Medical Image Retrieval: A Multi-modal Approach

Yu Cao¹*, Shawn Steffey¹, Jianbiao He²*, Degui Xiao³, Cui Tao⁴, Ping Chen⁵, Henning Müller ⁶

¹Department of Computer Science, The University of Massachusetts Lowell
Lowell, MA 01854, USA; Email: ycao@cs.uml.edu; Phone: (978)934-3628; Fax: (978)934-3551

²School of Information Science and Engineering, Central South University, Changsha,
P.R. China 410083; Email: jbhe@mail.csu.edu.cn

³College of Computer Science and Electronic Engineering, Hunan University,
Changsha, P.R. China 410082; Email: dgxiao@hnu.edu.cn

⁴School of Biomedical Informatics, The University of Texas, Health Science Center at
Houston, Houston, TX 77030; Email: Cui.Tao@uth.tmc.edu

⁵Department of Computer Science, The University of Massachusetts Boston,
Boston, MA 02125, USA; Email: Ping.Chen@umb.edu

⁶Information Systems, University of Applied Sciences Western Switzerland (HES-SO),
Medical Informatics, University Hospitals and University of Geneva, Switzerland;
Email: henning.mueller@hevs.ch

*Co-corresponding Authors

Title of journal manuscript is submitted to: Journal title “Cancer Informatics”;
Supplement title “Translational informatics systems and methods”

How our paper is relevant to the Journal of Cancer Informatics: Medical imaging is
becoming a vital component of the war on cancer. Imaging informatics provides critical
services and infrastructure to support cancer diagnosis, treatment, and recovery.
Medical image retrieval is one of the few translation informatics components that cover
a broad range of imaging informatics tasks (e.g., image manipulation, image
management, and image integration) for cancer care and cancer research. Developing
new medical image retrieval algorithms and systems promise to contribute greatly to
improving the quality of cancer care for each individual patient with lower cost.
**Abstract**

Medical imaging is becoming a vital component of war on cancer. Tremendous amount of medical image data is captured and recorded in digital format during cancer care and cancer research. Facing such an unprecedented volume of image data with heterogeneous image modalities, it is necessary to develop effective and efficient content-based medical image retrieval (CBMIR) systems for cancer clinical practice and research. While substantial progress has been made in different areas of content-based image retrieval (CBIR) research, direct applications of existing CBIR techniques to the medical images produced unsatisfactory results, due to the unique characteristics of medical images. In this paper, we develop a new multi-modal medical image retrieval approach based on the recent advances in statistical graphic model and deep learning. Specifically, we first investigate a new extended probabilistic Latent Semantic Analysis (pLSA) model to integrate the visual and textual information from medical images to bridge the semantic gap. We then develop a new Deep Boltzmann Machine (DBM)-based multi-modal learning model to learn the joint density model from multi-modal information in order to derive the missing modality. Experimental results with large volume of real-world medical images have shown our new approach is a promising solution for next generation medical imaging indexing and retrieval system.
Introduction

Medical imaging has been ranked as one of the most important medical developments of the past 1,000 years. Over the last ten years, tremendous amounts of medical image data have been captured and recorded in digital format during daily clinical practice, medical research and education. These data represent a rich source of information that is invaluable for diagnosis, treatment, recovery, rehabilitation, etc. This is particularly true for cancer related research and clinical practice: with the advancement of medical imaging, the body and the chaos of the cancer cells. For example, medical imaging is used to screen, diagnose, and stage cancer, to guide cancer treatments, to monitor cancer recurrence, and to support cancer research such as drug discovery and therapeutic innovation.

Advancing the appropriate use of information technology in medical imaging, as the newest frontier of medical imaging, promises to contribute greatly to improving the quality of cancer care for each individual patient with lower cost. Common information technology and computational tasks related to medical imaging include image acquisition, image manipulation, image management, and image integration. Medical image retrieval is one of the few computational components that covers a broad range of tasks including image manipulation, image management, and image integration. The goal of medical image retrieval is to find the most clinically relevant images in response to specific information needs represented as search queries.

Text-based information retrieval techniques are well researched. However, they are limited by the quality and quantity of the textual annotations of the images. Effective and efficient content-based approach can be used in conjunction with text-based methods to improve the accuracy and completeness of the search results. Motivated by the important potential clinical benefits, content-based medical image retrieval has become a very active research area over the last decade.

Among many state of the art techniques of content-based medical imaging retrieval research, one of the most promising directions is to correlate multi-modal information (e.g., text and image) or to use the combinations of textual and visual analysis techniques for more effective and efficient content-based medical image retrieval. In this paper, we focus on two modalities that are widely available in real-world clinic practice: visual information of a medical image and the corresponding text annotation of the medical image.

While substantial amount of research has been conducted in the area of medical image retrieval, real-world tools and applications that can access the medical image data by their content are rare in clinical practice. One of the major barriers we have identified is that the “Semantic Gap” exists between the low level features (e.g., low level visual and textual features) and high level medical concepts. The key to address these issue is to develop new semantic meaningful features from multi-modal information to bridge the “Semantic Gap” by enabling effective and efficient content-based medical image retrieval. Another challenge is that the real-world data is very noisy and some modality information (e.g., text annotation) maybe missing from input. To addressing the missing modality issue, new algorithms that can derive the missing modality information from existing known modality (e.g., deriving missing text annotation from known visual content).
The goal of our research in this paper is to develop, evaluate, and demonstrate new image and textual analysis techniques to generate semantic features for scalable semantic content analysis and medical image indexing to overcome the above limitation. Our new approach has great potential to improve the performance of medical image retrieval substantially and can be applied in clinical practice and healthcare applications. Our approach is inspired by recent advancement in statistical graphic models \textsuperscript{24,25}, and deep learning \textsuperscript{26}. Specifically, we first develop a new extended probabilistic Latent Semantic Analysis (pLSA) model to integrate the visual and textual information from medical images to bridge the semantic gap. While the proposed pLSA model is able to generate a representation space with desired feature (e.g., similarity in this feature space implies the similarity in corresponding medical concepts), we are still facing some additional issues if employing this model in real-world applications. More specifically, in real world clinical applications, the situations where some modalities are missing and noisy happen frequently. We plan to develop our second model to address these issues. More specifically, we then develop a new Deep Boltzmann Machine (DBM)-based \textsuperscript{27} multi-modal learning model to learn the joint density model from multi-modal information for deriving the missing modality information. By combining the two models, which are trained with a large volume of available medical image data, we are able to search the most relevant images for a given query. Experimental results with large volume, real-world medical images have shown our approach is a promising solution for next generation medical imaging indexing and retrieval system.

**Motivations**

As medical imaging is becoming an essential component for cancer care and research, many departments of cancer care and research would benefit directly from improvements in multi-modal content-based medical image retrieval research. Medical imaging is becoming even more important over the last ten years. One of the reasons behind this is the so called “Big Data” in medical imaging: tremendous amounts of medical image data, in the last few years, are captured and recorded in digital format during the daily clinical practice, medical research, and education. Driven by the aging population and technology advancements, the global diagnostic imaging market is expected to increase to $26.6 billion by 2016 \textsuperscript{2}. In 2010, over 5 billion medical imaging studies had been conducted worldwide \textsuperscript{3}. In 2011, the number of U.S. medical imaging procedures surpassed the 800 million mark \textsuperscript{4}. At the Mayo Clinic in Jacksonville, Florida, USA, a radiologist viewed 1,500 cross-sectional images per day in 1994 compared to 16,000 images per day in 2004 \textsuperscript{5}. The Radiology Department at University Hospitals of Geneva, Geneva, Switzerland, produced over 12,000 images per day in 2004, 40,000 images per day in 2006, 70,000 images per day in 2007, and over 117,000 images per day in 2009 \textsuperscript{13}. Images are ubiquitous in cancer care and research. The image viewers play a central role in many aspects of modern cancer care. This data provides an unprecedented opportunity for making smart and optimized healthcare decisions with improved outcomes while reducing costs.

Common computational tasks related to medical imaging include image acquisition, image manipulation, image management, and image integration \textsuperscript{2}. Medical image retrieval, with the goal of finding the most clinically relevant images in response to specific information needs represented as search queries, is one of the few computational components that cover a broad range of medical imaging computational...
tasks. Despite text-based information retrieval methods being both mature and well researched, they are limited by the quality of image annotations. Among other important limitations facing traditional text retrieval techniques are the fact that image annotations are subjective and context sensitive, and can be quite limited in scope or even completely absent. Manually annotating images is also label intensive, and can be very error-prone. Image annotations are quite noisy if they are automatically extracted from the surrounding text using natural language processing techniques, and there is much more information in an image than can be extracted using a limited number of words. Effective and efficient content-based approaches can be used in conjunction with text-based methods to improve the accuracy and completeness of the search results.

**Related Work**

The related work to this paper falls under three categories. The first category is content-based medical image retrieval research. The second category is multi-modal information fusion-based image retrieval research. The third category is deep learning research. We will introduce these three categories in more detail in the following paragraphs.

The first category, content-based medical image retrieval, has been a very active field in recent years. Many projects have advanced the work through a number of experiments and research publications. Several research projects have advanced the state of the art through a large number of publications and explorations of different aspects of content-based medical image retrieval over the last decade. In a database of computed tomography (CT) images of the chest called “automated search and selection engine with retrieval tools” (ASSERT) was developed. Image Retrieval in Medical Applications (IRMA) is a project that aims to develop high-level methods for content-based image retrieval with prototypical application to medical diagnostic tasks on a radiologic image archive. The Medical GNU Image Finding Tools (MedGIFT) project includes several axes around the retrieval of medical images from a variety of databases and image types as well as for several applications. Greenspan and Pinhas present a representation and matching framework for image categorization in medical image archives. Napel et al. develop a system to facilitate radiologic image retrieval that contain similar-appearing lesions. System evaluation was performed with a CT image database of liver and an external standard of image similarity. In Rahman et al. propose a unified medical image retrieval framework integrating visual and text keywords using a novel multi-modal query expansion (QE). Quellec et al. introduces a content-based heterogeneous information retrieval framework. In this paper, they propose a Bayesian network to recover missing information. The Medical Imaging Resource Center (MIRC) project was initiated by the RSNA Radiology Informatics Committee to construct a library of medical information, globally accessible to the imaging community over the Internet. An example of evaluation projects is a medical image retrieval task to benchmark and compare the performance of participating systems. Recently, researchers from U.S NLM/NIH developed new computer-aided techniques to identify and annotate the region of interests for a given medical image to facilitate biomedical document and image retrieval. Some medical image search prototypes, such as GoldMiner, were also developed. In paper, Kumar et al. presented a review present a review of state-of-the-art medical CBIR approaches in five main categories: (1) two-dimensional image retrieval; (2) retrieval of images with three or more dimensions; (3) the use of nonimage data to enhance the retrieval; (4)
multimodality image retrieval; and (5) retrieval from diverse datasets. Our system is different from the state of the art because our statistic graphic model and deep learning model make it possible to develop semantic features for bridging the “Semantic Gap”.

The second category of related work is called multi-modal fusion-based image retrieval. The research in this area is rooted in information fusion. Existing literature on multi-modal retrieval can roughly be classified into two categories: feature fusion and retrieval fusion. The first strategy (feature fusion strategy) generates an integrated feature representation from multiple modalities. For example, in $^{24}$, 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 $^{25}$, Lienhart et al propose a multi-layer probability Latent Semantic Analysis (pLSA) to solve the multi-modal image retrieval problem. The second strategy (retrieval fusion) refers to the techniques that merge the retrieval results from multiple retrieval algorithms. Our approach belongs to the first category (feature fusion). Our technique is different from Pham et al. $^{24}$ 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. $^{25}$ since we use a single pLSA model instead of multiple pLSA models.

The third category of related work, deep learning $^{26-39}$, aims to learn multiple levels of representation and abstraction that help infer knowledge from data such as images, videos, audio, and text, making astonishing gains in computer vision, speech recognition, multimedia analysis, and drug designing. The impact of deep learning is far reaching on applications in medical, social and commercial domains $^{40-42}$. Briefly speaking, there are two main classes of deep learning techniques: purely supervised learning algorithms (e.g., Deep Convolutional Network $^{43-44}$), unsupervised and semi-supervised learning algorithms (e.g., Denoising Autoencoders $^{45-46}$, Restricted Boltzmann Machines $^{47,48}$, Deep Boltzmann Machines $^{25}$). RBM was proposed as a significant improvement for Boltzmann Machines (BM) $^{49}$. Let’s give a brief background introduction on Boltzmann Machines (BM) $^{49}$, Restricted Boltzmann Machines (RBM) $^{47,48}$, and Deep Boltzmann Machines (DBM). In the following introduction, we will present the models following the terms and conventions introduced in prior research $^{27,27,47,49}$. Boltzmann Machines (BM) is a stochastic recurrent neural network and it is named after the Boltzmann distribution in statistical mechanics. BM is a network of units with an "energy" defined for the network. It also has binary units, but unlike Hopfield nets, Boltzmann machine units are stochastic. BM is a network of symmetrically coupled stochastic binary units. It includes a set of visible nodes $\{0,1\}$ and a set of hidden nodes $\{0,1\}$. The state $\{v, h\}$’s energy is defined as follows: $E(v, h; \theta) = \frac{1}{2}v^T L v - \frac{1}{2}h^T J h - v^T W h$, where $\theta = \{W, L, J\}$ are the parameters of the model. $W$, $L$, and $J$ indicate visible-to-hidden, visible-to-visible, and hidden-to-hidden symmetric interaction terms. In theory, the RM model is a general computational model that is suitable for many applications. However, in practice, learning in general BM is very inefficient. As a result, Restricted Boltzmann Machines (RBM) was proposed with a fast learning algorithm. Different from BM, RBM model does not allow visible-to-visible connections and hidden-to-hidden connections. In another word, in the energy model, we will set both $L = 0$ and $J = 0$. By setting both of them as zero, we can remove every intra-layer connection. The inference in RBM is
exact, which is also different general BM’s inference. While exact maximum likelihood learning in RBM is intractable, new learning algorithms, such as Contrastive Divergence, has been proposed to carry out the learning process very efficiently. When one RBM is trained, we can treat the activities of its hidden nodes as input for a higher-level RBM. By stacking multiple RBMs together, we could train many layers of hidden units efficiently. This method of stacking RBMs will produce a so called Deep Boltzmann Machines (DBM), which is one of the most common deep learning strategies. Among the state of the art of DBM techniques, the findings presented by papers 50, 51 are most similar to our proposed approach. In paper 50, the authors covered a semi-supervised learning method, as well as using both labeled and unlabeled training data. Their work showed that machine learning could be improved with unlabeled image data, provided associated text tags were also incorporated as another modality. Most recently, Srivastava and Salakhutdinov 51 presented method of accomplishing image recognition tasks for MIR Flickr dataset 52 of images and associated tags utilizing a DBM. In that paper 51, they demonstrate vast potential for DBMs and their competency for multi-modal learning. In this paper, we will focus on developing new semi-supervised learning algorithms using Deep Boltzmann Machines (DBM) 53, 54, 55. Our proposed solution is an extended model of DBM, which is a learning system through layers of binary stochastic variables. These layers are interconnected, but have no connections between nodes on the same layer. This allows for much faster processing of information than standard Boltzmann Machines, which are entirely interconnected.

**Our Multi-modal Approach**

In this section, we will introduce our approach in more detail. To the best of our knowledge, how to generate effective and efficient semantic features for large-scale medical image sets remains a challenging and unsolved problem. In this paper, we will focus on the development and evaluation of new semantic analysis techniques by investigating and evaluating innovative visual and textual modeling and analysis techniques for generating the semantic features. These semantic features have potential to bridge the semantic gap. Specifically, we develop two types of statistical graphic models that can fuse the distinct modalities to generate the semantic features, which can support the capturing of the real-world medical concepts effectively and efficiently. Furthermore, the proposed approach is able to handle the missing modality reliably. The semantic features are a new representation space with the following advantages: (1) the similarity in this representation space implies the similarity of the corresponding real-world medical concepts; (2) the representation space can be generated reliably even in the situations where there are missing and noisy modalities. In the following subsections, we will first introduce the extended pLSA model to fuse the multi-modal information (step 1). Then we will introduce our proposed DBM model (step 2), which can be used to derive the missing modality. Lastly, we will discuss how we will utilize the two proposed models from step 1 and 2 for medical image retrieval (step 3).
Step 1: Fusing the Multi-modal Information

Figure 1 depicts an overview of the first step. Our goal in this step is to build the graphic model and generate the latent topic representation for each image in the database. Given the images and their associated textual descriptions, our algorithms will generate a latent topic representation for each image.

Figure 1: Overview of our model in step 1

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. Please note, in our research, we employ the concept of visual bag-of-words (VBoW) model to extract the initial visual features from image. In this (VBoW) model, an image is represented as a visual document composed of visual elements (A.K.A. “visual words”). This model has been very popular in the last few years, due to its simplicity and scalability. Specifically, we first apply Scale-invariant feature transform (SIFT) feature extractions methods to identify the interesting points from the image. For each interesting point identified by SIFT method, we will extract the SIFT descriptor, which is a 128 dimension vector. We then run K-means clustering algorithm for all the SIFT descriptors collected from all the training images. The K centroids of the K-means algorithm are the “visual words” that can be used for late process. For each image, we compare the SIFT interesting point and its SIFT descriptor with each “visual words” and find out the closet “visual words”. By this way, we can generate a histogram of “visual words” as a feature representation of each image. For textual feature extraction, we employ the existing open source natural language processing package, Stanford NLP package, to extract textual features. Specifically, we employ the textual bag-of-word model and a vocabulary of the 1,000 most frequently used medical terms are used. In the following description, we present the models following the terms and conventions introduced in prior research.

Suppose we have \( D = \{d_1, \cdots, d_N\} \) images where \( d_i \) represents the \( i \)th image that contains both visual and textual information. We use two random variables \( (w_v, w_t) \) 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 \( N \), 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:

\[
P(w_v, w_t, d) = P(d) \cdot P(w_v, w_t | d) \tag{1}
\]

From Equation (1), we can perform further derivation by importing the latent variable \( z \).
\[ P(w_j, w_T, d) = \sum_{z} P(z)P(d | z)P(w_j, w_T | z). \quad (2) \]

We use the Expectation-Maximization (EM) algorithms for training. 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 \( P(z|d_i) \) for each image \( d_i \) (\( i \in [1, L] \), where \( L \) is the number of latent topics). More specifically, in our extended pLSA model, the E-step equation is listed as below.

\[ P(z | w_j, w_T, d) = \frac{P(z)P(d | z)P(w_j, w_T | z)}{\sum_{z} P(z)P(d | z)P(w_j, w_T | z')} \quad . \quad (3) \]

The formulas for the M-step are listed as follows

\[ P(w_j, w_T | z) \propto \sum_{d} n(w_j, w_T, d)P(z | w_j, w_T, d) \cdot \quad (4) \]

\[ P(d | z) \propto \sum_{w_j \in \mathcal{W}_j \text{ and } w_T \in \mathcal{W}_T} \sum_{d} n(w_j, w_T, d)P(z | w_j, w_T, d) \cdot \quad (5) \]

\[ P(z) \propto \sum_{d} \sum_{w_j \in \mathcal{W}_j \text{ and } w_T \in \mathcal{W}_T} \sum_{d} n(w_j, w_T, d)P(z | w_j, w_T, d) \cdot \quad (6) \]

During the retrieval stage similar operations are performed to the query image. More details are provided in Step 3 (“Retrieval” section). Our proposed extended pLSA model, compared with existing pLSA model, employs a three-dimension array. Therefore, compared with original pLSA model, the number of parameters to be estimated during the EM algorithms is also increased. However, as indicated by equations (3) to (6), the increasing of parameters will not cause the computation intractable. 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.

**Step 2: Deriving Missing Modalities**

While our model (extended pLSA model) proposed in the first step is able to generate a representation space with desired feature (similarity in this feature space implies the similarity in corresponding medical concepts), we are still facing some additional issues if employing this model in real-world applications. More specifically, in real world clinical applications, the situations where some modalities are missing and noisy happen frequently. We plan to develop our second model to address these issues. Our proposed approach is rooted from the recent advances in deep learning \cite{26, 62, 63}. The main innovation in this step is to learn a joint probability density model from the visual and textual information with the capacity of filling in missing modalities.

Figure 2. Overview architecture of our model based on DBM.

As shown in Figure 2, the novel part of our proposed approach is that of utilizing multi-modal inputs for analysis. This joint model is accomplished by training two separate DBMs, with the top hidden layers connected to a combined hidden layer to act as a joint representation for the associated learning. As show in Figure 2, we first extract both visual and textual features from the images. Then we train a visual-based DBM, as
shown in the middle left of Figure 2. We also train a text-based DBM, as shown in the middle right of Figure 2. Both of the two DBM has two hidden layers. In order to fuse the multi-modal information, we add one additional layer on top of these two DBMs as the joint representation of multi-modal data, as shown at the bottom of Figure 2.

As shown in Figure 2, our approach utilizes multiple layers of hidden variables, each layer connected to the neighboring layers through each and every node. One layer represents the visible data (In the case of image training, the image pixel data), all subsequent layers are hidden. The connections between nodes are weighted according to a probability function to be evaluated during the training sessions \(^{27}\). In order to derive the missing modality, we will first learn a joint density model from multi-modal information using the proposed DBM. Specifically, the proposed DBM includes a set of visible nodes \(v \in \{0,1\}^D\). It also includes several layers of hidden nodes \(h^1 \in \{0,1\}^E\), \(h^2 \in \{0,1\}^E\), ..., \(h^L \in \{0,1\}^E\). Please note, the basic building block of DBM is RBM, which is a special case of BM (Boltzmann Machine). In RBM, the connections exist only between nodes from different layers and there is no connections between nodes within the same layer. The joint distribution over the visible and hidden units is defined as \(P(v,h;\theta) = \frac{1}{Z(\theta)} \exp(-E(v,h;\theta))\) where \(Z(\theta)\) is defined as the partition function and \(E(v,h;\theta)\) is the state \(\{v,h\}\)’s energy. Since there is no connections between the same layer, the state \(\{v,h\}\)’s energy is defined as follows:

\[
E(v,h;\theta) = \frac{1}{2} v^T W h - b^T v - a^T h,
\]

where \(\theta = \{W,L,J\}\) are the parameters of the model. \(W, L,\) and \(J\) indicate visible-to-hidden, visible-to-visible, and hidden-to-hidden symmetric interaction terms sample the hidden modality from the conditional distributions, given the observed modalities. While there are many sampling algorithms that can obtain observations from probability distribution, we choose to employ the Gibbs sampling technique \(^{64}\), used by recent deep learning research \(^{27,65,66}\).

### Step 3: Retrieval

Once the two models (extended pLSA model shown in Figure 1 and the DBM model shown in Figure 2) are trained, we will represent the visual and textual information using the trained models. Specifically, we will determine the distribution of the visual-textual words over the latent topic generated from the new pLSA model. We will also generate the missing data using the conditional distribution over observed data.

To obtain the visual features we employ a bag-of-words (BoW) model \(^{61,67}\). 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.

The last step is to perform retrieval. The goal 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 $P(w_r, w_r | z)$, which is the visual-textual word distribution over a latent topic. Given a query image $d_q$, we need to calculate the $p(z|d_q) (k \in \{1, L\})$ where $L$ is the number of latent topics. To calculate $p(z|d_q)$, we apply Bayes rule to generate the following equation:

$$p(z|d_q) = \frac{p(d_q|z) p(z)}{p(d_q)}. \quad (7)$$

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 (which is introduced in Step 1: “Fusing the Multi-modal Information”), the value of $P(w_r, w_r | z)$ is fixed during the EM execution and this value is obtained from the training stage (which is introduced in Step 1: “Fusing the Multi-modal Information”). Once each $p(z|d_q)$ is calculated, we generate a histogram representation for the query image by concatenating each $p(z|d_q)$ 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.

**Experimental Results**

In this section, we will introduce the data sets used in our experiments (Sub-section entitled as “Data Sets”), the list of performance metrics (Sub-section entitled as “List of Definitions for Performance Metrics”), as well as the detailed implementation and experimental results (sub-section entitled as “System Implementation and Detailed Results”). In the sub-section entitled as “Data Sets”, we introduce the characteristics of the two data sets: ImageCLEF 2009 medical retrieval challenge, and ImageCLEF 2013 medical retrieval challenge. Then we introduce the list of definitions of performance metrics at the sub-section entitled as “List of Definitions for Performance Metrics”. Some sample metrics such as precision, recall, MAP, Ret_ret, and etc, used in our experiments are introduced in this sub-section. Finally, we introduce our system implementation and present the experimental results with detailed analysis at sub-section entitled as “System Implementation and Detailed Results”.

**Data Sets**

We employ two data sets that have been widely used in medical image retrieval research.

The first data set is medical images from the ImageCLEF 2009 medical retrieval challenge. 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. Figure 3 illustrates some sample queries, including both textual key words and the query images, used in this data set.

Figure 3. Sample queries (textual key words and the query images) from data set 1 (ImageCLEF 2009 medical retrieval challenge).

The second data set is medical images from the ImageCLEF 2013 medical image retrieval challenge. Similar to the ImageCLEF 2009 retrieval challenge, the images from 2013 challenge are also retrieved from open accessed biomedical literature. Instead of limiting the literature to the two radiology journals used in 2009 challenge, the 2013 challenge expand the literature to many other radiology journals in PubMed Central. As a result, the 2013 retrieval challenge contains 305,000 medical images, which represent one of the largest medical image collections available for research community. Similar to the 2009 challenge, the 2013 retrieval challenge also provides ad-hoc image-based retrieval examples. The examples include 30 textual queries with 2-3 sample images for each query. We use these examples for our validation. Figure 4 illustrates some sample queries, including both textual key words and the query images, used in this data set.

Figure 4. Sample queries (textual key words and the query images) from data set 2 (ImageCLEF 2013 medical retrieval challenge).

As shown in Figure 3 and 4, the number of key words in most of the search topics in the ImageCLEFmed datasets is between 3 and 5. For example, sample query 1 in Figure 3 has three key words: “Breast”, “cancer”, and “mammogram”. Sample query 1 in Figure 4 (images from ImageCLEFMed 2013) has three keywords: “osteoporosis”, “x-ray”, and “images”.

Please note, the total number of queries (25 queries from ImageCLEF 2009 and 30 queries from ImageCLEF 2013) is relative smaller than the total of images (74,902 images from ImageCLEF 2009 and 305,000 images from ImageCLEF 2013). This is a normal setting for every participant in ImageCLEF 2009 and 2013.

List of Definitions for Performance Metrics

In order to evaluate the results, we need to employ a sequence of metrics (a.k.a. performance measurements) to determine whether the returned results are relevant to a given query. The following performance measurements are used in our experiments: “Precision”, “Recall”, “Average Precision”, “Mean Average Precision (MAP)”, “bpref”, “P5”, “P10”, “P30”, and “rel_ret”.

Precision is defined as $\frac{relevant\_doc}{retrieved\_doc}$, where relevant documents are the returned results for this query from the ground truth, retrieved documents are the returned results for this query from my search engine. Therefore, precision is the number of elements in the array containing the intersection of ground truth results to our results divided by the number of elements in the array containing our results.

Recall is defined as $\frac{relevant\_doc}{relevant\_doc}$, where relevant documents and retrieved documents have the same meaning as in they did in the precision calculation. Therefore, recall is equal to the number of elements in the array containing the intersection of ground truth results to our results divided by the number of elements in the array containing ground truth results.
Average precision is defined as \[ \sum_{r=1}^{N} \frac{P(r) \times \text{rel}(r)}{|\text{relevant}\_doc|} \], where relevant documents have the same definition as in precision and recall, \( r \) is the rank, \( N \) is the number retrieved, \( P(r) \) is the precision of result at rank \( r \), and \( \text{rel}(r) \) is the relevance of result at rank \( r \). In other words, to calculate average precision, at each result, if the current result is relevant, we calculate the precision for every result up to the current result. Then we divide that number by the amount of results so far. After we have performed this calculation for every retrieved document, we divide it by the number of relevant documents to obtain average precision for that query. Average precision is useful because it places more weight on relevant documents since irrelevant documents are considered zero in the calculation. To determine mean average precision, or \( MAP \), we calculate average precision for several different queries then divide by the number of queries.

\( b_{pref} \) is the number of time that non-relevant documents are retrieved before relevant documents. \( b_{pref} \) is equal to \[ \frac{1}{R} \sum_{r} 1 - \frac{\text{num\_ranked\_higher\_than\_r}}{R} \] where \( r \) is the relevant retrieved documents, \( R \) is the relevant documents, and \( n \) is a member of the first \( R \) irrelevant retrieved documents.

Precision after five, ten, and thirty retrieved results a re represented by \( P5 \), \( P10 \), and \( P30 \) respectively. Performing these calculations illustrates how much weaker our returned results become as the rank decreases.

Finally, the \( rel\_ret \) measure is simply another name for the number of relevant documents in the retrieved documents according to the ground truth. This is the same number used in precision and recall.

**System Implementation and Detailed Results**

Since both the first data set (ImageCLEF 2009 challenge) and the second data set (ImageCLEF 2013 challenge) provide sample queries (25 queries and 30 queries from 2009 challenge and 2013 challenge, respectively), we use them as the ground truth. This ground truth is determined by a group of biomedical domain experts. Using the ground truth we could measure the accuracy of the results of the 55 queries in our system.

For the purpose of training the retrieval model, following the experimental setting in the state-of-the-art, we choose 25% of the images from ImageCLEF 2009 data set and 2013 data set, respectively as the data set. This means the total number of images used for model training is around 100,000.

For visual features extraction, we follow the traditional visual BoW approach. Briefly speaking, we first extract SIFT interesting point of the image and its corresponding SIFT descriptors from the 100,000 training images. We then apply the K-means clustering algorithm to all the SIFT descriptors. We experimented different K value and we choose K as 3,000 from the experimental results. Please note, based on the literature research and our own experiments, an optimal K value is largely application dependent. Next, we generate a histogram for each image. The number of the bins for the histogram is equal to the number of centroid from the K-means algorithm (in our context, the number of centroid is 3,000). The histogram is generated by comparing the SIFT interesting point and its SIFT descriptor with each centroid and identifying the closet centroid.
For textual feature extraction, we employ Stanford NLP package (an open source natural language processing package) to extract textual features. Traditional textual bag-of-word model was employed and 1,000 most frequently were used medical terms.

Our implementation is running in a server equipped with 128GB RAM, an eight-Core Intel Xeon E5-2600 v2 Series CPU, and 2 NVidia K-40 GPU. The most time consuming part is the training of DBM model. In our implementation, it took around 3 days to train the model. This is consistent with the state-of-the-art deep learning implementations, which usually take one week to train a deep model. The time for feature extraction, K-mean clustering, and Visual BoW generation are relative short. For example, the average time for extracting visual feature for one image is around 50 millisecond.

Table 1 Results of Our Approach for Multi-modal Retrieval Using the Two Data Set
Table 1 illustrates the results of the proposed approach when applied to the two data sets. The numbers in these tables are generated with the standard tool 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 0.2909. Other numbers, such as bpref, P_5, P_10, ret_ret, and etc, are also equivalent or better than the results from the state-of-the-art. More detailed performance comparison between the proposed approach and the state-of-the-art are introduce in the next paragraph.

Table 2 Performance Comparison between the Proposed Approach and the Image Retrieval Techniques with Single Modality
For performance comparison, we implemented other retrieval algorithms with single modality. The first compared algorithm, defined as algorithm A, used similar visual features and learning framework as our approach. It does not use the textual information. In fact, algorithm A is a standard technique for content-based image retrieval using SIFT Visual Bag-of-Words (BoW) since the visual features used in algorithm A is based on SIFT feature extraction algorithms. The second compared algorithm, defined as algorithm B, only used textual features. As shown in Table 2, 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.

Table 3 Performance Comparison between the Proposed Approach and other State-of-the-art Medical Image Retrieval Techniques
We further compare the proposed approach with the state of the art. The first compared algorithm, defined as algorithm C, was developed using similar multi-modal features and learning framework as introduced in literature. The second compared algorithm, defined as algorithm D, used similar multi-modal features and learning framework as introduced in . The last compared algorithm, defined as algorithm E, was developed by researching the techniques used by the ImageCLEF medical retrieval challenge 2013 participant. We carefully researched the best performer (in terms of MAP), which is the ITI (Image and Text Integration Project) group. This group is from the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, which is a research division of the U.S. National Library of Medicine. Table 3 shows the results (MAP) from different techniques. As shown in this table, our proposed approach outperforms algorithm C. This means our proposed feature fusion techniques based on the extended pLSA model is more suitable than feature
normalization and concatenation (which were used in algorithm C). The results of our proposed approach is only slightly better than the results from algorithm D. Our proposed approach used a single pLSA model but algorithm D employed multiple pLSA models. Therefore, the implementation of our proposed approach is much simpler than algorithm D. The average MAP in our approach is only slightly worse than algorithm E (best performer in the ImageCLEF medical retrieval challenge 2013). 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.

Table 4 Results of the Proposed Approach When Certain Percentage of Textual Modality Is Missing

One additional advantages of our approach, compared with existing methodology, is able to derive the missing modality using the models we developed (Step 2 of “Approach” section). In order to verify the performance of the proposed model for deriving missing modality, we purposely removed around the textual information for 10%, 15%, 20%, 25%, and 30% of the training data set while keeping all the other conditions unchanged. The average MAP we received under these settings are listed in Table 4. In this table, the first row is the percentage of missing textual information. The second row in this table shows the results (MAP) under different missing rate. As shown in this Table, the MAP values under different missing rate are just slightly worse than the MAP value with all the date sets ready. This verifies our effectiveness of our approach.

Conclusions

Our research aims to develop effective and efficient content-based medical image retrieval (CBMIR) systems for cancer clinical practice and research. This is very important because medical imaging is becoming a vital component of war on cancer. Direct applications of existing content-based image retrieval (CBIR) techniques to the medical images produced unsatisfactory results, due to the unique characteristics of medical images. In this paper, we developed a new multi-modal medical image retrieval approach based on the recent advances in statistical graphic model and deep learning. We have investigated a new extended probabilistic Latent Semantic Analysis (pLSA) model to integrate the visual and textual information from medical images. We also developed a new Deep Boltzmann Machine (DBM)-based multi-modal learning model to learn the joint density model from multi-modal information in order to derive the missing modality. To verify the effectiveness of the proposed approach, we validated our system with large volume of real-world medical images. The experimental results have shown that the proposed approach is a promising solution for next generation medical imaging indexing and retrieval system. In the future, we plan to refine our proposed approach with larger data sets and to include medical ontology into our approach. We also plan to explore the possibility of integrating our proposed approach into clinical practice.
Acknowledgements
The authors would like to thank Chyeeka Brown and Huijuan Xu for contributing on part of the system implementation.

Funding sources
The research is supported in partial by National Science Foundation of the United States (Award No. 1156639, 1229213, and 1415477), and National Natural Science Foundation of China (Award No: 61272147 and 61272062)

Author Contributions
Conceived and designed the experiments: YC, JH, DX, CT. Analyzed the data: SS, JH, PC, HM. Wrote the first draft of the manuscript: YC, SS. Contributed to the writing of the manuscript: YC, JH, DX, CT, PC, HM. Agree with manuscript results and conclusions: YC, JH, DX, CT, PC, HM. Jointly developed the structure and arguments for the paper: YC, JH, DX, CT, PC, HM. Made critical revisions and approved final version: YC, SS, JH, DX, CT, PC, HM. All authors reviewed and approved of the final manuscript.

Disclosures and Ethics
As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.
References


Tables

Table 1
Results of Proposed Approach for Multi-modal Retrieval Using the Two Data Set

<table>
<thead>
<tr>
<th>rel_ret</th>
<th>map</th>
<th>gm_map</th>
<th>Rprec</th>
<th>bpref</th>
<th>recip_rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.902</td>
<td>0.2909</td>
<td>0.2019</td>
<td>0.3101</td>
<td>0.3206</td>
<td>0.6421</td>
</tr>
<tr>
<td>P 5</td>
<td>P 10</td>
<td>P 15</td>
<td>P 20</td>
<td>P 30</td>
<td>P 100</td>
</tr>
<tr>
<td>0.5620</td>
<td>0.5510</td>
<td>0.5309</td>
<td>0.5270</td>
<td>0.4647</td>
<td>0.3281</td>
</tr>
</tbody>
</table>

Table 2
Performance Comparison between the Proposed Approach and the Image Retrieval Techniques with Single Modality

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Proposed Multimodal Approach</th>
<th>Algorithm A (Only using Visual Modality)</th>
<th>Algorithm B (Only using Textual Modality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2909</td>
<td>0.0101</td>
<td>0.2013</td>
</tr>
</tbody>
</table>

Table 3
Performance Comparison between the Proposed Approach and the State-of-the-art Image Retrieval Techniques

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Proposed Multimodal Approach</th>
<th>Algorithm C (LSA-based Technique)</th>
<th>Algorithm D (Multilayer pLSA-based Technique)</th>
<th>Algorithm E (Techniques from the Best Performer at ImageCLEF 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2909</td>
<td>0.0912</td>
<td>0.2825</td>
<td>0.3010</td>
</tr>
</tbody>
</table>

Table 4
Results of the Proposed Approach When Certain Percentage of Textual Modality Is Missing

<table>
<thead>
<tr>
<th>Percentage of Missing Textual Modality</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2709</td>
<td>0.2601</td>
<td>0.2505</td>
<td>0.2459</td>
<td>0.2319</td>
</tr>
</tbody>
</table>
Figure legends

Figure 1. Overview architecture of the proposed model in Step 1

Figure 2. Overview architecture of the proposed model based on DBM

Figure 3. Sample queries (textual; key words and the query images) from data set 1 (ImageCLEF 2009 medical retrieval challenge)

Figure 4. Sample queries (textual; key words and the query images) from data set 2 (ImageCLEF 2013 medical retrieval challenge)
Figures

Figure 1. Overview architecture of the proposed model in Step 1

Figure 2. Overview architecture of the proposed model based on DBM
Figure 3. Sample queries (textual; key words and the query images) from data set 1 (ImageCLEF 2009 medical retrieval challenge)

Sample query 1: Breast cancer mammogram

Sample query 2: MR Images of rotator cuff

Figure 4. Sample queries (textual; key words and the query images) from data set 2 (ImageCLEF 2013 medical retrieval challenge)

Sample query 1: osteoporosis x-ray images

Sample query 2: thyroid CT images
Supplementary data
Insert supplementary data here, OR if supplementary data is in separate files, list files here