Multi-Modal Query Expansion Based on Local Analysis for Medical Image Retrieval*

Md Mahmudur Rahman, Sameer K. Antani, Rodney L. Long, Dina Demner-Fushman, and George R. Thoma

U.S. National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA
{rahmanmm,santani,rlong,ddemner,gthoma}@mail.nih.gov

Abstract. A unified medical image retrieval framework integrating visual and text keywords using a novel multi-modal query expansion (QE) is presented. For the content-based image search, visual keywords are modeled using support vector machine (SVM)-based classification of local color and texture patches from image regions. For the text-based search, keywords from the associated annotations are extracted and indexed. The correlations between the keywords in both the visual and text feature spaces are analyzed for QE by considering local feedback information. The QE approach can propagate user perceived semantics from one modality to another and improve retrieval effectiveness when combined in multi-modal search. An evaluation of the method on imageCLEFmed'08 dataset and topics results in a mean average precision (MAP) score of 0.15 over comparable searches without QE or using only single modality.

1 Introduction

Medical image retrieval based on multi-modal sources has been recently gaining popularity due the large amount of text-based clinical data available in the form of case and lab reports. Improvement in retrieval performances has been noted by fusing evidence from the textual information and the visual image content in a single framework. The results of the past ImageCLEFmed [1] tracks suggest that the combination of visual and text based image searches provides better results than using the two different approaches individually [1]. While there is a substantial amount of completed and ongoing research in both the text and content based image retrieval (CBIR) in medical domain [2, 3], much remains to be done to see how effectively these two approaches can complement each other in an integrated interactive framework based on query reformulation. s To increase

* This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and the Lister Hill National Center for Biomedical Communications (LHNCBC). We thank the CLEF [1] organizers for making the dataset available for the experiments.

[1] http://imageclef.org/2008/medical

the effectiveness and reduce the ambiguity due to the word mismatch problem in text information retrieval, a variety of query reformulation strategies based on term co-occurrence or term similarity have been investigated [4–7]. These techniques exploit term (keyword) dependency as term clustering in document collection based on either global or local analysis [5]. In a global analysis, all documents in the collection are analyzed to determine a global thesaurus-like structure that defines term relationships. This structure is then utilized to select additional terms for QE. In local analysis, the top retrieved documents for a query are examined at query time without any assistance from the user, in general to determine the terms for QE.

On the other hand, due to the nature of the low-level continuous feature representation in majority of the CBIR systems [8, 9], the idea of QE cannot be directly applied and is relatively new in this domain [10, 11]. For example, a context expansion approach has been recently explored in [10] by expanding the key regions of the (image) queries using highly correlated environmental regions according to an image thesaurus. In [11], the authors attempt to automatically annotate and retrieve images by applying QE in its relevance model based on a set of training images. Here, images are modeled with a *bag-of-concepts* (e.g., *bag-of-words* in text) approach of vector space model (VSM) in information retrieval [12]. These approaches are either data dependent over the entire collection or dependent on the associated keywords. In this paper, we explore a fundamentally different QE technique in a unified multi-modal framework, which is based on the correlation analysis of both visual and text keywords and relies only on the local feedback information. The aim of this framework is to develop methods that can combine and take advantage of complementary nformation from both modalities through application of cross-modal QE mechanism.

The proposed approach and an evaluation of its efficacy are presented as follows: in Section 2, we briefly describe the image representation approach in visual and text keyword spaces. In Section 3, we describe the proposed query expansion strategy based on local analysis. The experiments and analysis of the results are presented in Section 4 and finally Section 5 provides the conclusions.

2 Image Representation in Visual and Text Keyword Spaces

In a heterogeneous collection of medical images, it is possible to identify specific local patches that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in grey level radiological images, differential color and texture structures in microscopic pathology and dermoscopic images, etc. The variation in these local patches can be effectively modeled as visual keywords by using supervised learning based classification techniques, such as the support vector machine (SVM) [13]. In its basic formulation, the SVM is a binary classification method that constructs a decision surface and maximizing the inter-class boundary between the samples. A number of methods have been proposed for multi-class classification by solving many two-class problems and

combining their predictions. For visual keyword generation, we utilize one such voting-based multi-class SVM known as *one-against-one* or pairwise coupling (PWC) [14].

In order to perform the learning, a set of L labels are assigned as $C = \{c_1, \cdots, c_i, \cdots, c_L\}$, where each $c_i \in C$ characterizes a visual keyword. The training set of the local patches that are generated by a fixed-partition based approach and represented by a combination of color and texture moment-based features. For SVM training, the initial input to the system is the feature vector set of the patches along with their manually assigned corresponding concept labels. Images in the data set are annotated with visual keyword labels by fixed partitioning each image I_j into l regions as $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \Re^d$ is a combined color and texture feature vector. For each $\mathbf{x}_{k_j}$, the visual keyword probabilities are determined by the prediction of the multi-class SVMs as [14]

$$p_{ik_j} = P(y = i \mid \mathbf{x}_{k_j}),\ 1 \leq i \leq L. \tag{1}$$

Finally, the category label of x_{k_j} is determined as c_m, which is the label of the category with the maximum probability score. Hence, the entire image is thus represented as a two-dimensional index linked to the visual keyword labels. Based on this encoding scheme, an image I_j is represented as a vector of weighted visual keywords as

$$\mathbf{f}_j^I = [w_{1_j}, \cdots, w_{i_j}, \cdots w_{L_j}]^{\mathrm{T}} \tag{2}$$

where each w_{i_j} denotes the weight a visual keyword $c_i, 1 \leq i \leq L$ in image I_j, depending on its information content. The popular *"tf-idf"* term-weighting scheme [12] is used in this work, where the element w_{i_j} is expressed as the product of local and global weights.

For the text-based image search, it is necessary to transform the annotation files in XML formats into an easily accessible representation known as the *index*. In this case, information from only relevant tags are extracted and preprocessed by removing stop words that are considered to be of no importance for the actual retrieval process. Subsequently, the remaining words are reduced to their stems, which finally form the set $T = \{t_1, t_2, \cdots, t_N\}$ of index terms or keywords of the annotation files. Next, the annotation files (document) are modeled as a vector of keywords as

$$\mathbf{f}_j^D = [\hat{w}_{1_j}, \cdots, \hat{w}_{i_j}, \cdots \hat{w}_{N_j}]^{\mathrm{T}} \tag{3}$$

where each $\hat{w}_{i_j}$ denotes the *"tf-idf"* weight of a keyword $t_i, 1 \leq i \leq N$ in the annotation of image I_j.

3 Multi-modal QE Based on Local Analysis

Query expansion based on local feedback and cluster analysis has been one of the most effective methods for expanding queries in text retrieval domain [4, 5, 7]. Generally, this approach expands a query based on the information from the top retrieved documents for that query without any assistance from the user. The correlated terms are identified and ranked in order of their potential

contribution to the query and are re-weighted and appended to the query [4]. Before presenting our query expansion method, some basic terminologies need to be defined as follows:

Definition 1. *Let us consider q as a multi-modal query, which has an image part as I_q and a text part as D_q. The similarity between q and a multi-modal item j, which also has also two parts (e.g., image I_j and context D_j), is defined as*

$$Sim(q, j) = \omega_I Sim_I(I_q, I_j) + \omega_D Sim_D(D_q, D_j) \tag{4}$$

Here, ω_I and ω_D are normalized inter-modality weights within the text and image feature spaces. In this framework, the individual image $\mathrm{Sim_I}(I_q, I_j)$ and text $\mathrm{Sim_D}(D_q, D_j)$ based similarities are computed based on the Cosine distance measure [12].

Definition 2. *For the given query q, the set S_l of retrieved images along with associated annotations is called the* local image set. *Also, the set $C_l \subseteq C$ of all distinct visual keywords $c_i \in C_l$ and the set $T_l \subseteq T$ of all distinct text keywords $t_i \in T_l$ in the local image set S_l is called the* local vocabulary of visual and text keywords *respectively.*

Since, the correlated terms for an expansion are those present in the local cluster, we first need to generate such clusters from S_l and thereafter from C_l and T_l. To generate the cluster, we rely on a local correlation matrix that is built based on the co-occurrence of keywords inside images and associated annotations. Let $\mathbf{A_l^I} = [a_{uv}]$ be a $|C_l| \times |C_l|$ matrix in which the rows and columns are associated with the visual keywords in C_l. Each entry a_{uv} expresses a normalized correlation factor between visual keywords c_u and c_v as

$$a_{uv} = n_{uv}/(n_u + n_v - n_{uv}) \tag{5}$$

where n_u be the number of images in S_l that contain the keyword c_u, n_v be the number of images that contain the keyword c_v, and n_{uv} be the number of the top retrieved images in S_l that contain both keywords. The entry a_{uv} measures the ratio between the number of images where both c_u and c_v appear and the total number of images in S_l where either c_u or c_v appear and its value ranges to $0 \leq a_{uv} \leq 1$. If c_u and c_v have many co-occurrences in images, the value of a_{uv} increases and the images are considered to be more correlated. In a similar fashion, we can generate a $|D_l| \times |D_l|$ matrix $\mathbf{A_l^D}$ in which the rows and columns are associated with the keywords in T_l. The global version of this matrix, which is termed as the *connection matrix*, is utilized in a fuzzy information retrieval approach in [16].

By generating the above matrices, we can use them to build local correlation clusters for the multi-modal query expansion. Let $f_i(n)$ be a function that takes the i-th row and return the ordered set of n largest values a_{ij} from $\mathbf{A}_l^I$ ($\mathbf{A}_l^D$), where j varies over the set of keywords and $i \neq j$. Then $f_i(n)$ defines a local correlation cluster around the visual keyword c_i (t_i) as shown as a blue star in Figure 1). Here, the concept c_j (t_j) is located within a neighborhood $f_i(n)$ associated with the concept c_i (t_i) as shown as the red stars in Figure 1).

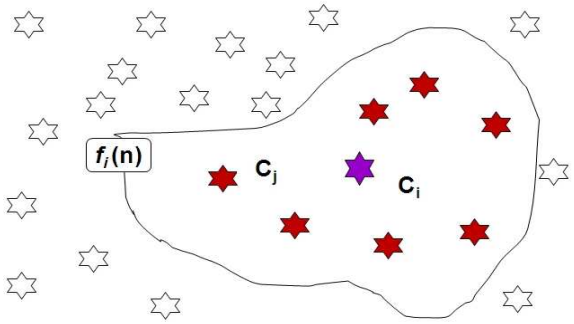

Fig. 1. Visual keyword c_j as a neighbor of the keyword c_i based on a local cluster

Now, the keywords that belong to clusters associated with the query q can be used to expand the original query. Often these neighbor concepts are correlated by the current query context [12]. The steps of the query expansion process for visual keywords based on the correlation cluster on $\mathbf{A}_l^I$ are given in Algorithm 1. Similar steps are also applied for the textual query expansion based on the correlation cluster on $\mathbf{A}_l^D$.

Algorithm 1 Query Expansion through Local Analysis

1: Initialize a temporary expanded query vector of query image I_q as $\mathbf{f}_q^{\mathrm{e}} = [\hat{w}_{1_q}\ \hat{w}_{2_q} \cdots \hat{w}_{i_q} \cdots \hat{w}_{L_q}]^{\mathrm{T}}$ where each $\hat{w}_{i_q} = 0$.
2: For an original query vector $\mathbf{f}_q^{\mathrm{o}} = [w_{1_q}\ w_{2_q} \cdots w_{i_q} \cdots w_{L_q}]^{\mathrm{T}}$ of I_q, perform initial retrieval.
3: Consider, the top ranked K images as the local image set S_l and construct the local correlation matrix $\mathbf{A}_l^I$ based on equation (5)
4: **for** $i = 1$ to L **do**
5: **if** $w_{i_q} > 0$ **then**
6: Consider the i-th row in $\mathbf{A}_l^I$ for the visual keyword c_i.
7: Return $f_i(n)$, the ordered set of n largest values m_{ij}, where $i \neq j$, therefore $c_j \in C_l - \{c_i\}$.
8: **for** each c_j **do**
9: Add and re-weight the corresponding element in query vector as $\hat{w}_{j_q} += w_{i_q} - ((w_{i_q} - 0.1) \times k/n)$, where k is the position of c_j in the rank order.
10: **end for**
11: **end if**
12: **end for**
13: Obtain the re-formulated or modified query vector as $\mathbf{f}_q^{\mathrm{m}} = \mathbf{f}_q^{\mathrm{e}} + \mathbf{f}_q^{\mathrm{o}}$.
14: Perform the image-based search with the modified query vector $\mathbf{f}_q^{\mathrm{m}}$.
15: Continue the process, i.e., steps 3 to 14 until no more changes are noticed.

Based on the Step 9 of the Algorithm 1, weights are assigned in such a way that a top ranked keyword in a ordered set gets the largest weight value than the

next one in the set. After this expansion process, new keywords may have been added to the original query based on the Step 9 and the weight of an original query concept may have been modified had the concept belonged to the top ranked concepts based on the Step 13 of the algorithm.

Since, a query can be represented with both visual and text keywords, it can be initiated either by a keyword-based search or by a image-query-by-example (QBE) search. The query expansion approach can be used to automatically expand a textual query using related keywords obtained from top retrieved (based on user's feedback) annotation files of associated images. In a similar fashion, the visual QBE can be reformulated using visual keywords from top retrieved relevant images based on a CBIR or text search in the previous iteration. The flexibility in such a search process implicitly creates a semantic network that links text keywords with visual keywords and vice versa.

4 Experiments and Results

To evaluate the retrieval effectiveness, experiments are performed in a benchmark medical image collection from ImageCLEFmed'08 [1]. This collection contains more than 67,000 images of different modalities from the RSNA journals[2] Radiology and Radiographics. For each image, the text of the figure caption is supplied as free text. In some cases, however, the caption is associated with a multi-part image. The contents of this collection represent a broad and significant body of medical knowledge, which makes the retrieval more challenging.

The proposed methods are evaluated on the 30 query topics developed by imageCLEFmed organizers. Each topic is a short sentence or phrase describing the search topic with one to three "relevant" images. The query topics are equally subdivided into three categories: visual, mixed, and semantic [1]. On completion of the imageCLEFmed'08 task, a set of relevant results for all topics was ere created by considering top retrieval results of all submitted runs of the participating groups. Retrieval results are evaluated using uninterpolated (arithmetic) *Mean Average Precisions (MAP)* and *Precision* at rank 20 (P 20).

For the visual keyword generation based on SVM learning, 30 local concept categories are manually defined, such as tissues of lung or brain of CT or MRI, bone of chest, hand, or knee X-ray, microscopic blood or muscle cells, dark or white background, etc. The training set consists of less then 1% images of the entire collection. Each image in the training set is partitioned into an 8×8 grid generating 64 non-overlapping regions, which is proved to be effective to generate the local patches. Only the regions that conform to at least 80% of a particular concept category are selected and labeled with the corresponding category label due to the consideration of robustness to noise. For the SVM training, we utilized the radial basis function (RBF). A 10-fold cross-validation (CV) is conducted to find the best tunable parameters C and γ of the RBF kernel. After finding the best values of the parameters $C = 200$ and $\gamma = 0.02$ of the RBF kernel with a CV

[2] http://www.rsnajnls.org

accuracy of 81.01%, they are utilized for the final training to generate the SVM model file. We utilized the *LIBSVM* software package [15] for implementing the multi-class SVM classifier. For text based indexing, we only consider the keywords (after removing stop words and stemming) from the *"ArticleTitle"* tag of the XML formats of each abstract, which are linked by *one-to-one* or *one-to-many* relationship with images in the collection.

The performances are compared with and without using any QE in different feature spaces, i.e., visual, text, and multi-modal as shown in Table 1. For the automatic simulation of QE, we considered top 20 retrieved images from the previous iteration as the local feedback for the next iteration and selected three additional visual and text keywords from the local clusters for each query keywords in both the visual and text feature spaces. For multi-modal retrieval, the search is initiated simultaneously based on both text and image parts of a query and later the individual results are linearly combined (with weight $\omega_I = 0.3$ and $\omega_D = 0.7$) for a final ranked result list. It is clear from Table 1 that the

Table 1. Retrieval results of different methods

Method	**Modality**	**Query Expansion**	**MAP**	**P20**
Visual Keyword	Image	Without	0.025	0.0717
Visual Keyword-QE	Image	With	0.028	0.0767
Text Keyword	Text	Without	0.1253	0.1469
Text Keyword-QE	Text	With	0.1311	0.1491
Visual-Text Keyword	Image+Text	Without	0.1426	0.1522
Visual-Text Keyword-QE	Image+Text	With	0.1501	0.1564

retrieval performance was improved for the QE-based approaches (even after using only one iteration of feedback) compared to the case when no QE is utilized with image representation in visual and text keyword spaces. The proposed QE method performed well in all cases, i.e., whether it was applied to a single modality or was applied in the multi-modal search with a linear combination scheme. In general, we achieved around 4-5% increase in MAP scores for all QE based searches compared to searches without any expansion. For example, for the search with multi-modal QE (e.g., Visual-Text Keyword-QE) where visual and textual expansions are performed together, we achieved the best MAP score of 0.15. Finally, from the results, we can conjecture that there exists enough correlation between the visual-visual and text-text keywords, which can be exploited with QE or modification process. It is also evident that combining both the visual and text keyword-based features as well as using QE techniques can significantly improve retrieval performance. A comparison with results from approaches by other imageCLEFmed'08 participants is not possible due to lack of evidence on their use of any query-expansion methods or kinds of relevance feedback techniques in a multi-modal framework.

5 Conclusions

This paper investigates a novel multimodal query-expansion (QE) technique for medical image retrieval inspired by approaches in text information retrieval. The proposed technique exploits correlations between the visual and text keywords using a local analysis approach. We observe that there exists enough correlation between keywords within each modality and exploiting this property reduces the keyword mismatch problem. Furthermore, a standard image dataset has provided enough reliability for objective performance evaluation that demonstrates the efficacy of the proposed method.

References

1. H. Müller, T. Deselaers,T.M. Deserno, J. Kalpathy-Kramer, E. Kim, and W. Hersh, "Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks," *8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007), Proceedings of LNCS*, vol. 5152, 2008.
2. 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.
3. T. C. Wong, *Medical Image Databases*, New York, LLC: Springer Verlag, 1998.
4. J. Xu, and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Trans. on Info. Sys.*, 18(1):79–112, 2000.
5. J. Xu, and W. B. Croft, "Query Expansion Using Local and Global Document Analysis," *Proc. 19th Annual Int'l ACM SIGIR Conf. on Research and Develop. in Info. Retrieval*, pp. 4–11, 1996.
6. C. J. Crouch, "An approach to the automatic construction of global thesauri," *Info. Process. and Management*, 26(5):629–40, 1990.
7. R. Attar, and A. S. Fraenkel, "Local feedback in full-text retrieval systems," *J. ACM* , 24(3):397–417, 1977.
8. Y. Liua, D. Zhang, G. Lu, and W. Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, 40(1):262–82, 2007.
9. R. Datta, D. Joshi, J. Li and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, 40(2):1–60, 2008.
10. X. J. Wang, W. Y. Ma, and X. Li, "Exploring Statistical Correlations for Image Retrieval," *Multimedia Systems*, vol. 11(4), pp. 340–51, 2006.
11. J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models," *Proc. of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 119–26, 2003.
12. R. B. Yates and B. R. Neto, *Modern Information Retrieval*, Addison Wesley, 1999.
13. V. Vapnik, *Statistical Learning Theory*, New York, NY, Wiley; 1998.
14. T. Hastie and R. Tibshirani, "Classification by pairwise coupling," In: Jordan, M.I., Kearns, M.J., Solla, A.S. (eds.): *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.
15. C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.
16. O. Yasushi, M. Tetsuya, and K. Kiyohiko, "A fuzzy document retrieval system using the keyword connection matrix and a learning method," *Fuzzy Sets and Systems*, 39(2):163–79, 1991.