Multi-modal Medical Image Retrieval

Yu Cao[*a], *Henning Müller*[b], *Charles E. Kahn, Jr.*[c], *Ethan Munson*[d]
[a] Department of Computer Science & Engineering, University of Tennessee at Chattanooga, CA, USA;
[b]University Hospitals and University of Geneva, Geneva, Switzerland
[c] Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA
[d] Department of EECS, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
*Contact Author: yu-cao@utc.edu

1. DESCRIPTION OF PURPOSE

Images are ubiquitous in biomedicine and the image viewers play a central role in many aspects of modern health care. Tremendous amounts of medical image data are captured and recorded in digital format during the daily clinical practice, medical research, and education (in 2009, over 117,000 images per day in the Geneva radiology department alone). Facing such an unprecedented volume of image data with heterogeneous image modalities, it is necessary to develop an effective and efficient medical image retrieval system for clinical practice and research. Traditionally, medical image retrieval systems rely on text-based retrieval techniques that use the captions associated with the images, and most often, the access is by patient ID, only. Since the 1990s, we have seen increasing interests in content-based image retrieval for medical applications. One of the promising directions in content-based medical image retrieval is to correlate multi-modal information (e.g., text and image information) to provide better insights.

In this paper, we concentrate our efforts on how to retrieve the most relevant medical images using multi-modal information. Specifically, we use two modalities: the visual content of the images (represented by visual features) and the textual information associated with the images. The core idea for multi-modal retrieval is rooted in information fusion. Existing literature on multi-modal retrieval can roughly be classified into two categories: feature fusion and retrieval fusion. The feature fusion strategy generates an integrated feature representation from multiple modalities. The retrieval fusion strategy refers to the techniques that merge the retrieval results from multiple retrieval algorithms. Our proposed approach belongs to the first category (feature fusion) and is largely inspired by Pham et al. [1] and Leinhart et al.[2]. In [1], the features from different modalities are normalized and concatenated to generate the feature vectors. Then, the Latent Semantic Analysis (LSA) is applied on these features for image retrieval. In [2], Lienhart et al propose a multi-layer probability Latent Semantic Analysis (pLSA) to solve the multi-modal image retrieval problem. Our proposed approach is different from Pham et al. [1] in that we do not simply concatenate the features from different modalities. Instead, we represent the features from different modalities as a multi-dimensional matrix and incorporate these feature vectors using an extended pLSA model. Our method is also different from Lienhart et al. [2] since we use a single pLSA model instead of multiple pLSA models. The major contribution of our work is the new representation of an image using visual-textual "words". These "words" are generated from the visual descriptors and textual information using the extended pLSA model.

2. METHODS

Figure 1 depicts an overview of our method. There are two components in our system. The first one is the "Training Component" and the second is the "Retrieval Component". The "Training Component" is shown in the left side of the figure. The goal of this part is to build the model and generate the latent topic representation for each image in the database. The input of this component includes the images and their textual descriptions. Our algorithms will generate a latent topic representation for each image. For the "Retrieval Component" shown in the right side of the figure our method takes a query image as input and generates the latent topic representation for this image. Finally, we compare the distance between the images in the database and the query image by performing a histogram intersection between the latent topic representations. The images in the database are ranked based on the similarity score. In the following sub-sections, we first introduce the structure of the extended pLSA model. Then we present our algorithms for learning the model parameters, followed by a detailed description on retrieving the images for a given query image

2.1. Structure of the Extended pLSA Model

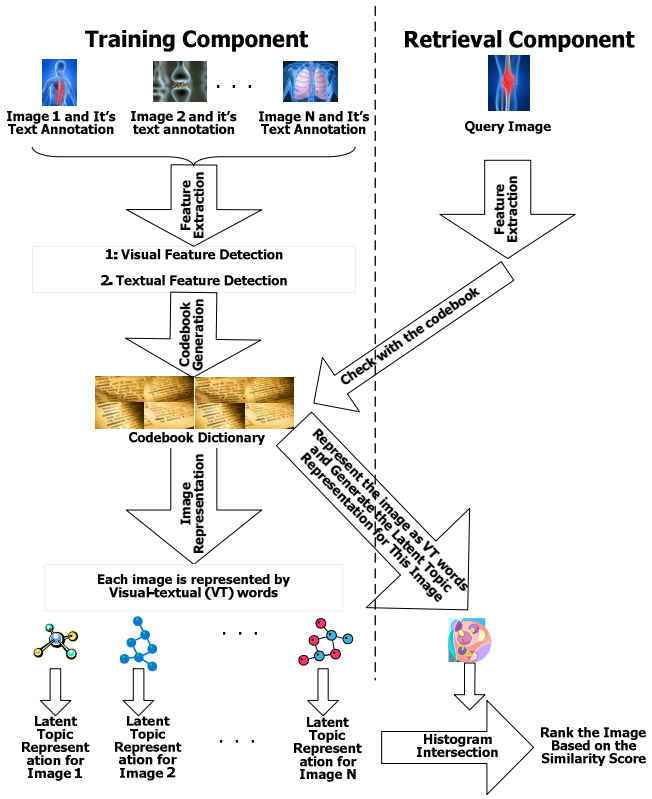

Figure 1: System overview of the proposed approach

We use an extended pLSA model to encode the visual and textual information for each image. The original pLSA method is based on an aspect model, which is a latent variable model for general co-occurrence data (e.g., document-word frequency matrix). It models the distribution of words in the document as a mixture of a few aspects. It was recently employed by the computer vision community to solve the problems of image retrieval and object class recognition. We extend the pLSA model by employing two random variables to represent the visual and textual features. In the following descriptions, we present the models follow the terms and conventions introduced in prior research [1-3].

Suppose we have D ($D = \{d_1, \cdots d_N\}$) images where d_i represents the *i*th image that contains both visual and textual information. We use two random variables (w_t and w_v) to represent the visual and textual words, respectively. We assume that the visual vocabulary is represented as $W_V = \{w_{V\_1}, \cdots w_{V\_M}\}$ while the textual vocabulary is $W_T = \{w_{T\_1}, \cdots w_{T\_K}\}$. The corpus of the image database can be summarized in a three-dimensional co-occurrence matrix $\overline{N}$, whose degree is $M \times K \times N$. The entries $n(w_{V\_m}, w_{T\_k}, d_n)$ in this matrix represent how often the term w_{V_m} and w_{T_k} occurred in image d_n. A latent topic variable z is used to associate the occurrence of words w_V and w_T to image d. The joint probability model over $W_V \times W_T \times D$ is represented by the following equation:

$$P(w_V, w_T, d) = P(d) \cdot P(w_V, w_T \mid d). \quad (1)$$

From Equation (1), we can perform further derivation by importing the latent variable z.

$$P(w_V, w_T, d) = \sum_{z \in Z} P(z) P(d \mid z) P(w_V, w_T \mid z). \quad (2)$$

We use the Expectation-Maximization (EM) algorithms for both training and retrieval. EM alternates two steps: (1) an expectation (E) step where posterior probabilities are computed for the latent variables, (2) a maximization (M) step, where parameters are updated. In the final stage of the training component we compute the value of $P(z_l|d_i)$ for each image d_i ($l \in (1, L)$, where L is the number of latent topics). During the retrieval stage similar operations are performed to the query image. More details are provided in Section 2.3. Finally, we use a histogram intersection (or potentially other distance measures) to measure the similarity between the query image and the images in the database.

2.2. Training

The goal of the "Training Component" is to determine the distribution of the visual-textual words over the latent topic. It includes two parts: visual feature extraction and textual feature extraction. To obtain the visual features we employ a bag-of-words (BoW) method, which is described in other research [3, 4]. Textual features are extracted from the text annotations associated with the images. We apply the existing vector-space model to the textual annotations. Some necessary pre-processing (e.g., removing stop words and stemming) are performed. Now, each image is represented by a two-dimensional matrix that indicates the co-occurrence of the visual-textual words in this image. Therefore, the entire training data is represented by a three-dimensional matrix. Then we apply the EM algorithm to this three-dimensional co-occurrence table and obtain the model parameters.

2.3. Retrieval

The goal of the retrieval component is to compute the similarity score between the database images and the query image. The first step is to extract the visual and textual features from the query image. Based on the features and the codebook (which is generated during the training stage) we could project the query image on the simplex spanned by the $P(w_V, w_T \mid z)$, which is the visual-textual word distribution over a latent topic. Given a query image d_q, we need to calculate the $p(z_k|d_q)$ ($k \in (1, L)$) where L is the number of latent topics. To calculate $p(z_k|d_q)$, we apply Bayes rule to generate the following equation:

$$P(z_k|d_q) = \frac{P(d_q|z_k) \cdot P(z_k)}{P(d_q)}. \quad (7)$$

In order to obtain the likelihood and the prior in Equation (7), an EM algorithm that is similar to the one used in the training stage is employed. Different from the EM method for training, the value of $P(w_V, w_T \mid z)$ is fixed during the EM execution and this value is obtained from the training stage. Once each $p(z_k|d_q)$ is calculated, we generate a histogram representation for the query image by concatenating each $p(z_k|d_q)$ value. Distance metrics such as the histogram intersection are employed to compute the similarity between the query image and the database images. Finally, the database images are ranked based on the similarity score.

3. RESULTS

Table I
Results of Proposed Approach for Multi-modal Retrieval

rel_ret	map	gm_map	Rprec	bpref	recip_rank
1804	0.2919	0.2031	0.3181	0.3196	0.6482

P_5	P_10	P_15	P_20	P_30	P_100
0.5600	0.5440	0.5387	0.5240	0.4627	0.3048

To show the effectiveness of our approach we use the medical images from the ImageCLEF 2009 medical retrieval challenge [5], which is a widely used dataset for medical image retrieval. It contains 74,902 radiological images from two leading peer-reviewed journals (Radiology and RadioGraphics). These images are linked with their existing textual annotations (the captions of the images) extracted from the journal papers. Therefore, this image collection represents a

wide range of medical knowledge. The ImageCLEF challenge also provides 25 realistic search topics. Each search topic contains both the textual key words and the query images. In our implementation, we use these realistic search topics as our queries. Table 1 shows the results of our proposed approach. The numbers in this table are generated with the standard tool [6] used by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results. The overall performance is encouraging with a Mean Average Precision (MAP) at *0.29*. For performance comparison, we implemented other retrieval algorithms. The first compared algorithm, defined as algorithm A, used similar visual features and learning framework as our proposed approach. It does not use the textual information. The second compared algorithm, defined as algorithm B, only used textual features. The average MAP of algorithm A and B are 0.01 and 0.21 respectively. These experiments show that the proposed method is more effective because of the integration of both visual and textual features. The average MAP in our proposed approach is not as good as the best performer in the ImageCLEF medical retrieval challenge 2009, whose average MAP is 0.37 [5]. Other measurements, such as "ret_ret", "bpref", early precision, are very close to the best performer. One of the possible reasons is the usage of the medical ontology (e.g., Unified Medical Language System) by the best performer in the ImageCLEF challenge. We believe that further improvements can be achieved by employing a medical ontology. This will be one of our future works. Interested readers can test our current implementation at http://impact.csufresno.edu:8080/.

4. NEW OR BREAKTHROUGH WORK TO BE PRESENTED

The main contribution of this paper is the proposed generative model-based approach for medical image retrieval using both visual and textual information. This approach is based on a new unsupervised learning method using an extended probabilistic Latent Semantic Analysis (pLSA) model. Each image is represented as a collection of visual-textual words, which are generated by fusing the visual features and associated textual information using the pLSA model. The probability distributions of the visual-textual words are learned from training images. Experimental results show the effectiveness of the proposed approach. To the best of our knowledge, no similar research has been reported in the medical image retrieval field and we expect our research could provide useful insights for further investigation.

5. CONCLUSIONS

In this paper, we have demonstrated a new method to integrate the visual and textual information for the purpose of multi-modal medical image retrieval. An extended pLSA model is developed to fuse the visual and textual information. The EM algorithm is used for both learning and the retrieval stage. Experiments on a large scale, real-world medical image dataset validate the proposed methods. In the near future, we plan to adapt more sophisticated visual analysis techniques and model the spatial layout of the local features. We will also explore new methods to integrate a medical ontology into our multi-modal retrieval framework.

6. REFERENCES

[1] T.-T. Pham, N. E. Maillot, J.-H. Lim, and J.-P. Chevallet, "Latent semantic fusion model for image retrieval and annotation," in *Proc. of the sixteenth ACM conference on Conference on information and knowledge management (CIKM)*, Lisbon, Portugal, 2007, pp. 439-444.

[2] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer pLSA for multimodal image retrieval," in *Proc. of the ACM International Conference on Image and Video Retrieval (CIVR)*, Island of Santorini, Greece, 2009, pp. 1-8.

[3] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, P.R.China, 2005, pp. 370- 377.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, 2006, pp. 2169-2178.

[5] H. Muller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. ı. Radhouani, B. Bakke, C. E. K. Jr, and W. Hersh, "Overview of the CLEF 2009 medical image retrieval track," in *10th Workshop of the Cross-Language Evaluation Forum*, 2009, pp. 1-11.

[6] "trec_eval: A standard tool used by the TREC community for evaluating an ad hoc retrieval run," in *http://trec.nist.gov/trec_eval/*. Washington DC, 2010.