Text- and Content-based Medical Image Retrievals in the VISCERAL Retrieval Benchmark

Fan Zhang[1], Yang Song[1], Weidong Cai[1], Adrien Depeursinge[2], and Henning Müller[2]

[1] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, NSW, Australia
[2] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

Abstract. Text– and content-based retrievals are the most widely used approaches for medical image retrieval. They capture the similarity between images from different perspectives: text–based methods rely on manual textual annotations or captions associated with images; content-based approaches are based on the visual content of the images themselves such as colors and textures. Text-based retrieval can better meet the high-level expectations from humans but is limited by the time-consuming annotations. Content-based retrieval can automatically extract the visual features for high-throughput processing; however, its performance is less favorable than the text-based approaches due to the gap between low-level visual features and high-level human expectations. In this Chapter, we present the participation from our joint research team of USYD/HES-SO in the VISCERAL retrieval task. Five different methods are introduced, of which two are based on the anatomy-pathology terms, two are based on the visual image content, and the last one is based on the fusion of the aforementioned methods. The comparison results given the different methods indicated that the text-based methods outperformed the content-based retrieval and the fusion of text and visual content generated the best performance overall.

1 Introduction

Medical image data produced has been growing rapidly in quantity, content and dimension, due to an enormous increase in the number of diverse clinical exams performed in digital form and to the large range of image modalities and protocols available [1–5]. Retrieving a set of images that are clinically relevant to the query from a large image database has been the focus of medical research and clinical practice [6–9]. The relevance between images is normally computed in two manners, i.e., text- and content-based. The text-based approach is performed given the manual clinical / pathological descriptions, which require that the experts manually index the images with alphanumerical keywords if no text is already available with the images. The content-based retrieval is based on the image visual content information, which automatically extracts the rich visual

properties / features to characterize the images [10–12]. While the text-based retrieval is the more common method, the content-based approach is attracting more interests due to the fact that medical image data have expanded rapidly in the past two decades [13, 15–17]. The combination of the two approaches suggests a potential direction of medical image retrieval on performance improvements [18].

In the VISCERAL retrieval benchmark [20], we conducted medical image retrieval based on multimodal and multidimensional data. The similarities between medical cases are computed based on extracts of the medical records, radiology images and radiology reports. The VISCERAL retrieval dataset consists of 2311 volumes originated from three different modalities of CT, MRT1 and MRT2. The volumes are from different human body regions such as the abdomen, thorax and the whole body. Within the whole dataset, 1815 volumes are provided with anatomy-pathology terms extracted from the radiologic reports. A total of 10 topics were used as queries. Each of them was annotated with the 3D bounding box of the region of interest (ROI), binary mask of the main organ affected and the corresponding anatomy-pathology terms. A brief introduction of our participation has been reported in [19] and more on the VISCERAL data in general and the evaluation approach can be found in [20]. Our experimental results are reported with text-based retrieval that utilized the anatomy-pathology terms, with visual content-based retrieval that made use of the visual content features, and with information fusion that combined the above results.

The structure is as follows: in Section 2, we introduce the text, visual content and fusion retrieval methods that were used in our participation; in Section 3, we present the experimental results and discussion; and we provide the conclusion in Section 4.

2 Methods

A general framework of image retrieval consists of the following steps [13, 14]: feature extraction, similarity calculation, and relevance feedback, as illustrated in Fig. 1. For our methods, the feature extraction is conducted by analyzing the anatomy-pathology term (Sections 2.1 and 2.2) and the image content information (Section 2.3). The similarity is computed by measuring the Euclidean distance between the feature vectors. The relevance feedback is extracted based on the neighborhood information among the cases for retrieval result refinement (Section 2.4).

2.1 Term Weighting Retrieval

Medical image retrieval is conventionally performed with text-based approaches, which rely on manual annotation with alpha-numerical keywords. The anatomy-pathology term files provided in the VISCERAL retrieval benchmark lists the pathological terms and affected anatomies were extracted from the German radiology reports and mapped to RadLex. The co-occurence of different anatomy-

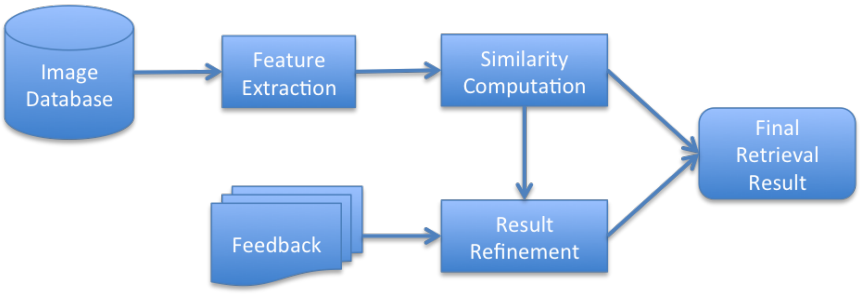

Fig. 1. Image retrieval framework.

pathology terms on the same cases can be used to evaluate the terms' effectiveness of finding the similarity between subjects, for example of some 'stop words' that occur widely but have little influence on describing the similarities. Our text-based methods are based on the co-occurrence matrix between the terms and cases.

For our first text-based method, we used term frequency–inverse document (case) frequency (TF-IDF) [21] to weight the terms for each case. TF-IDF can find the rare terms that carry more information than frequent ones, and is thus widely applied in term weighting problems. Formally, a case-term co-occurrence matrix $OCC_{NT\times NC}$ is constructed according to the anatomy-pathology terms on different cases, where the element $occ(t, c)$ refers to the number of occurrences of term T_t on case C_c, NC is the number of cases and NT is the number of terms. Term frequency $TF(t, c)$ evaluates the frequency of the term T_t occurred on the case C_c, which is

$$TF(t,c) = \frac{occ(t,c)}{\sum_{t\in[1,NT]} occ(t,c)}. \tag{1}$$

Inverse document (case) frequency $IDF(t)$ indicates whether the term T_t is common or rare across all cases, which is

$$IDF(t) = \log(\frac{\sum_{c\in[1,NC]} occ(t,c)}{1+occ(t,c)}). \tag{2}$$

TF-IDF measure of T_t for C_c is then computed, as

$$TF\text{-}IDF(t,c) = TF(t,c) * IDF(t). \tag{3}$$

Case C_c is finally formulated as a vector of TF-IDF measures of all terms, as

$$V_{TF\text{-}IDF}(c) = \{TF\text{-}IDF(t,c)|t \in [1,NT]\}. \tag{4}$$

The Euclidean distance between the vectors are then computed, followed by a k-NN method for retrieval.

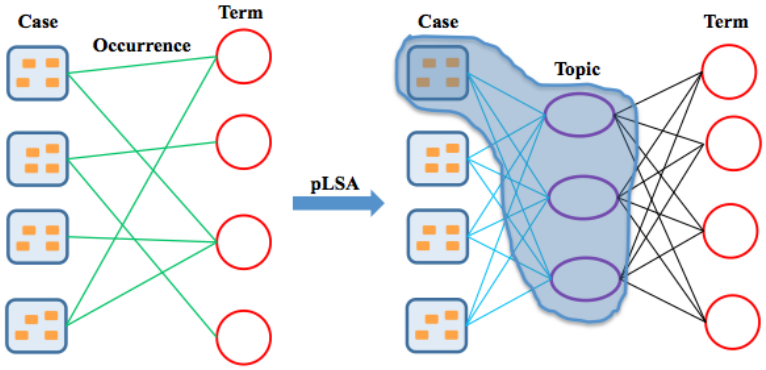

Fig. 2. Latent topic generation with pLSA.

2.2 Semantics Retrieval

While the TF-IDF method merely utilizes the direct co-occurrence relationship between the terms and cases, this relationship can be further used to infer the semantic information and can provide a more discriminative description of these terms for similarity computation. The latent semantic topic model is one of the most representative methods that can automatically extract the semantic information based on the co-occurrence relationship. It assumes that each image can be considered as a mixture of latent topics and the latent topic is a probability distribution of terms. In this study, we applied probabilistic Latent Semantic Analysis (pLSA)[22], which is a widely used latent topic extraction technique, for learning the latent semantics.

The schema of pLSA is shown in Fig. 2. pLSA considers that the observed probability of a term T_t occurring on a case C_c can be expressed with a latent or unobserved set of latent topics $Z = \{z_h | h \in [1, H]\}$ where H is the number of latent topics, as:

$$P(t|c) = \sum_h P(t|z_h) \cdot P(z_h|c). \tag{5}$$

The probability $P(z_h|c)$ describes the distribution of latent topics given a certain case. The latent topics Z can be learnt by fitting the model with Expectation-Maximization (EM) [23] algorithm that maximizes the likelihood function L:

$$L = \prod_t \prod_c P(t|c)^{occ(t,c)}. \tag{6}$$

After the latent topic extraction, each case is represented as the probability vector of the extracted latent topics,

$$V_{pLSA}(c) = \{P(z_h|c) | h \in [1, H]\}, \tag{7}$$

where each element is the probability of the latent topic given this case. The similarity between different cases is then measured by the Euclidean distance

between the probability vectors. During the experiments, 20 latent topics were used, i.e, $H = 20$.

2.3 BoVW Retrieval

Unlike the aforementioned text-based methods, the visual content-based retrieval computes the similarity between images based on their visual characteristics, such as the texture and color. In the literature, there are many methods that can automatically extract the visual features to characterize the medical images [2, 24–26]. The bag-of-visual-words (BoVW) [27, 28] method, which is one of the popular methods for visual content-based image retrieval, is applied as our first content-based retrieval method. The BoVW model represents an image with a visual word frequency histogram that is obtained by assigning the local visual features to the closest visual words in the dictionary. Rather than matching the visual feature descriptors directly, the BoVW-based approaches compare the images according to the visual words that are assumed to have higher discriminative power [27].

Specifically, the scale invariant feature transform (SIFT) [29] descriptors are extracted from the image to obtain a collection of local patch features for each image / case[1]. The entire patch feature set computed from all images in the database is then grouped into clusters, e.g., with k-means method. Each cluster is regarded as a visual word W and the whole cluster collection is considered as the visual dictionary $D = \{W_d | d \in [1, ND]\}$, where ND is the size of dictionary. Following that, all patch features in one image are assigned to the visual words, generating a visual word frequency histogram to represent this image (case) as,

$$V_{BoVW}(c) = \{fre(d, c) | d \in [1, ND]\}, \tag{8}$$

where $fre(d, c)$ is the frequency of visual word W_d on case C_c. Finally, the similarity between images is computed based on these frequency histograms for retrieval.

In our experiments, the SIFT [29] descriptors were extracted from each scan of the 3D volume from the axial view. A visual dictionary of size 100, i.e., $ND = 100$ that is sufficient for capturing local visual details and does not introduce too much noise, was computed with k-means. During the retrieval, given the ROI of a query case, we traversed all sub-regions (of the same size as the ROI) in a candidate volume. The sub-region that has the smallest Euclidean distance from the query ROI in terms of the visual word frequency histograms was regarded as the most similar area of the candidate to the query ROI. The distance between the two regions represented the similarity between the query and candidate images.

2.4 Retrieval Result Refinement

While the first two steps form a basic retrieval process, relevance feedback refines the retrieval results if the top-ranked items are not fully satisfactory. Relevance

[1] Image corresponds to the case as used in Sections 2.1 and 2.2.

feedback is based on the preferences upon the initial retrieval results, which can be provided by the users. However, providing manual feedback can be quite challenging due to the huge amount of image data. The relevance can also be affected since manual interpretation sometimes could be error-prone. The neighborhood among images on the other hand can be used as a form of relevance feedback and are expected to be beneficial for image retrieval.

Based on the results of the BoVW method, we further conducted a retrieval result refinement process based on our recent work [30]. In our method, we assume that the similarity relationship between the initial retrieved results and the remaining candidates can be used as relevance feedback for result refinement. For a given query image, we first get a ranked list of initial retrieval results based on the BoVW model. Then, the similarities between the retrieved items and all candidates are used to evaluate their *preference* and *relativity*.

Formally, a preference score $pref(C_{cr})$ for the retrieved item C_{cr} is defined to evaluate the preference upon C_{cr} regarding to the query, i.e., relevance and irrelevance. A relativity score $rel(C_{cc})$ is appointed to the candidate image C_{cc} indicating the similarity of C_{cc} to the query. The two values are computed conditioned on each other regarding the query case C_{cq}: the relativity score $rel(C_{cc})$ of C_{cc} would be high if it is similar to the highly preferred retrieved item C_{cr}, and the preference score $pref(C_{cr})$ of C_{cr} would be high if it is close to the more relevant candidate C_{cc}. The relativity score of C_{cc} is formulated as the sum of preference scores of its neighbouring retrieved items, similarly to the preference score of C_{cr}. Denoting rel and $pref$ are the vectors of relativity and preference scores, we have the following formulations:

$$pref(C_{cr}) = \sum_{C_{cc}:A(C_{cc},C_{cr})=1} rel(C_{cc}), \tag{9}$$

$$rel(C_{cc}) = \sum_{C_{cr}:A(C_{cc},C_{cr})=1} pref(C_{cr}), \tag{10}$$

where A is a matrix indicating the bipartite neighborhood relationship between retrieved items and candidates, i.e., $A(C_{cc}, C_{cr}) = 1$ if C_{cc} is the neighbour of C_{cr}; otherwise, $A(C_{cc}, C_{cr}) = 0$. Eqs. (9) and (10) can be alternatively solved iteratively as shown in Algorithm 1.

For our experiments, we selected the top 30 volumes based on the BoVW outputs as the initial results. Then, a bipartite relationship between the initial results and the remaining candidates, which represented the neighbourhood, was constructed by keeping the top 30 candidates for each initial result. The iterative ranking method [30] was applied to recompute the similarity score of each candidates with an iteration number T of 20.

2.5 Fusion Retrieval

It is often suggested that the combination of textual and visual features can improve the retrieval performance [18]. Many fusion strategies have been proposed

Algorithm 1 Pseudo code for preference and relativity computation

Input: Number of iterations T, neighborhood matrix A.
Output: Preference and relativity values.

initialize $rel_0 = 1$ and $pref_0 = 1$.
for each t in $[1, T]$ **do**
  for each C_{cr} **do**
    Compute $pref_t(C_{cr})$ based on $rel_{t-1}(C_{cc})$ using Eq.(9);
  end for;
  for each C_{cc} **do**
    Compute $rel_t(C_{cc})$ based on $pref_t(C_{cr})$ using Eq.(10);
  end for;
  $L2$-normalize $pref_t(C_{cr})$ and $rel_t(C_{cc})$
end for;
return $pref_T$ and rel_T.

in the past such as maximum combination, sum combination [32], and Condorcet fuse [33].

Given the results from the text- and content-based retrievals, we conducted the fusion retrieval by using the sum combination method, which has been effective for textual and visual feature fusion [31]. To do this, a normalization step was firstly incorporated to normalize the similarity scores obtained from the aforementioned results, as:

$$S' = \frac{S - S_{min}}{S_{max} - S_{min}}, \tag{11}$$

where S_{min} and S_{max} are the lowest and highest similarity scores obtained within a certain method. The sum combination was then adopted to compute a fusion score for each candidate, as:

$$S_F = \sum_{r \in [1,4]} S'_r, \tag{12}$$

where $r \in [1, 4]$ represents the first four methods. The ones with the higher scores were for the results of fusion retrieval.

3 Results and Discussion

To evaluate the performance of retrieval results, medical experts were invited to perform relevance assessment of the top ranked cases for each run. Various evaluation measures were used considering the top-ranked X cases, including: the precision for top-ranked 10 and 30 cases (P@10, P@30), mean uninterpolated average precision (MAP), bpref measure, and the R-precision.

Fig. 3 displays the retrieval result for each of the topics given the aforementioned measures. The performances were diverse across the cases. It can be generally observed that better results were obtained for topics 1 and 7 when

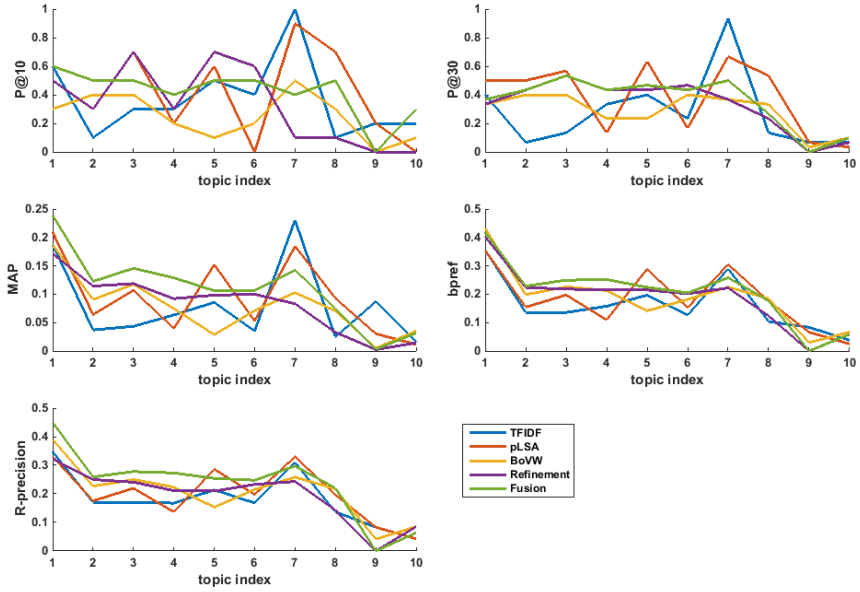

Fig. 3. Retrieval results of the 10 topics given different evaluation measures.

compared to the other topics but the results for topics 9 and 10 were unfavorable. The differences were due to the different affected regions. Our methods computed the similarity between cases using the entire volumes, instead of focusing on the local details. Therefore, for cases that have a small affected region, e.g., case 10, the similarity tended to be inaccurate.

Table 1 shows the average results of the measures across the 10 queries, with the first five rows from our resutls and the last three rows showing the best results from all participants of the VISCERAL retrieval benchmark. Within our text-based approaches, pLSA generated better performance when compared to the TF-IDF method, by further using the latent semantic information inferred from the co-occurrence relationship between cases and terms. Regarding the content-based retrieval, we obtained better results when applying the result refinement. Across the four methods, better performance was obtained from the text-based retrieval when compared to the content-based retrieval. The content-based methods use the visual content characteristics that may have large variation between the relevant cases but small difference between the irrelevant ones. The SIFT feature used in our experiments is widely known for capturing the local image content information but it sometimes can be hard for SIFT to recognize the subtle visual difference between different images. In addition, while the size of the dictionary was set to 100 in our experiments, it can be varied for different datasets and potentially affect the retrieval performance. The text-based approaches on the other hand compare the different cases directly based on the pathological terms and affected anatomies. Thus, the text-based retrieval ob-

Table 1. Average results of the different measures across the 10 queries.

	P@10	P@30	MAP	bpref
TFIDF	0.370	0.277	0.081	0.162
pLSA	0.410	**0.380**	0.094	0.183
BoVW	0.250	0.283	0.078	0.190
Refinement	0.330	0.330	0.083	0.188
Fusion	**0.420**	0.353	**0.110**	**0.207**
Text	0.570	0.497	0.194	0.322
Image	0.330	0.330	0.083	0.188
Mixed	0.688	0.638	0.283	0.340

tained the more favorable retrieval results. While the anatomy-pathology terms provide an overall description for the similarity computation, the visual content feature can better capture the local anatomical differences between cases. Therefore, the fusion approach achieved the overall best result, which is in accordance with findings in the literature. Regaring the comparisons across all VISCERAL retrieval benchmark participations, we had the best performance with the result refinement among all image based methods. The results from the text and fusion methods were less unfavorable since only co-occurrence information between the terms were used. Further analysis of the terms in the benchmark relating to the entire anatomy-pathology RadLex term collection would be helpful for retrieval improvements.

4 Conclusions

In this Chapter, we introduce the approaches from our joint research team of USYD/HES-SO to address the VISCERAL Retrieval task, including the TF-IDF and pLSA methods for text-based retrieval, the BoVW and its result refinement for content-based retrieval, and the fusion retrieval of the above methods. The experimental results are in accordance with finds in the literature, i.e., the text-based approaches typically perform better than purely visual content-based methods and the combination of text– and content–based retrieval can achieve improved retrieval performance.

References

1. Doi, K.: Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. Physics in Medicine and Biology 51, R5-R27 (2006)
2. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics 73, 1-23 (2004)
3. Cai, W., Feng, D., Fulton, R.: Content-Based Retrieval of Dynamic PET Functional Images. IEEE Transactions on Information Technology in Biomedicine 4(2), 152-158 (2000)

4. Song, Y., Cai, W., Eberl, S., Fulham, M.J., Feng, D.: Thoracic Image Case Retrieval with Spatial and Contextual Information. IEEE International Symposium on Biomedical Imaging (ISBI), 1885-1888 (2011)
5. Kumar, A., Kim, J., Cai, W., Fulham, M., Feng, D.: Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. Journal of Digital Imaging 26(6), 1025-1039 (2013)
6. Zhang, S., Yang, M., Cour, T.,Yu, K., Metaxas, D.: Query Specific Rank Fusion for Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(4), 803-815. (2014)
7. Müller, H., Antoine R., Arnaud G., Jean-Paul V., Antoine G.: Benefits of Content-based Visual Data Access in Radiology 1. Radiographics 25(3), 849-858 (2005)
8. Song, Y., Cai, W., Zhou, Y., Wen, L., Feng D.: Pathology-centric Medical Image Retrieval with Hierarchical Contextual Spatial Descriptor. IEEE International Symposium on Biomedical Imaging (ISBI), 202-205 (2013)
9. Song, Y., Cai, W., Eberl, S., Fulham, M.J., Feng, D.: A Content-based Image Retrieval Framework for Multi-Modality Lung Images. IEEE International Symposium on Computer-Based Medical System (CBMS), 285-290 (2010)
10. El-Naqa, I., Yang, Y., Galatsanos, N.P., Nishikawa, R.M., Wernick, M.N., 2004. A similarity learning approach to content-based image retrieval: ap- plication to digital mammography. IEEE Transactions on Medical Imaging 23, 1233–1244.
11. Zhang, F., Song, Y., Cai, W., Lee, M-Z., Zhou, Y., Huang, H., Shan, S., Fulham, M.J., Feng, D.: Lung Nodule Classification With Multi-level Patch-based Context Analysis. IEEE Transactions on Biomedical Engineering 61(4), 1155-1166 (2014)
12. Foncubierta-Rodríguez, A., Depeursinge, A., Müller, H.: Using multiscale visual words for lung texture classification and retrieval. Medical Content-Based Retrieval for Clinical Decision Support, 69-79 (2012)
13. Cai, W., Kim, J., Feng, D.: Content-based medical image retrieval. Elsevier. book section 4, 83–113 (2008)
14. Squire, D., Müller, W., Müller, H., Raki, J.: Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. The 11th Scandinavian Conference on Image Analysis, 143-149 (1999)
15. Müller, H., Deserno, T. M. : Content-based medical image retrieval. Biomedical Image Processing, Springer, 471-494 (2011)
16. Haas, S., Donner, R., Burner, A., Holzer, M., Langs, G.: 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)
17. Zhang, F., Song, Y., Cai, W., Hauptmann, A. G., Liu, S., Liu, SQ., Feng, D., Chen, M.: Ranking-based Vocabulary Pruning in Bag-of-Features for Image Retrieval. First Australasian Conference on Artificial Life and Computational Intelligence (ACALCI 2015), Lecture Notes in Artificial Intelligence 8955, 436-445, 2015.
18. Müller, H., Kalpathy-Cramer, J.: The ImageCLEF Medical Retrieval Task at ICPR 2010—Information Fusion to Combine Visual and Textual Information. Recognizing Patterns in Signals, Speech, Images and Videos, 99-108 (2010)
19. Zhang, F., Song, Y., Cai, W., Depeursinge, A., Müller, H.: USYD/HES-SO in the VISCERAL Retrieval Benchmark. The 37th European Conference on Information Retrieval (ECIR 2015) Workshop on Multimodal Retrieval in the Medical Domain (2015)

20. Hanbury, A., Müller, H., Langs, G., Weber, M. A., Menze, B. H., Fernandez, T.S.: Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. CLEF conference, Springer Lecture Notes in Computer Science (2012)
21. Jones, K. S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11-21 (1972)
22. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177-196 (2001)
23. Heinrich, G.: Parameter estimation for text analysis. Technical report (2005)
24. Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S.: Towards large scale histopathological image analysis: Hashing-based image retrieval. IEEE Transactions on Medical Imaging 34, 496–506 (2015)
25. Yang, W., Lu, Z., Yu, M., Huang, M., Feng, Q., Chen, W.: Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single- and multiphase contrast-enhanced CT images. Journal of Digital Imaging 25, 708–719 (2012)
26. Song, Y., Cai, W., Eberl, S., Fulham, M.J., Feng, D.: 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)
27. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. IEEE International Conference on Computer Vision (ICCV), 1470-1477 (2003)
28. Liu, S., Cai, W., Song, Y., Pujol, S., Kikinis, R., Feng, D.: 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).
29. Lowe, D. G.: Object recognition from local scale-invariant features. IEEE International Conference on Computer Vision (ICCV), 1150-1157 (1999)
30. Cai, W., Zhang, F., Song, Y., Liu, S., Wen, L., Eberl, S., Fulham, M., Feng, D.: Automated feedback extraction for medical imaging retrieval. IEEE International Symposium on Biomedical Imaging (ISBI), 907-910 (2014)
31. Zhou, X., Depeursinge, A., Müller, H.: Information Fusion for Combining Visual and Textual Image Retrieval. International Conference on Pattern Recognition (ICPR), 1590-1593 (2010)
32. Fox, E. A., Shaw, J. A.: Combination of multiple searches. Text REtrieval Conference, 243–252 (1993)
33. Montague, M., Aslam, J. A.: Condorcet fusion for improved retrieval. Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM), 538–548 (2002)