# Ranking-based Vocabulary Pruning in Bag-of-Features for Image Retrieval

Fan Zhang[1], Yang Song[1], Weidong Cai[1], Alexander G. Hauptmann[2], Sidong Liu[1], Siqi Liu[1], Dagan Feng[1,3], and Mei Chen[4]

[1] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, Australia
[2] School of Computer Science, Carnegie Mellon University, United States
[3] Med-X Research Institute, Shanghai Jiaotong University, China
[4] Intel Science and Technology Center on Embedded Computing, Carnegie Mellon University, United States

**Abstract.** Content-based image retrieval (CBIR) has been applied to a variety of medical applications, e.g., pathology research and clinical decision support, and bag-of-features (BOF) model is one of the most widely used techniques. In this study, we address the problem of vocabulary pruning to reduce the influence from the redundant and noisy visual words. The conditional probability of each word upon the hidden topics extracted using probabilistic Latent Semantic Analysis (pLSA) is firstly calculated. A ranking method is then proposed to compute the significance of the words based on the relationship between the words and topics. Experiments on the publicly available Early Lung Cancer Action Program (ELCAP) database show that the method can reduce the number of words required while improving the retrieval performance. The proposed method is applicable to general image retrieval since it is independent of the problem domain.

**Keywords:** Image retrieval, bag-of-features model, vocabulary pruning

## 1 Introduction

Content-based image retrieval (CBIR), i.e., searching for images similar to the query under certain similarity metric, has been an active research field [1–3]. It can be a powerful tool for diagnosis assistance and decision support [4–17]. Most of the state-of-the-art approaches build upon the bag-of-features (BOF) model [18–21], which represents one image as a frequency histogram of visual words based on a vocabulary obtained by quantizing the local features of all images in the database.

In general, the main steps involved in BOF-based CBIR include feature extraction, vocabulary construction, BOF generation, and similarity calculation [22]. Firstly, feature extraction is conducted by computing local descriptors for the regions of interest (ROIs). Then, a codebook is built offline within the feature space. The obtained codebook is usually referred to as a visual vocabulary, and the cluster centers are visual words. The BOF representation of an image is obtained by assigning a visual word to each of the feature descriptors, resulting in a frequency histogram of the visual words to calculate the similarity [23–25].

Current work on BOF model mostly tackles the feature design and neighbor identification problems [26–31]. In this study, we focus on the vocabulary construction step, particularly on vocabulary pruning. Visual vocabulary is usually redundant, over-complete and noisy, which leads to a high-dimensional feature space [32]. It reduces the retrieval accuracy due to the sparse data problem, and increases the computational cost. Therefore, it is preferable to obtain a more meaningful and compact vocabulary. The supervised method [33] is considered as an option by giving the prior knowledge, e.g., the prefixed vocabulary size. It has limited adaptability since the words used are different under various imaging conditions. An unsupervised approach was proposed in [32] by extracting the latent associations between image set and vocabulary with probabilistic Latent Semantic Analysis (pLSA). The visual words were ranked according to the conditional probability upon the extracted hidden topics, and a significance threshold was selected to eliminate the unimportant words. However, this method does not achieve an observable retrieval accuracy improvement; and we hypothesize that it is due to the pruning based on the conditional probabilities only without considering the relative significance among the words.

We present an unsupervised ranking algorithm to prune the vocabulary to improve the BOF-based image retrieval. We suggest that the hidden topics should normally not be of equal importance, and the words should not be equally linked to each topic. We thus propose to model the mutually reinforced relationship between the visual words and hidden topics to calculate their significance values. The proposed method was evaluated on the publicly available Early Lung Cancer Action Program (ELCAP) [34] database as a case study. The experimental results showed that it can improve image retrieval accuracy.

# 2 Method

## 2.1 Dataset

In this study, ELCAP database, which contains 50 sets of low dose computed tomography (CT) images, was used for evaluation. A total of 379 slices are provided with the centroids of lung nodules annotated, which are divided into four different types based on their relative locations to the surrounding anatomical structures (e.g., pleural surfaces, vessels, etc. [35]): well-circumscribed (W-15%), vascularized (V-16%), juxta-pleural(J-30%) and pleural-tail (P-39%). Example images of the four type nodules are shown in Figure 1(a), with the nodules displayed in the center.

## 2.2 Method outline

The overall BOF-based retrieval with the proposed vocabulary pruning method is illustrated in Figure 1. The Scale Invariant Feature Transform (SIFT) descriptor is firstly extracted with each pixel in the area around the nodule centroid as the keypoint, so that both the nodule and surrounding anatomical structures are included. Next, we use k-means clustering to construct the vocabulary, and obtain the frequency histogram of the visual words for each image. The Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme is then applied followed by the L2 normalization on the frequency histogram matrix, i.e., a $N \times M$ matrix where $N$ is the number of images and $M$ is the size of vocabulary that is the number of clusters obtained by k-means clustering, and k-nearest neighbor (k-NN) method is used to perform the retrieval. In our proposed method, instead of using the entire vocabulary generated by k-means, we would like to prune the vocabulary by keeping the most useful words. In particular, as show in Figure 1(e), we use pLSA to extract a total of $K$ hidden topics and design a ranking method to compute the significance of each word. The words with higher significance values sv are reserved as the pruned vocabulary.

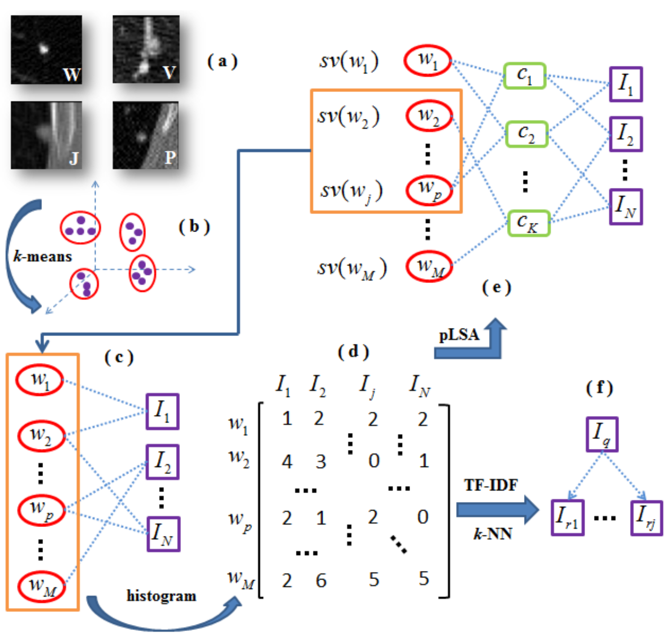


**Fig. 1.** Outline of BOF-based retrieval with the proposed vocabulary pruning method: (a) sample nodule images from ELCAP database; (b) vocabulary construction with k-means clustering; (c) feature assignment to visual words; (d) frequency histogram matrix; (e) hidden topic discovery and visual works ranking; (f) retrieval results.

### 2.3 Vocabulary pruning with pLSA

pLSA, which was originally used in linguistic studies, can be used to extract the hidden topics to bridge the semantic gap between documents (images) and words (visual words) [36–38]. It is a general model assuming that documents (images) can be interpreted by a set of hidden variables c, i.e., the hidden topics, each of which is the probability distribution upon the words (visual words). Given the document-word co-occurrence matrix, i.e., the frequency histogram matrix, the conditional probability of the words upon each hidden topic, $p(w|c)$, can be learned (see [39] for details).

The hidden topics are object categories [40] describing the common characteristics of different ROIs, e.g., nodules, pleural surfaces, and vessels in lung nodule images, so that the words would be more meaningfully linked to the ROIs than the individual images. Therefore, the conditional probability can be used to measure the significance / meaningfulness of the visual words. For a given hidden topic c, we consider the word w meaningless if its conditional probability $p(w|c)$ is below a certain significance threshold [32]. A word is removed for vocabulary pruning if it is meaningless to all topics.

### 2.4 Vocabulary pruning with the proposed ranking method

While the conditional probability $p(w|c)$ provides an effective criterion to evaluate the significance of visual words, it is insufficient to perform vocabulary pruning with the above scheme. Firstly, the extracted topics are not equally important, and the words linked to the more important topics should have higher significance values. Secondly, the words should be evaluated based on the overall relationship with all topics rather than individual ones, which means a word might be removed even if it is meaningful for some topics especially if these topics are not important. Based on these motivations, we propose a ranking method to calculate the significance value $sv$ of the hidden topics and visual words by analyzing the overall mutual interactions between them. This is the main difference from [6], which uses $p(w|c)$ directly as the significance value.

Our method is based on the underlying algorithm that the significance values of topics and words are calculated conditioned on each other. This means that, a topic c with higher significance value $sv(c)$ tends to connect with words of higher significance, and a word w with higher significance value $sv(w)$ tends to connect with topics of higher significance. This mutual relationship can be formulated as:

$$sv(c_q) = \sum_{w_p \in L(c_q)} sv(w_p) \tag{1}$$

$$sv(w_p) = \sum_{c_q : w_p \in L(c_q)} sv(c_q) \tag{2}$$

where $L(c)$ is the list of words that are meaningful for topic $c$, which is obtained according to the conditional probability as described in Section 3.

Next, the significance values are updated iteratively based on Eqs. (1) and (2), as shown in Figure 2. During each iteration, the significance values of all words and topics are calculated based on each other from the overall perspective.

This helps to determine the significance of a word based on all topics collectively, and the significances of various topics can be differentiated based on the related words. Across the iterations, the significance of a certain word, e.g., $sv(w_p)$, is diffused to the topics at the current iteration and gathered at the next iteration for updating the other words. This helps to encode the mutual relationship into the significance values in an iterative manner. Based on the experiments, we observe that the significance values tend to converge with more iterations. 20 iterations were chosen to balance between efficiency and performance.

**Inputs**:

$L$ for each topic, number of iteration $T$

**Outputs**:

$sv$ of each word and each topic

**Steps**:

Initialize $sv_0(\boldsymbol{w})=1$ and $sv_0(\boldsymbol{c})=1$.<br>
for $t = 1 : T$<br>
  for $q = 1 : K$<br>
    Compute $sv_t(c_q)$ based on $sv_{t\text{-}1}(\boldsymbol{w})$ using Eq.(1);<br>
  for $p = 1 : M$<br>
    Compute $sv_t(w_p)$ based on $sv_{t\text{-}1}(\boldsymbol{c})$ using Eq.(2);<br>
  Normalize $sv_t(\boldsymbol{c})$ and $sv_t(\boldsymbol{w})$;

**Return**:

$sv_T(\boldsymbol{w})$ and $sv_T(\boldsymbol{c})$

**Fig. 2.** The pseudo code of significance value computation for words and topics. Bold $\boldsymbol{w}$ and $\boldsymbol{c}$ represent the sets of words and topics, respectively.

Finally, all words are ranked according to their significance values obtained at the final iteration, and vocabulary pruning is performed by removing the words below a significance threshold $th \in (0, 100)$, i.e., the top $th\%$ words are reserved.

## 3 Experiments

In the experiments, we conduct leave-one-case-out cross-validation and compute the average retrieval accuracy (recall) of all queries to evaluate the performance. To obtain a fair comparison, we selected the first 100 descriptors near the centroid for each image producing the same number of descriptors per image, and extracted vocabularies with different sizes from 500 to 2000 with an interval of 100. The list of meaningful words for each topic consisted of the top 10% with higher conditional probabilities, which generated the best performance in general.

We first discuss the effect of the pruned vocabulary. Figure 3 shows the average accuracy of the first four retrieved items over different percentages of words reserved. The statistics generated by extracting different numbers of topics

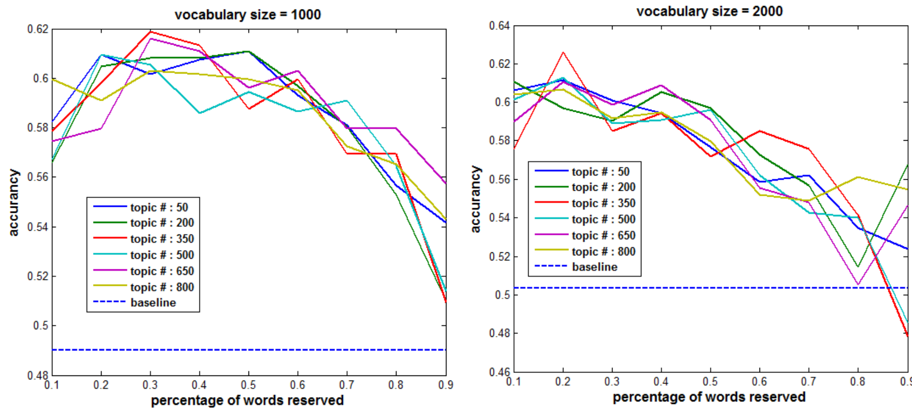


**Fig. 3.** Evaluation of the pruned vocabularies obtained by reserving different percentages of high significance value words in the original vocabularies. The curves show the accuracy distribution over various topic numbers, and the result from baseline (standard BOF model) is also given.

at two different original vocabulary sizes (1000 and 2000) are displayed. The standard BOF approach on the original vocabulary is regarded as the baseline. It can be observed that considerable improvements were obtained by pruning the vocabulary. Typically, the best performance was achieved when 60% to 80% of the words were pruned from the overall perspective.

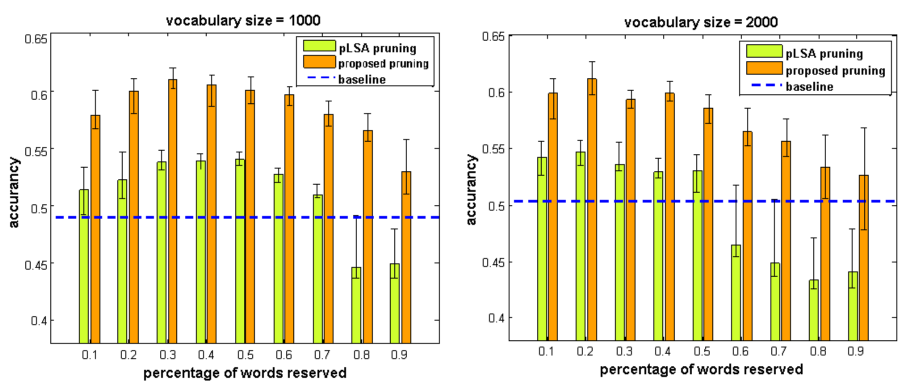


**Fig. 4.** Comparison with pLSA pruning approach. The bars indicate the average accuracy, and the error bars show the lowest and highest.

Figure 4 shows the comparisons with vocabulary pruning using pLSA only. For each pruned vocabulary, the average, minimum and maximum of retrieval accuracy are displayed. The words in the pLSA approach were reserved accord-

**Table 1.** Average retrieval results (varying original vocabulary sizes and hidden topic numbers) regarding various numbers of outputs from the baseline, pLSA pruning and proposed approaches.

| Output number | Baseline | pLSA pruning | Proposed pruning |
|---|---|---|---|
| 1 | 0.630 ± 0.129 | 0.675 ± 0.076 | **0.731 ± 0.066** |
| 2 | 0.687 ± 0.034 | 0.680 ± 0.043 | **0.720 ± 0.017** |
| 3 | 0.566 ± 0.031 | 0.579 ± 0.045 | **0.639 ± 0.025** |
| 5 | 0.454 ± 0.027 | 0.488 ± 0.052 | **0.546 ± 0.037** |
| 8 | 0.392 ± 0.025 | 0.427 ± 0.057 | **0.481 ± 0.046** |
| 10 | 0.370 ± 0.019 | 0.403 ± 0.058 | **0.456 ± 0.050** |
| 15 | 0.342 ± 0.022 | 0.366 ± 0.056 | **0.417 ± 0.051** |
| 20 | 0.328 ± 0.023 | 0.346 ± 0.052 | **0.395 ± 0.050** |

ing to the conditional probability upon the topics. The average accuracies across all pruned vocabularies of pLSA pruning are 0.5094 and 0.4969, which are similar to that of the baseline, which are 0.4901 and 0.5033 respectively. This is in accordance with the finding in [6] that pLSA can be used to reduce the vocabulary but with no obvious effect on the retrieval accuracy. Using our approach, the retrieval performances were 0.5843 and 0.5738 on average with about 8% improvement over the baseline, which suggests the advantage of the proposed ranking-based significance value computation method.

The overall performance of the proposed method regarding various numbers of retrieval output is given in Table 1. The average accuracy and standard deviation across all dictionaries (from 500 to 2000), and topics (from 50 to 800) are listed. Overall, the proposed pruning method outperforms the standard BOF and pLSA pruning by about 8% and 10%, respectively.

## 4 Conclusions and future work

We propose an unsupervised ranking-based vocabulary pruning method, which improves the performance of BOF-based image retrieval. The experimental results on lung nodule image retrieval show that the proposed method can identify the most meaningful visual words to describe the image content so that the retrieval quality is significantly enhanced even the vocabulary is pruned significantly. The reduction of the vocabulary leads to a low-dimensional feature representation, which reduces the computational cost and is more applicable to large scale data analysis.

The method is currently used on medical image analysis, and we would expect a better performance if a customized BOF model is used, e.g., more sophisticated feature design and better regions of interest detection. In addition, we are currently extending the proposed method to general image analysis due to the domain independent characteristic.

## 5 Acknowledgments

This work was supported in part by ARC grants.

## References

1. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349-1380 (2000)
2. R. Torres and A. Falcao: Content-based image retrieval: Theory and applications. Revista de Informática Teórica e Aplicada, 13(2):161-185 (2006)
3. S. Zhang, M. Yang, T. Cour, K. Yu and D. Metaxas: Query Specific Rank Fusion for Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
4. H. Müller, N. Michoux, D. Bandon and A. Geissbuhler: A Review of Content-based Image Retrieval Systems in Medical Applications Clinical Benefits and Future Directions. International Journal of Medical Informatics, 73(1):1-23 (2004)
5. W. Cai, J. Kim and D. Feng: Content-based Medical Image Retrieval. Biomedical Information Technology, Chapter 4, 83-113 (2008)
6. A. Kumar, J. Kim, W. Cai, M.J. Fulham and D. Feng: Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data. Journal of Digital Imaging, 26(6):1025-1039 (2013)
7. Y. Song, W. Cai, S. Eberl, M.J. Fulham and D. Feng: Discriminative Pathological Context Detection in Thoracic Images based on Multi-level Inference. The 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 191-198 (2011)
8. S. Liu, W. Cai, L. Wen and D. Feng: Multi-channel Brain Atrophy Pattern Analysis in Neuroimaging Retrieval. IEEE International Symposium on Biomedical Imaging (ISBI), 206-209 (2013)
9. C.B. Akgül, D.L. Rubin, S. Napel, C.F. Beaulieu, H. Greenspan and B. Acar: Content-based image retrieval in radiology: current status and future directions. Journal of Digital Imaging 24, 208-222 (2011)
10. Y. Song, W. Cai, H. Huang, Y. Wang and D. Feng: Object Localization in Medical Images based on Graphical Model with Contrast and Interest-Region Terms. The 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Medical Computer Vision (2012)
11. S. Liu, S.Q. Liu, S. Pujol, R. Kikinis, D. Feng, W. Cai: Propagation graph fusion for multi-modal medical content-based retrieval. The 13th International Conference on Control, Automation, Robotics and Vision (ICARCV) (2014)
12. Y. Song, W. Cai, S. Eberl, M.J. Fulham and D. Feng: Thoracic Image Case Retrieval with Spatial and Contextual Information. IEEE International Symposium on Biomedical Imaging (ISBI), 1885-1888 (2011)
13. X. Zhang, W. Liu, M. Dundar, B. Sunil and S. Zhang: Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval. IEEE Transactions on Medical Imaging (2014)
14. W. Cai, D. Feng and R. Fulton: Content-Based Retrieval of Dynamic PET Functional Images. IEEE Transactions on Information Technology in Biomedicine, 4(2):152-158 (2000)

15. H. Che, S. Liu, W. Cai, S. Pujol, R. Kikinis and D. Feng: Co-neighbor Multi-view Spectral Embedding for Medical content-based Retrieval. IEEE International Symposium on Biomedical Imaging (ISBI), 911-914 (2014)
16. Y. Song, W. Cai, Y. Zhou, M.J. Fulham and D. Feng: Volume-of-interest retrieval for PET-CT images with a conditional random field alignment. The Journal of Nuclear Medicine, 55 (Supplement 1):2065 (2014)
17. S. Liu, W. Cai, L. Wen, D. Feng, S. Pujol, R. Kikinis, M.J. Fulham, S. Eberl: Multi-channel neurodegenerative pattern analysis and its application in Alzheimer's disease characterization. Computerized Medical Imaging and Graphics, 38 (4), 436-444 (2014)
18. Y. Song, W. Cai, S. Eberl, M.J. Fulham and D. Feng: A Content-based Image Retrieval Framework for Multi-Modality Lung Images. IEEE International Symposium on Computer-Based Medical System (CBMS), 285-290 (2010)
19. S. Haas, R. Donner, A. Burner, M. Holzer and G. Langs: Superpixel-based Interest Points for Effective Bags of Visual Words Medical Image Retrieval. Second MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MCBR-CDS), 58-68 (2012)
20. Y. Song, W. Cai, Y. Zhou, L. Wen and D. Feng: Pathology-centric Medical Image Retrieval with Hierarchical Contextual Spatial Descriptor. IEEE International Symposium on Biomedical Imaging (ISBI), 202-205 (2013)
21. Y. Song, W. Cai, S. Eberl, M.J. Fulham and D. Feng: Structure-Adaptive Feature Extraction and Representation for Multi-Modality Lung Images Retrieval. The International Conference on Digital Image Computing: Techniques and Applications (DICTA), 152-157 (2010)
22. J. Yang, Y. G. Jiang, A. G. Hauptmann and C.W. Ngo: Evaluating Bag-of-visual-words Representations in Scene Classification. Proceedings of the International Workshop on Multimedia Information Retrieval, 197-206 (2007)
23. Y. Song, W. Cai and D. Feng: Hierarchical Spatial Matching for Medical Image Retrieval: The Annual ACM International Conference on Multimedia Workshop on Medical Multimedia Analysis and Retrieval (ACM MMAR), 1-6 (2011)
24. S. Liu, W. Cai, Y. Song, S. Pujol, R. Kikinis and D. Feng: A Bag of Semantic Words Model for Medical Content-based Retrieval. The 16th International Conference on MICCAI Workshop on Medical Content-Based Retrieval for Clinical Decision Support (2013)
25. Y. Song, W. Cai, S. Eberl, M.J. Fulham and D. Feng: Thoracic Image Matching with Appearance and Spatial Distribution. The 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 4469-4472 (2011)
26. R. Arandjelovic and A. Zisserman: All about VLAD. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1578-1585 (2013)
27. D. Qin, S. Gammeter, L. Bossard, T. Quack and L. Van Gool: Hello Neighbor: Accurate Object Retrieval with K-reciprocal Nearest Neighbors. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 777-784 (2011)
28. W. Cai, F. Zhang, Y. Song, S. Liu, L. Wen, S. Eberl, M.J. Fulham and D. Feng: Automated Feedback Extraction for Medical Imaging Retrieval. IEEE International Symposium on Biomedical Imaging (ISBI), 907-910 (2014)
29. J. Sivic and A. Zisserman: Video Google: A Text Retrieval Approach to Object Matching in Videos, IEEE International Conference on Computer Vision (ICCV), 1470-1477 (2003)

30. S. Liu, W. Cai, L. Wen, S. Eberl, M.J. Fulham, D. Feng: A robust volumetric feature extraction approach for 3D neuroimaging retrieval. IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBS), 5657-5660 (2010)
31. W. Cai, S. Liu, Y. Song, S. Pjuol, R. Kikinis, D. Feng: A 3D Difference-of-Gaussian based lesion detector for brain PET. IEEE International Symposium on Biomedical Imaging (ISBI), 677-680 (2014)
32. A. Foncubierta-Rodríguez, A. G. S. d. Herrera and H. Müller: Medical Image Retrieval using Bag of Meaningful Visual Words: Unsupervised Visual Vocabulary Pruning with pLSA. Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare, 75-82 (2013)
33. M. Bilenko, S. Basu and R. J. Mooney: Integrating Constraints and Metric Learning in Semi-supervised Clustering. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML), 11-18 (2004)
34. ELCAP Public Lung Image Database. Available: http://www.via.cornell.edu/databases/lungdb.html
35. S. Diciotti, G. Picozzi, M. Falchini, M. Mascalchi, N. Villari and G. Valli: 3-D Segmentation Algorithm of Small Lung Nodules in Spiral CT Images. IEEE Transactions on Information Technology in Biomedicine, 12(1):7-19 (2008)
36. U. Castellani, A. Perina, V. Murino, M. Bellani, G. Rambaldelli, M. Tansella and P. Brambilla: Brain Morphometry by Probabilistic Latent Semantic Analysis. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 177-184 (2010).
37. A. Cruz-Roa, F. Gonzalez, J. Galaro, A. Judkins, D. Ellison, J. Baccon, A. Madabhushi and E. Romero: A Visual Latent Semantic Approach for Automatic Analysis and Interpretation of Anaplastic Medulloblastoma virtual slides, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 157-164 (2012)
38. F. Zhang, Y. Song, W. Cai, M-Z. Lee, Y. Zhou, H. Huang, S. Shan, M.J. Fulham and D. Feng: Lung Nodule Classification With Multi-level Patch-based Context Analysis. IEEE Transactions on Biomedical Engineering, 61(4):1155-1166 (2014)
39. T. Hofmann: Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50-57 (1999)
40. A. Bosch, A. Zisserman and X. Munoz: Scene Classification via pLSA. European Conference on Computer Vision (ECCV), 517-530 (2006)