance among these three approaches, with a small standard deviation and the upper and lower extremes close to the maximum and minimum, but the retrieval accuracy was unfavorable when compared to LDA. The lower accuracy was due to the lowered discriminative ability of PairLDA in the latent topic space. The LDA method learnt the latent topic considering a single image, which can emphasize the most discriminative topics in the latent topic space. The PairLDA approach extracted the latent topics in the context of image pairs and adjusted the topics for all pairs, thus reduced the difference between individual image pairs. On the other hand, adjusting the topics for all image pairs could reduce the influence of the trivial topics, hence PairLDA was more stable when compared to LDA.

Better retrieval performances were achieved by the latter three methods that constructed the similarity relationship based on the CCA-correlation. While accuracy improvements from pLSA and LDA topics were relatively small, variations of retrieval accuracies across different dictionaries and topics became smaller. For example, there were fewer outliers with CCA-pLSA compared to pLSA, and the upper and lower extremes of CCA-LDA were similar to its maximum and mini-mum values. These were due to the fact that CCA-correlation is able to make the topics correlated closely with variable trans-formation. However, pLSA and LDA topics were generated independently and thus did not lead to considerable accuracy improvement. PairLDA topics, however, were generated by pairing the images, which is more suitable for CCA-correlation generation that works on the correlated variables. Therefore, although PairLDA individually obtained lower accuracy when compared to the LDA approach, the combination of CCA-correlation and PairLDA (CCA-PairLDA) obtained the best retrieval results across all of these LTM-based approaches.

C. Retrieval Accuracy, Recall and Precision

Fig. 9 shows the retrieval accuracies using our CCA-PairLDA and the BoVW approach, with varying numbers of outputs on the two datasets. Here, the mean ± standard deviation of the accuracies across the 5-folds cross-validation were reported. It can be observed that higher retrieval accuracies were achieved with CCA-PairLDA. Furthermore, while BoVW had lower accuracies when the number of outputs was small, CCA-PairLDA obtained relatively consistent accuracies across the different numbers of retrieval outputs.

Tables I and II give the recall and precision comparisons between the BoVW and CCA-PairLDA approaches on the two datasets with different numbers of outputs as $tk = 1$, 9, 19 and 29 across the different categories. For a given output number, the mean ± standard deviation of the recalls and precisions were displayed. Overall, our method outperformed the control method with higher recalls and precisions across different groups. Furthermore, our method obtained more balanced recalls and precisions on different groups. For example of the ELCAP dataset, type W obtained lower recalls and precisions with the BoVW method due to the fact that the type W nodules are very similar to types V and P and are usually retrieved incorrectly. Our CCA-PairLDA generated more balanced recalls and precisions across the three types by correctly retrieving type W nodules, specifically when the output numbers were relatively large. For the ADNI dataset, MCI is usually considered as the transitional state from CN to AD. Both CN and AD subjects were inclined to be incorrectly retrieved as MCI, resulting in very low recalls of CN and AD in particular for the large output numbers. Our methods can better relieve this problem when compared to the BoVW method with higher recalls of these two stages and higher precisions overall. We also tested our method on binary brain image classification task (AD v.s. normal control) with the 1-NN method, and obtained an accuracy of 0.773 ± 0.053. It is close to the result from Simpson et al [57]; however, we expect improved performance if we have more advanced features specific to the brain anatomical information as used by them, which will be explored in our future work.

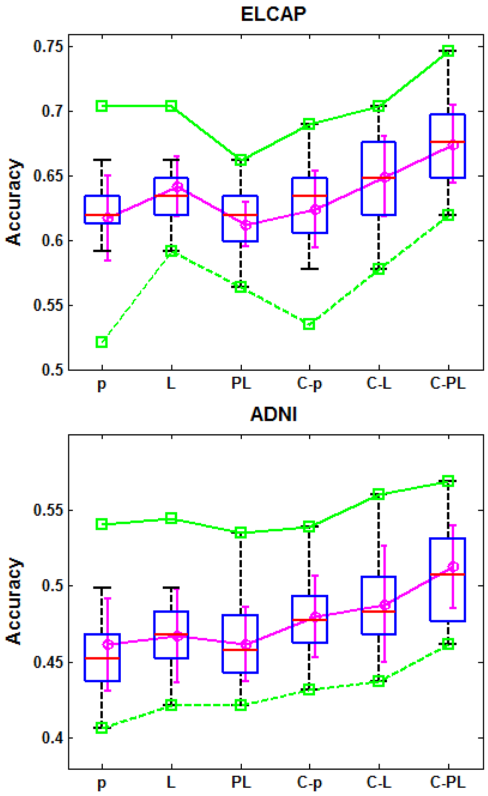

Fig. 8. Comparisons of different LTM-based approaches: pLSA (p), LDA (L), PairLDA (PL), CCA-pLSA (C-p), CCA-LDA (C-L), CCA-PairLDA (C-PL). 9 different statistical values are displayed: maximum and minimum (green lines), mean (mauve circle), standard deviation (mauve error bar), upper and lower extremes (black error bar), upper and lower quartiles (blue rectangle) and median (red line). The upper and lower extremes are the highest and lowest values not considered outliers[3].

In Figs. 10 and 11, we displayed the visual retrieval results from the BoVW and CCA-PairLDA approaches. Given these queries, both of the two methods can correctly retrieve the cases with the same class of the query as the most related results. However, the CCA-PairLDA tended to have better performance as more results were included. This was due to the reason that CCA-PairLDA represented the images with latent association instead of merely with visual appearance. In this way, we can find the cases that may be visually different but within the same category. For instance of the

[3] The points are regarded as outliers if they are greater than $q3+ot(q3-q1)$ or less than $q1-ot(q3-q1)$, where $q1$ and $q3$ are the lower and upper quartiles. The $ot = 1.5$ was used in Fig. 8., corresponding to approximately $\pm 2.7\sigma$ and 99.3 coverage if the data are normally distributed, where σ is the variance.

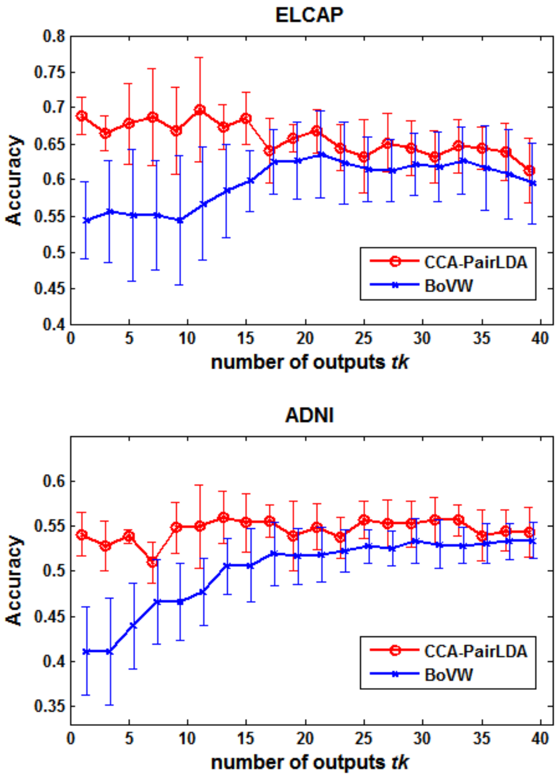

Fig. 9. The retrieval accuracy curves given different retrieval outputs.

TABLE I
RECALLS AND PRECISIONS ACROSS THE 4 TYPES ON THE ELCAP DATASET

tk	class	*Recall*		*Precision*	
		BoVW	CCA-PairLDA	BoVW	CCA-PairLDA
1	W	0.356 ± 0.137	**0.639 ± 0.072**	0.280 ± 0.098	**0.567 ± 0.073**
	V	0.460 ± 0.106	**0.593 ± 0.093**	0.633 ± 0.100	**0.815 ± 0.099**
	J	0.506 ± 0.122	**0.641 ± 0.089**	0.483 ± 0.082	**0.580 ± 0.065**
	P	0.667 ± 0.035	**0.774 ± 0.080**	0.651 ± 0.046	**0.766 ± 0.038**
9	W	0.267 ± 0.094	**0.472 ± 0.087**	0.234 ± 0.120	**0.519 ± 0.109**
	V	0.413 ± 0.196	**0.497 ± 0.106**	0.577 ± 0.097	**0.695 ± 0.050**
	J	0.500 ± 0.122	**0.587 ± 0.191**	0.529 ± 0.099	**0.702 ± 0.093**
	P	0.723 ± 0.070	**0.861 ± 0.057**	0.658 ± 0.052	**0.700 ± 0.101**
19	W	0.333 ± 0.157	**0.480 ± 0.174**	0.385 ± 0.144	**0.619 ± 0.155**
	V	0.507 ± 0.248	**0.565 ± 0.091**	0.693 ± 0.081	**0.724 ± 0.109**
	J	0.500 ± 0.125	**0.503 ± 0.123**	0.643 ± 0.061	**0.697 ± 0.172**
	P	0.830 ± 0.037	**0.850 ± 0.070**	0.657 ± 0.043	**0.669 ± 0.043**
29	W	0.300 ± 0.139	**0.405 ± 0.113**	0.452 ± 0.149	**0.526 ± 0.097**
	V	0.553 ± 0.252	**0.596 ± 0.327**	**0.746 ± 0.096**	0.736 ± 0.049
	J	0.400 ± 0.144	**0.444 ± 0.107**	**0.666 ± 0.124**	0.656 ± 0.053
	P	**0.880 ± 0.057**	0.859 ± 0.111	0.617 ± 0.041	**0.658 ± 0.087**

brain images, given the MCI query, although BoVW obtained a more visually similar case for the second result, our method correctly found one from the same category of MCI.

D. Retrieval Method

Our CCA-PairLDA is a feature extraction method that presents the image in a semantic association space and can be used with different retrieval methods. We compared with several retrieval methods to show the effectiveness of our CCA-PairLDA feature on medical image similarity computation. We conducted the comparison between the BoVW and our CCA-PairLDA features, with the various retrieval methods: k-NN, large margin nearest neighbor (LMNN) [58] and iterative ranking (ITRA) [59]. Specifically, the k-NN retrieval is the classical retrieval method. The LMNN retrieval is a supervised method using distance metric learning to identify the most related neighbors before conducting the k-NN retrieval. The ITRA retrieval refines the retrieval results from k-NN by calculating the ranking scores of the retrieved items and remaining candidates. Fig. 12 displays the mean ± standard deviation of accuracies for each method given different outputs with as $tk = 1$, 9, 19 and 29. The BoVW based methods involved the parameter W, and the CCA-PairLDA method contained the parameters W and K. For the LMNN method[4], we applied the default settings for distance metric learning (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

TABLE II
RECALLS AND PRECISIONS ACROSS THE 3 STAGES ON THE ADNI DATASET

tk	class	*Recall*		*Precision*	
		BoVW	CCA-PairLDA	BoVW	CCA-PairLDA
1	CN	0.280 ± 0.088	**0.600 ± 0.027**	0.263 ± 0.062	**0.424 ± 0.017**
	MCI	0.505 ± 0.091	**0.669 ± 0.028**	0.475 ± 0.058	**0.665 ± 0.016**
	AD	0.229 ± 0.095	**0.426 ± 0.061**	0.280 ± 0.106	**0.678 ± 0.049**
9	CN	0.214 ± 0.101	**0.335 ± 0.065**	0.299 ± 0.103	**0.586 ± 0.089**
	MCI	0.722 ± 0.098	**0.842 ± 0.040**	0.486 ± 0.031	**0.588 ± 0.014**
	AD	0.077 ± 0.045	**0.258 ± 0.045**	0.269 ± 0.150	**0.481 ± 0.054**
19	CN	0.131 ± 0.086	**0.162 ± 0.049**	0.371 ± 0.213	**0.508 ± 0.172**
	MCI	0.867 ± 0.063	**0.910 ± 0.012**	0.503 ± 0.019	**0.557 ± 0.018**
	AD	0.061 ± 0.046	**0.212 ± 0.046**	0.383 ± 0.302	**0.504 ± 0.153**
29	CN	0.080 ± 0.065	**0.106 ± 0.031**	0.469 ± 0.314	**0.774 ± 0.215**
	MCI	**0.935 ± 0.044**	0.925 ± 0.068	0.507 ± 0.013	**0.543 ± 0.003**
	AD	0.039 ± 0.055	**0.211 ± 0.084**	0.402 ± 0.371	**0.454 ± 0.081**

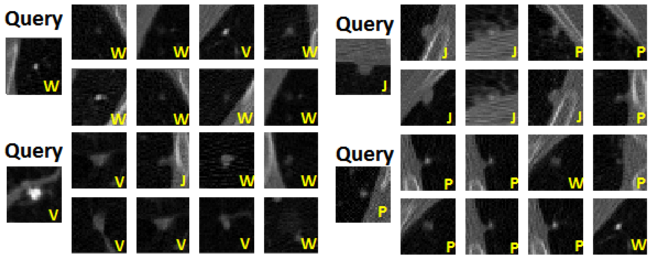

Fig. 10. Visual retrieval results of the BoVW (upper row) and CCA-PairLDA (lower row) features given the K-NN methods on the ELCAP dataset. The top four ranked images are displayed.

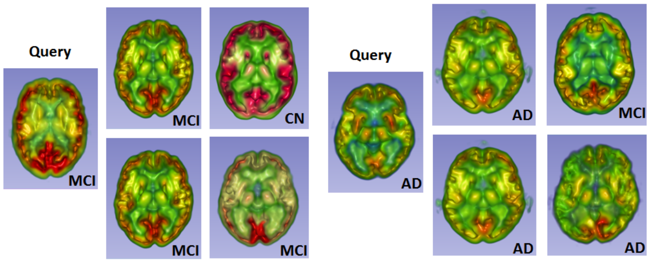

Fig. 11. Visual retrieval results of the BoVW (upper row) and CCA-PairLDA (lower row) features given the K-NN methods on the ADNI dataset. The top two ranked images are displayed.

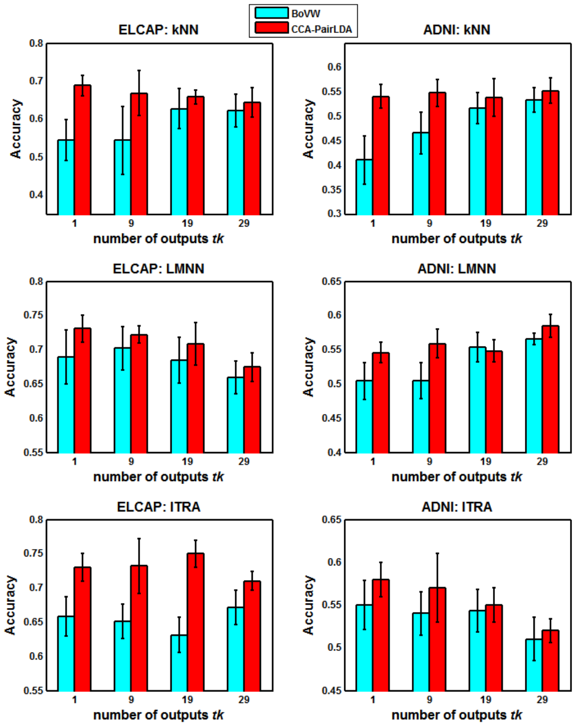

Fig. 12. Comparison of different retrieval methods between the BoVW and CCA-PairLDA features.

results and neighbours for bipartite graph construction at 10 and the iteration number at 20.

It can be observed that higher retrieval accuracies were obtained with our CCA-PairLDA feature when compared to the BoVW feature with the different retrieval approaches. Although the BoVW approach can be used to bridge the gap between the low-level visual appearance and high-level semantic understanding by grouping the similar local features, our CCA-PairLDA can provide more powerful semantic descriptions by further inferring the latent topics using the co-occurrence relationship between the images and words. Furthermore, improvements of retrieval performance using different retrieval methods were different. LMNN and ITRA achieved larger improvements compared to k-NN based on the BoVW feature, especially when the number of outputs was small, e.g., $tk = 1$ and 9. The improvements were due to that LMNN incorporated a learning process and ITRA involved the retrieval result refinement. However, for our CCA-PairLDA feature, relatively smaller improvements can be observed with LMNN and ITRA over k-NN. This was because CCA-PairLDA involved the CCA-correlation generation in a supervised way, leading to a smaller improvement when further learning process was introduced by LMNN. ITRA used the relationship information between the image pairs of the initial retrievals and remaining candidates, which was utilized during the Pair-LDA topic extracting stage in our method, thus the ITRA refinement did not obtain obvious improvements. These observations showed that the retrieval improvements of our CCA-PairLDA method over BoVW across these retrieval methods were attributed more to the feature extraction than the retrieval methods. In addition, the retrieval accuracies with our CCA-PairLDA feature were relatively consistent across the various retrieval methods, indicating that our feature extraction method can be generally effective for different retrieval approaches.

V. Conclusions and Future Work

We have presented a CCA-PairLDA feature representation method for medical image similarity computation. Our method compared the images in a semantic association space where the semantic descriptions of the two images can be closely correlated. The method has two main components: a PairLDA topic extraction and a CCA-correlation generation. Experimental results on two datasets (ELCAP and ADNI) showed that our method achieved high retrieval accuracies.

Future work will include applying our method to large scale data analysis, and we will test our method on other imaging domains such as the lung tissue classification in high-resolution computed tomography (HRCT) images [11], the thoracic tumor retrieval in positron emission tomography computed tomography (PET-CT) images [60] and the brain image classification of AD and normal controls [57]. In addition, we will further investigate if a more sophisticated design of low-level local feature will help to provide a better retrieval performance with our CCA-PairLDA feature representation, e.g., the deformation-based features of voxel- and tenser-based morphometry features of the brain images. We will also explore incorporating more domain-specific anatomical information and inter- and intra-category disease characteristics into our feature model for further improvement, e.g., of the binary AD classification.

References

[1] H. Müller *et al.*, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[2] T. M. Lehmann *et al.*, "Content-based image retrieval in medical applications," *Methods of Information in Medicine*, vol. 43, no. 4, pp. 354–361, 2004.

[3] C. B. Akgül *et al.*, "Content-based image retrieval in radiology: current status and future directions," *Journal of Digital Imaging*, vol. 24, no. 2, pp. 208–222, 2011.

[4] W. Cai *et al.*, "Content-based retrieval of dynamic PET functional images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 2, pp. 152–158, 2000.

[5] M. I. Daoud *et al.*, "Tissue classification using ultrasound-induced variations in acoustic backscattering features," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 2, pp. 310–320, 2013.

[6] F. Riaz *et al.*, "Impact of visual features on the segmentation of gastroenterology images using normalized cuts," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1191–1201, 2013.

[7] Y. Song *et al.*, "Lesion detection and characterization with context driven approximation in thoracic FDG PET-CT images of NSCLC studies," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 408–421, 2014.

[8] A. Farag *et al.*, "Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2011, pp. 169–172.

[9] M. Gangeh *et al.*, "A texton-based approach for the classification of lung parenchyma in CT images," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 6363, 2010, pp. 595–602.

[4] The LMNN package was downloaded from http://www.cse.wustl.edu/~kilian/code/lmnn/lmnn.html

[10] A. Depeursinge *et al.*, "Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 665–675, 2012.

[11] Y. Song *et al.*, "Feature-based image patch approximation for lung tissue classification," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 797–808, 2013.

[12] W. Cai *et al.*, *Content-based medical image retrieval*. Elsevier, 2008, book section 4, pp. 83–113.

[13] Y. Song *et al.*, "Locality-constrained subcluster representation ensemble for lung image classification," *Medical image analysis*, vol. 22, no. 1, pp. 102–113, 2015.

[14] R. Kwitt *et al.*, "Endoscopic image analysis in semantic space," *Medical Image Analysis*, vol. 16, no. 7, pp. 1415–1422, 2012.

[15] B. André *et al.*, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1276–1288, 2012.

[16] A. Depeursinge *et al.*, "Predicting visual semantic descriptive terms from radiological image data: preliminary results with liver lesions in CT," *IEEE Transactions on Medical Imaging*, no. 99, pp. 1–1, 2014.

[17] C. Kurtz *et al.*, "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications," *Medical Image Analysis*, vol. 18, no. 7, pp. 1082–1100, 2014.

[18] M. Batet *et al.*, "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 118–125, 2011.

[19] A. Foncubierta Rodríguez *et al.*, "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*, 2013, pp. 75–82.

[20] U. Castellani *et al.*, "Brain morphometry by probabilistic latent semantic analysis," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 6362, 2010, pp. 177–184.

[21] A. Cruz-Roa *et al.*, "A visual latent semantic approach for automatic analysis and interpretation of anaplastic medulloblastoma virtual slides," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 7510, 2012, pp. 157–164.

[22] J. Caicedo *et al.*, "Histopathology image classification using bag of features and kernel functions," *Artificial Intelligence in Medicine*, vol. 5651, pp. 126–135, 2009.

[23] U. Avni *et al.*, "X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words," *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 733–746, 2011.

[24] W. Yang *et al.*, "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*, vol. 25, no. 6, pp. 708–719, 2012.

[25] F.-F. Li *et al.*, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2005, pp. 524–531.

[26] A. Bosch *et al.*, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[27] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[28] D. M. Blei *et al.*, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[29] G. Heinrich, "Parameter estimation for text analysis," Report, 2005. [Online]. Available: http://www.arbylon.net/publications/text-est.pdf

[30] T. L. Griffiths *et al.*, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[31] D. Mimno *et al.*, "Polylingual topic models," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 880–889.

[32] F. R. Bach *et al.*, "A probabilistic interpretation of canonical correlation analysis," Report, 2005. [Online]. Available: http://www.stat.berkeley.edu/~jordan/688.pdf

[33] ELCAP *et al.*, "ELCAP public lung image database," 2003. [Online]. Available: http://www.via.cornell.edu/databases/lungdb.html

[34] C. R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[35] F. Zhang *et al.*, "Latent semantic association for medical image retrieval," in *The International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2014.

[36] C. Lu *et al.*, "The topic-perspective model for social tagging systems," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1835891: ACM, 2010, Conference Proceedings, pp. 683–692.

[37] D. Ramage *et al.*, "Clustering the tagged web," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2009, Conference Proceedings, pp. 54–63.

[38] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 1150–1157.

[39] A. Vedaldi *et al.*, "VLfeat: An open and portable library of computer vision algorithms," in *Proceedings of the International Conference on Multimedia (ACMMM)*, 2012, pp. 1469–1472.

[40] F. Zhang *et al.*, "Lung nodule classification with multi-level patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1155–1166, 2014.

[41] ——, "Context curves for classification of lung nodule images," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2013, pp. 1–7.

[42] S. Liu *et al.*, "Multifold bayesian kernelization in Alzheimer's diagnosis," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013, pp. 303–310.

[43] ——, "Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders," *Brain Informatics*, vol. 2, no. 3, pp. 1–14, 2015.

[44] ——, "Multimodal neuroimaging computing: the workflows, methods, and platforms," *Brain Informatics*, vol. 2, no. 3, pp. 1–15, 2015.

[45] M. Jenkinson *et al.*, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[46] R. A. Heckemann *et al.*, "Automatic morphometry in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 56, no. 4, pp. 2024–2037, 2011.

[47] J. Mazziotta *et al.*, "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm)," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[48] J. A. Schnabel *et al.*, "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*. Springer, 2001, pp. 573–581.

[49] S. Liu *et al.*, "Multi-channel neurodegenerative pattern analysis and its application in Alzheimer's disease characterization," *Computerized Medical Imaging and Graphics*, vol. 38, no. 6, pp. 436–444, 2014.

[50] W. Cai *et al.*, "3D neurological image retrieval with localized pathology-centric CMRGlc patterns," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 3201–3204.

[51] S. Liu *et al.*, "Generalized regional disorder-sensitive-weighting scheme for 3D neuroimaging retrieval," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011, pp. 7009–7012.

[52] M. Toews *et al.*, "Feature-based morphometry: discovering group-related anatomical patterns," *NeuroImage*, vol. 49, no. 3, pp. 2318–2327, 2010.

[53] W. Cai *et al.*, "A 3D difference-of-Gaussian-based lesion detector for brain PET," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2014, pp. 677–680.

[54] P. G. Batchelor *et al.*, "Measures of folding applied to the development of the human fetal brain," *IEEE Transactions on Medical Imaging*, vol. 21, no. 8, pp. 953–965, 2002.

[55] J. Ashburner *et al.*, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[56] A. Fedorov *et al.*, "3D Slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

[57] I. J. Simpson *et al.*, "Ensemble learning incorporating uncertain registration," *IEEE transactions on medical imaging*, vol. 32, no. 4, pp. 748–756, 2013.

[58] K. Q. Weinberger *et al.*, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[59] W. Cai *et al.*, "Automated feedback extraction for medical imaging retrieval," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2014, pp. 907–910.

[60] Y. Song *et al.*, "Pathology-centric medical image retrieval with hierarchical contextual spatial descriptor," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 198–201.