THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

An Intramural Research Division of the U.S. National Library of Medicine

A Report to the Board of Scientific Counselors September 2010

Combining Text and Visual Features for Biomedical Information Retrieval

Dina Demner-Fushman, M.D., Ph.D.
Sameer K. Antani, Ph.D.
Matthew Simpson
Md. Mahmudur Rahman, Ph.D.

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Table of Contents

List of Figures

Figures in the Main Section

Figures in the Appendix

List of Tables

Tables in the Main Section

Tables in the Appendix

Glossary

ARRS	American Roentgen Ray Society
ASM	Active Shape Modeling
CBIR	Content-Based Image Retrieval
CEDD	Color Edge Direction Descriptor
CLD	Color Layout Descriptor
CLEF	Cross Language Evaluation Forum (http://www.clef-campaign.org)
CCV	Color Coherence Vector
CT	Computerized Tomography
DICOM	Digital Image Communications
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EHR	Electronic Health Record
EHD	Edge Histogram Descriptor
FCTH	Fuzzy Color Texture Histogram
GLCM	Gray Level Co-occurrence Matrices
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ImageCLEFmed	Medical Image Retrieval extension to CLEF (http://imageclef.org)
IR	Information Retrieval
IRMA	Image Retrieval for Medical Applications
ITSE	Image and Text Search Engine
LIRE	Lucene Image Retrieval Engine
MAP	Mean Average Precision
MeSH	Medical Subject Headings
MPEG	Motion Picture Expert Group
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MTI	Medical Text Indexer system
NLP	Natural Language Processing
PACS	Picture Archiving and Communications Systems
PICO	Patient/Problem, Intervention, Comparison, Outcomes
PSO	Particle Swarm Optimization
RIDeM	Repository for Informed Decision Making
RoC	Receiver Operating Characteristic
ROI	Region of Interest
RSNA	Radiological Society of North America
SPIRS	Spine Pathology and Image Retrieval System
SVM	Support Vector Machines
TREC	Text REtrieval Conference
UMLS	Unified Medical Language System

Combining text and visual features for biomedical information retrieval

1 Introduction

The search for relevant and actionable information is key to achieving clinical and research goals in biomedicine. Biomedical information exists in different forms: as text and illustrations in journal articles and other documents, in "images"[1] stored in databases, and as patients' cases in electronic health records. Our objectives in this project may be formulated as seeking better ways to retrieve information from these entities, by moving beyond conventional text-based searching to combining both text and visual features in search queries. The approaches to meeting these objectives use a combination of techniques and tools from the fields of Information Retrieval (IR), Content-Based Image Retrieval (CBIR), and Natural Language Processing (NLP).

Our first objective is to improve the retrieval of biomedical literature by targeting the visual content in articles, a rich source of information not typically exploited by conventional bibliographic or full-text databases. We index these figures (including illustrations and images) using (i) text in captions and where they are mentioned in the body of the article ("mentions"), (ii) image features, and, if available, (iii) annotation markers within figures such as arrows, letters or symbols that are extracted from the image and correlated with concepts in the caption. These annotation markers can help isolate regions of interest (ROI) in images, the ROI being useful for improving the relevance of the figures retrieved. It is hypothesized that augmenting conventional search results with relevant images offers a richer search.

Figure 1: Reaction to intradermal adalimumab 1 to 2 days after the fourth dose

For example, in scientific publications, images are used to elucidate the text and can be easily understood in context. For example, Figure 1 and its caption are fairly informative in the context of the paper [1] "Eosinophilic cellulitis-like reaction to subcutaneous etanercept injection". Taken out of context, the caption provides little information about the image, and the image does not provide enough information about the nature of the skin reaction. This example illustrates both the problem of finding text that provides sufficient information about the image without introducing irrelevant information, and the potential benefits of combining information provided by the text and image. An even greater problem is determining what information about and in an image is sufficient for clinical decision support.

Sandusky and Tenopir find as an outcome of their survey [2] exploring the value of indexing and providing access to figures and tables along with the citation that:

[1] In the context of this work, an "image" includes not only biomedical images, but also illustrations, charts, graphs, and other visual material appearing in biomedical journals, electronic health records, and other relevant databases.

> *"Scientists find free text searching of abstracts or full text frustrating because results sets often include articles in which the query terms are not central to the article's purpose. ... Scientific journal-article components such as tables and figures are often among the first parts of an article scanned or read by a researcher after obtaining the complete text of the article. … The presence of individual figure and table components in the results set along with a collection of thumbnails in the enhanced abstract brings additional, highly salient information to the user prior to examination of the article's full text."*

Taking the retrieval of biomedical literature a step further, within the first objective our goal is to find information relevant to a patient's case from the literature and EHR databases and then link it to the patient's health record. The case is first represented in structured form using both text and image features, and then literature and EHR databases are searched for similar cases.

Our second objective is to find semantically similar images in image databases, an important step in communication of public health messages[1] and differential diagnosis. We explore approaches that automatically combine image and text features in contrast to typical visual decision support systems (for example, VisualDx®) that use only text driven menus. Such menu driven systems guide a physician to describe a patient and then present a set of images from which a clinician can select the ones most similar to the patient's, and access relevant information manually linked to the images.

Our methods use text and image features extracted from relevant components in a document, database, or case description to achieve our objectives. For the document retrieval task, we rely on the Essie search engine. Essie is a phrase-based text search engine with UMLS®-based [3] term and concept query expansion and probabilistic relevancy ranking that exploits document structure. To use Essie, we create structured representations of every full-text document and all its figures. These structured "documents" presented to the user as search results include typical fields found in MEDLINE® citations (e.g., titles, abstracts and MeSH® terms), the figures in the original documents, and image-specific fields extracted from the original documents (such as captions segmented into parts pertaining to each pane in a multi-panel image, ROI described in each caption, and modality of the image). In addition, patient-oriented outcomes extracted from the abstracts are provided to the user.

Automatic image annotation and retrieval objectives can be achieved in the following ways: (i) using image analysis alone [4]; (ii) by indexing the text assigned to images [5,6]; and (iii) using a combination of image and text analysis [7]. One approach is to compute image similarity [8], the traditional CBIR task of finding images that are overall visually similar to a query image, using machine learning classifiers [9] (e.g., Support Vector Machine) and fusion of class probabilities. These classifiers are trained on a variety of image features such as wavelets, edge histograms and those recommended by the MPEG-7 committee[2]. Additional steps include describing an image by automatically detecting its modality (for example, CT, MRI, X-ray, ultrasound, etc.) and

[1] To support communication of public health messages, the Centers for Disease Control and Prevention (CDC) provides a universal electronic gateway to CDC's pictures – Public Health Image Library (PHIL) http://phil.cdc.gov/phil/about.asp

[2] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

generating "visual keywords", i.e., text keywords assigned to patches in an image. These visual keywords are used to find similar images, and then IR techniques (e.g., tf-idf) are used on the visual keywords to improve the relevance of visually similar images. We are also exploring methods to automatically detect and recognize overlays on images (arrows, text labels) as a means to correlate image ROIs with concepts extracted from the image caption.

To prepare documents for indexing and retrieval, we combine our tools and those publicly available in a pipeline that starts with acquiring data and ends in generation of citations enriched with image-related information (henceforth, "enriched citations"). The initial separate text and image processing pathways merge in image annotation and multimodal indexes for use with specialized multimodal information retrieval algorithms (See Figure 2). The images and text data used for processing are obtained from different sources. For example, research toward improving access to biomedical literature is supported by full-text archives such as PubMedCentral®[1] and BioMedCentral[2]. The initiative to aid differential visual diagnosis uses images from annotated image collections and images published in the literature.

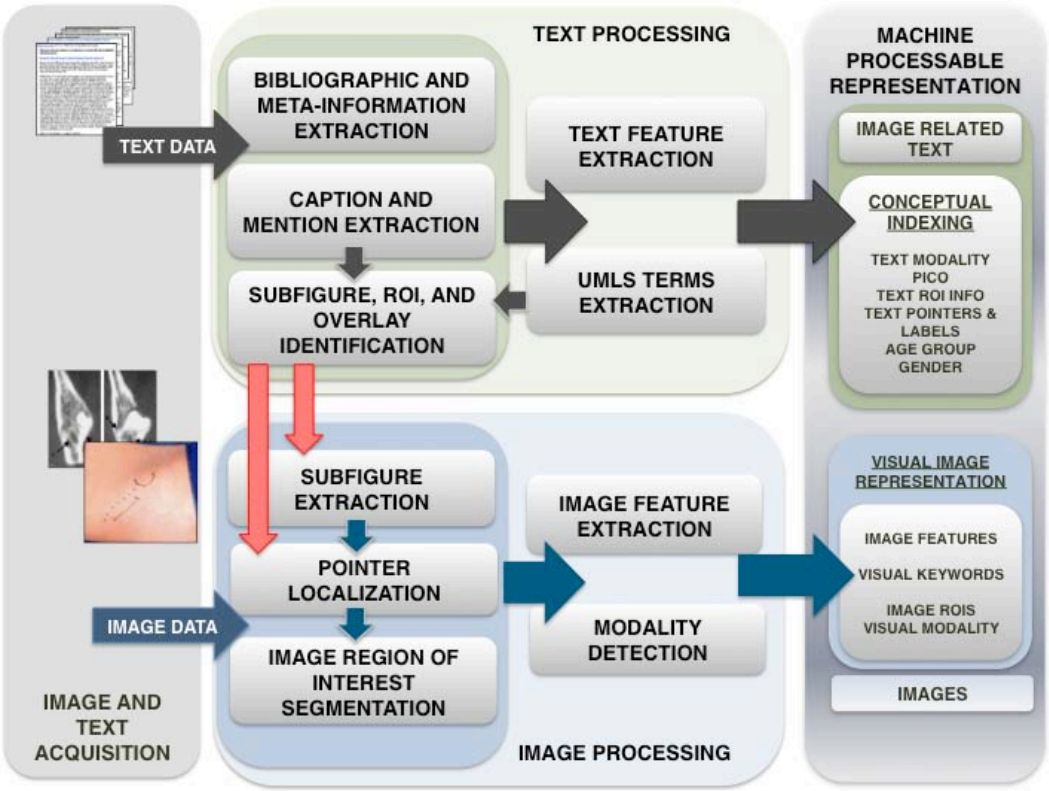


Figure 2: Overview of image and text processing steps for creating enriched citations. In the context of this work, an "image" includes not only biomedical images, such as CT, MRI, X-ray, and other modalities, but also illustrations, figures, charts, graphs and other visual material appearing in biomedical journals, electronic health records, and image databases.

[1] http://www.ncbi.nlm.nih.gov/pmc/
[2] http://www.bmc.org

To evaluate and demonstrate our techniques, we have developed the Image and Text Search Engine (ITSE), a hybrid system combining Essie with CEB's image similarity engine. Using this framework we explore alternative approaches to the problem of searching for information using a combination of visual and text features: (i) starting a text-based search of an image database, and refining the search using image features; (ii) starting a visual search using the (clinical) image of a given patient, and then linking the image to relevant information found by using visual and text features; and, (iii) merging the results of independent text and image searches.

These techniques were tested in the medical retrieval tasks of the ImageCLEF 2009 contest. Our approaches were shown to be the best in two of three categories (image retrieval using only visual features, and case retrieval) and in the top four for ad-hoc retrieval among over a dozen teams from around the world, including several from the industry.

This report is organized as follows. Section 2 briefly describes related research by other investigators. This is followed by Project objectives and significance. Our two research initiatives are described in Sections 5 and 6. In the Appendices, we describe image processing and text processing methods and tools that are common to the initiatives. Image processing methods are discussed in Appendix A, and the text processing steps appear in Appendix B. Our Image and Text Search Engine (ITSE) is discussed in Appendix C.

2 Related Work

Several ongoing research efforts are dedicated to augmenting text results with images. Some of these efforts aim to retrieve images by matching query text terms in the citations to the articles and the figure captions. We list five efforts related to our goals. A comprehensive study of other text and image retrieval search engines is covered in a CEB internal report [10]. Most systems do not use image features to find similar images or combine visual and text features for biomedical information retrieval. Our goals include improving relevance of multi-modal (text and image) information retrieval by including lessons learned from these efforts.

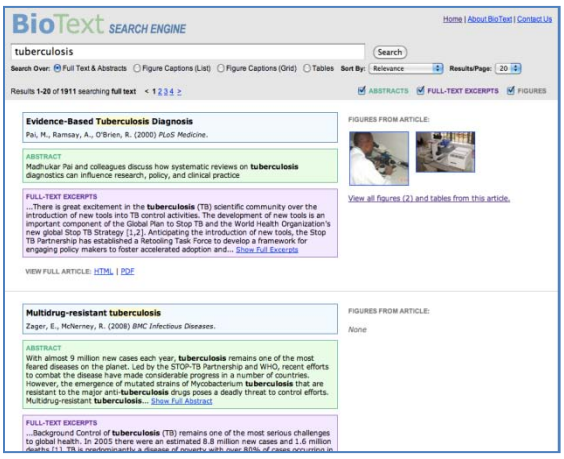


Figure 3: BioText search engine from University of California at Berkeley searches full text, figure captions, and table captions and presents retrieved results in various layouts.

The BioText[1] [11] search engine, shown in Figure 3, searches over 300 open access journals and retrieves figures as well as text. BioText uses the Lucene text search engine[2] to search full-text or abstracts of journal articles, as well as image and table captions. Retrieved results (displayed in a list or grid view) can be sorted by date or relevance. This search engine has influenced our user interface design.

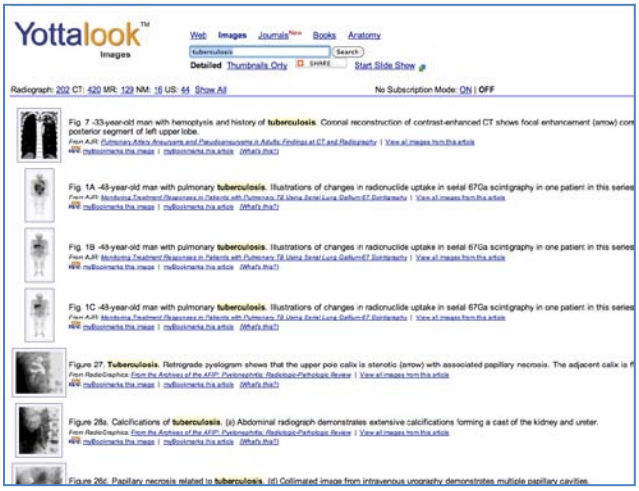


Figure 4: Screenshot of the YottaLook search engine. It searches the Web, image databases, journal articles, and books and teaching files for relevant text or image content.

Yottalook[3] allows multilingual searching to retrieve information (text or medical images) from the Web and journal articles. The goal of the search engine (shown in Figure 4) is to provide information to clinicians at the point of care. The results can be viewed as thumbnails or details. This site sets an example in the breadth of its searches, capabilities to filter results on image modality and other criteria, being current with social media, and connecting with the users' myRSNA accounts (offered by the Radiological Society of North America -- RSNA) that allows saving search results.

Other related work includes the Goldminer[4] search engine developed by the American Roentgen Ray Society (ARRS) that retrieves images by searching figure captions in the peer-reviewed journal articles appearing in the RSNA journals *Radiographics* and *Radiology*. It maps keywords in figure captions to concepts from the Unified Medical Language System® (UMLS) Metathesaurus®. Users have the options to search by age/modality/sex for images where such information is available. Results are displayed in a list or grid view.

The FigureSearch search engine, a component of the askHermes[5] system [12], uses a supervised machine-learning algorithm for classifying clinical questions and the Lucene search engine for information retrieval. Ad hoc clinical questions posed by users of the Web site are classified into queries using a Naïve Bayes classifier and logistic regression. The search engine searches

[1] http://biosearch.berkeley.edu/
[2] http://lucene.apache.org/
[3] http://www.yottalook.com/index_img.php
[4] http://goldminer.arrs.org/
[5] FigureSearch by askHermes http://snake.ims.uwm.edu/articlesearch/index.php?mode=figure.

published medical literature to generate a list view of the results with relevant images, abstracts, and summaries.

The Yale Image Finder (YIF)[1] [13] searches text within biomedical images, captions, abstract, and title to retrieve images from biomedical journal papers. YIF uses optical character recognition (OCR) to recognize text in images in both landscape and portrait modes.

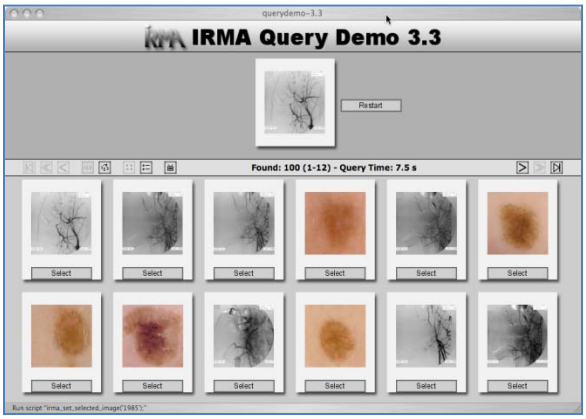


Figure 5: Screen capture of the Image Retrieval for Medical Applications (IRMA) system developed at Aachen University RWTH. IRMA uses image features to compute visual similarity between medical images.

The IRMA system[2], developed at Aachen University of Technology, Germany, aims to integrate text and image-based features for medical image retrieval. The system, shown in Figure 5, primarily uses visual features, but uses a limited number of text labels that describe the anatomy, biosystem, the imaging direction, and modality of the image. When medical images are categorized, they can belong to several different classes at the same time with different probabilities. We have collaborated with the developers of the IRMA system, and enhanced their image retrieval system (that uses features computed on the gross image) with our image features and similarity computation techniques applied to local image regions. This geographically distributed multi-scale image retrieval system [14] has been recognized by the Internet2 consortium with its IDEA Award in 2008[3] and our paper describing its application to spine image retrieval was selected as a best-paper finalist in MEDINFO 2007 [15].

Commercial Systems

There is increasing commercial interest in multi-modal information retrieval in the biomedical domain as evidenced from the teams participating in the ImageCLEFmed contests. Participants include researchers from Siemens, GE Medical Systems, Xerox, and other industrial organizations. Publishers such as Springer also provide a text-based image retrieval Web site[4] that searches figure captions and retrieves images from various journals published by Springer.

[1] http://krauthammerlab.med.yale.edu/imagefinder/
[2] http://www.irma-project.org
[3] https://lists.internet2.edu/sympa/arc/i2-news/2008-04/msg00005.html
[4] http://www.springerimages.com/

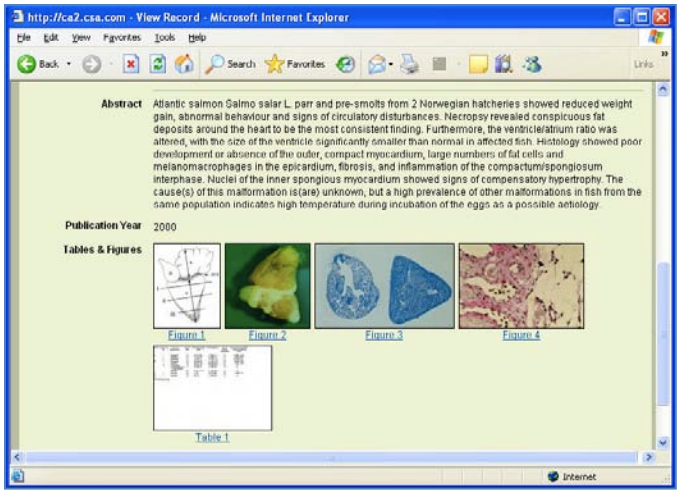


Figure 6: Screen capture showing ProQuest's Illustrata search engine that shows thumbnail images of all figures in the retrieved articles. This example shows images from an article in their life-sciences collection.

ProQuest, a major provider of access to more than 125 billion digital pages of the world's scholarship in medicine, sciences, technology, business, and other disciplines, provides a search tool called Illustrata (shown in Figure 6) that makes searchable tables, figures, graphs, charts and other illustrations from the scholarly research and technical literature. They assert that "*because of the visual impact of the results …scientists… can quickly determine whether or not to spend time reading the source documents*". While their effort is limited to searching for figures using only text queries, this vision captures the promised benefits of our research.

Other commercial image search engines include those developed by Google[1], Gazopa[2], and Flickr[3]. None of these use a combination of text and image features.

3 Project Objectives

Our objectives in this project may be formulated as seeking better ways to improve information retrieval from collections of full-text biomedical articles, images, and patient cases, by moving beyond conventional text-based searching to combining both text and visual features to:

1. Build text processing and image processing tools to index images and image-related text, and enable searching of the literature by textual, visual and hybrid search queries.
2. Build tools employing a combination of text and image features to enrich traditional bibliographic citations with relevant biomedical images, charts, graphs, diagrams and other illustrations, as well as with patient-oriented outcomes from the literature.

In addition to developing these tools, we propose to test them in two related initiatives that seek to:

1. Improve the retrieval of the biomedical literature by targeting the visual content in articles. Within this broad goal, we initially focus on finding information relevant to a

[1] http://images.google.com
[2] http://www.gazopa.com
[3] http://www.flickr.com

patient's medical case in the literature, and then linking it to the health record, and clinical question answering.

2. Improve the retrieval of semantically similar images from the literature and from image databases, with the goal of reducing the "semantic gap" that is a significant hindrance to the use of image retrieval for practical clinical purposes.

4 Project Significance

There is considerable evidence, some of it cited in the introduction, for a strong need to supplement traditional bibliographic citations with relevant visual material. Database services required to deliver such information would be essentially unaffordable if they are to be manually created. The automated techniques outlined in this report offer building blocks for the development of advanced information services that enable users to search by textual as well as visual queries, and retrieve citations enriched by relevant images, charts, graphs, diagrams, and other illustrations, not only from the journal literature, but also drawn from patient records and independent image databases. In addition to promoting greater, and more targeted access to the biomedical literature, our techniques would enhance visual diagnoses and clinical decision support.

5 Initiative 1: Improve retrieval of biomedical literature

5.1 Background

Text-based approaches to retrieval of biomedical literature have been well researched. Specialized retrieval systems have been developed for retrieving biomedical articles. Retrieval of biomedical literature provided by most widely used specialized biomedical search engines (such as PubMed®) is based on bibliographic citations (the titles and abstracts of scientific publications and MeSH terms, among other metadata.) Large-scale evaluations of retrieval of biomedical text within the TREC Genomics track showed that citation-based retrieval has achieved considerable sophistication, and that further significant improvements will require additional sources of information. The Genomics track turned to investigating the value of the full text of scientific articles, and demonstrated that automatic indexing of the entire text does not necessarily lead to significant improvements in retrieval [16]. However, there is evidence that augmenting MEDLINE citations with other relevant text can improve retrieval. For example, figure captions were instrumental in finding documents containing experimental evidence and discussions of the Drosophila genes and their products [17]. In this work, Regev et al. noticed that the evidence is often in the figures and used captions as substitutes. Shatkay et al. examined the possibility of integrating information derived directly from image data with text for biomedical document categorization, and concluded that this method has potential [18]. Also, Divoli [19] et al

> *"found evidence … that bioscience literature search systems such as PubMed should show figures from articles alongside search results. … Full text and captions should be searched along with the article title, metadata, and abstract. Finally, for a subset of users and information needs, allowing for explicit search*

within captions for figures and tables is a useful function, but it is not entirely clear how to cleanly integrate this within a more general literature search interface."

These investigations suggest strongly that figure captions and information derived directly from image data should improve retrieval of literature.

Use of images and their associated text in providing evidence for clinical decision support has yet to be evaluated in the context of information retrieval. To that end, the goals of our initiative are threefold: 1) determine information needs for which searching enriched citations is beneficial; 2) further explore the integration of information derived directly from image data with text for retrieval purposes; and 3) determine how best to display search results for different users and information needs.

5.2 Methods

To address these three goals, we developed a search engine ITSE (described in Appendix C.) At present, we restrict the information needs to answering clinical questions and linking relevant biomedical literature to patients' cases. As first steps in linking relevant biomedical literature to patients' cases, we participated in the ImageCLEFmed[1] case-based retrieval task. Case-based retrieval refers to the task of automatically finding clinical case reports that are similar to a given patient's case. In their pilot task introduced in 2009, participants were given clinical case descriptions and asked to retrieve the most relevant related cases and supporting articles.

In our approach to answering clinical questions and the case-based retrieval task, we represent the articles in the collection by enriched citations (as described in Appendix B.4). We use MetaMap to extract the UMLS concepts from the questions and case descriptions to form queries (as described in Appendix B.5). Our preliminary results achieved a Mean Average Precision (MAP) of 0.34 for the case-based information requests (the highest in the 2009 case-based retrieval evaluation) which indicated that our approach warrants further research to better understand the utility of incorporating image-related text into our enriched citations.

The bottleneck in this research is the lack of test collections for developing the approaches. To study the utility of enhanced citations, we need a collection of clinical questions, patients' cases and documents with at least partial judgments on their relevance to the questions and cases. Such collections do not yet exist in the public domain. Therefore, the first essential step in this research was to create a test collection.

5.2.1 Creating a test collection using "found data"

Fortunately, the American Academy of Family Physicians (AAFP) founded the Clinical Inquiries (CI) network that accepts clinical questions submitted by physicians and provides high-quality peer-reviewed answers. A CI article usually poses a brief clinical question and then summarizes an evidence-based answer using knowledge from supporting references. We considered articles in the reference sections of the publications that answer each clinical question to be relevant to

[1] http://www.imageclef.org/2010/medical

the question and marked each cited reference judged relevant. We mined the online version of the Journal of Family Practice[1] for 50 of the most recent publicly available questions having at least two cited articles that we could obtain using NCBI E-Utilities. To obtain relevant documents for our test collection, we began by downloading the full text HTML Clinical Inquiries articles from the Web site of the journal. We then parsed the HTML documents and extracted the list of references from each article. We used the NCBI ESearch utility to find PubMed identifiers (PMIDs) of as many references as possible, and then downloaded the cited articles, using the ELink utility to obtain the primary LinkOut provider for each PMID, and added the articles to the collection.

In looking for resources to help us develop and evaluate case retrieval approaches, we found that the "Photo Rounds" (PR) articles in the Journal of Family Practice typically present a detailed description of a clinical case, along with relevant images. The first part of each article presents the case and the (clearly separated) rest of the article describes a differential diagnosis while providing supporting evidence with references. We augmented the Clinical Inquiries collection applying the methods described above to the Photo Rounds articles.

Table 1: The number of articles per source in the test collection.

Source	Articles
American Journal of Public Health	589
Annals of Family Medicine	129
Antimicrobial Agents and Chemotherapy	1411
Archives of Disease in Childhood	347
BMJ	331
Gut	353
Heart	441
Radiographics	1285
Radiology	4421
Thorax	308
Total	9561

We obtained 232 references for the CI articles (avg. 5 relevant articles per question) and 212 articles referenced by the 50 PR articles, averaging 4 relevant documents per information request. To approximate a real-life document collection, we added numerous other articles (that lack relevance judgments) from various sources to the collection. Table 1 enumerates the articles obtained from each additional source. With the exception of the Radiology and Radiographics journals, we downloaded from the journals' Web sites the full text HTML of all articles from the two most recent complete years of publication (2008–2009). The articles from the two radiology journals were obtained through participation in ImageCLEFmed 2009.

[1] http://www.jfponline.com/

5.3 Experiments and Results

To study the effect of image-related text, we separately indexed the citations and image-text enhanced citations with the text search engine part of ITSE (i.e., Essie). We extracted the elements of the clinical scenario from the text of each clinical question and case description, and then created the type-based and concept-based queries as described in Appendix B.5.

Since clinical questions are rather brief, only 65 terms could be extracted from the 50 questions (avg. 1 term per query). However, the case descriptions, being longer, yielded 1091 terms (avg. 22 terms per query). The queries were run against the two types of citations (traditional and enriched). The results were evaluated using the *trec_eval* package[1][2] developed for evaluation of retrieval results within TREC. Since the number of documents judged relevant is small in comparison to the size of our collection, we used the binary preference[3] (bpref) retrieval evaluation metric computed by *trec_eval*, which is more robust than Mean Average Precision when given incomplete relevance judgments [20]. We followed the method outlined by Smucker et al [21] to compute two-sided Fisher randomization tests in order to measure the statistical significance of our retrieval results. The randomization (or permutation) test is considered more reliable than the Wilcoxon signed-rank test and more general than the paired Student's t-test.

In evaluating the importance of image-related text, we sought to determine (1) whether the inclusion of image-related text improves document retrieval and (2) if the concept- and type-based queries produce significantly different retrieval results. Therefore, we performed 8 batch retrieval runs (2 retrieval tasks x 2 citation types x 2 query generation strategies) over our test collection.

Table 2 summarizes the batch retrieval results for the 50 ad-hoc clinical questions and the case-based retrieval results. The average bpref is given for the concept- and type-based queries on both the traditional and enriched citations. For the ad-hoc clinical questions, the concept- and type-based query strategies resulted in nearly identical average bpref scores, and there was no statistically significant difference in bpref with the inclusion of image-related text. Since there was on average only one term extracted from each clinical question, there was essentially no difference between the concept- and type-based queries for these information requests. For the case-based retrieval, the use of the concept-based query generation strategy resulted in a substantially higher average bpref than did the type-based strategy. Most notably, the average bpref on the enriched citations (0.738) was a 10% increase over the average bpref on the traditional citations (0.668) at the 0.0002 significance level (p).

[1] http://trec.nist.gov/trec_eval/index.html
[2] http://trec.nist.gov/trec_eval/index.html
[3] Bpref (which stands for binary preference) is a retrieval effectiveness metric designed for evaluations with incomplete relevance data. Bpref measures the effectiveness of a system on the basis of judged documents only. It is a function of the number of times the judged non-relevant documents are ranked above relevant documents.

Table 2: Retrieval results for the "Photo Rounds" (case –based retrieval) and "Clinical Inquiries" (retrieval for clinical question answering).

	Queries	bpref		%increase in bpref	Significance (p)
		Traditional citation	Enriched citation		
Case – based retrieval	Concept-based (all concepts ORed)	0.668	0.738	**10.40**	**0.0002**
	Type-based (concepts within semantic types ORed, semantic typed ANDed)	0.004	0.015	2.80	0.5027
Question answering	Concept-based (all concepts ORed)	0.784	0.793	1.21	0.5004
	Type-based (concepts within semantic types ORed, semantic typed ANDed)	0.689	0.707	2.67	0.1232

The improvement in case-based retrieval is likely due to the reliance on medical images for clinical case descriptions (which are often accompanied by, or refer to, images in order to provide evidence in support of a particular diagnosis.) This suggests that image-related text should be utilized to improve retrieval accuracy for case-based information needs. The second significant finding from this evaluation is that the concept-based query strategy achieved better retrieval results than did the type-based strategy, especially for the case-based information requests.

5.3.1 Evaluating the quality and usefulness of the extracted terms

To evaluate the quality of our term extraction methods, we enlisted a family physician trained in informatics (and whose opinions on determining accuracy of image indexing terms were found to be consistent with a group of experts [22]) to manually judge terms for correctness and "usefulness" in constructing a clinical query. The extracted terms were judged correct if the automatically extracted terms had the correct UMLS semantic type and negation status. The annotator also added useful terms (for each image) that were not extracted. To evaluate the usefulness of the extracted terms for retrieval, we performed batch queries for each case description. As shown in Table 3, our extraction algorithm achieved a precision of 0.83 at 0.76 recall in identifying useful terms.

Table 3: Quality and utility of extracted terms.

Used Terms	**Extraction**		**Retrieval**	
	Precision	Recall	bpref	%improvement in bpref
All extracted			0.738	
Only useful	0.835	0.762	0.792	7.343
Only correct	0.837	0.760	0.792	7.343
Useful & correct	0.776	0.748	0.792	7.343

In retrieving relevant cases, we saw a moderate 7.3% improvement in bref ($p < 0.05$) when terms that were neither useful nor correct were eliminated. This indicates our approach to case-based retrieval can benefit from improving the quality of our case representation.

5.4 Summary and Next Steps

The most significant result from our evaluation is that the use of image-related text significantly improved document retrieval for the case-based retrieval, but had little effect for the clinical question answering. Additionally, this result indicates that combining textual approaches with

techniques from CBIR is a promising direction for further improving case-based retrieval. We plan to further explore (i) starting a search with the text description of a case and refining the search using the images of a given patient; (ii) starting a visual search using the images of a given patient and refining the search using the text of the retrieved visually similar cases, (iii) merging the results of independent text and image searches.

Another finding is that the occurrence of a few specific terms in a query is more important to the retrieval of similar cases than co-occurrence of terms from several semantic types. Determining better automatic query strategies is another direction to be pursued in our future work. This direction includes developing methods for combining textual and image features in queries and obtaining user feedback to improve results obtained from initial automatically generated queries.

Further studies are needed to determine if ad-hoc retrieval in some other clinical decision support areas (for example, finding nursing procedures relevant to care plan development) will benefit from image-related text as much as our case retrieval and whether images will allow for a better user experience.

The study of improving user experience will include developing mechanisms for submitting patients' images to our search engine and personalization of search results.

Our preliminary studies and the test collection assembled for the studies provide a solid foundation for these near-future steps.

6 Initiative 2: Improve semantic image retrieval

6.1 Background

As presented in Sections 1 and 2, the importance of medical illustrations in clinical decision making has motivated the development of large databases of medical images, such as the Public Health Image Library (PHIL) and GoldMiner, as well as active research in image retrieval within the yearly ImageCLEF medical image retrieval tasks and by individual researchers. The challenge is to find images that are "semantically" similar, and not merely similar in appearance. There is increasing interest in semantic image retrieval where image semantics are derived solely from the visual appearance of an image. Biomedical image collections, however, present unique challenges, where subtle differences determine retrieval accuracy between otherwise highly similar images. For example, a PA chest x-ray of a tuberculosis patient appears overall very similar to a PA chest x-ray of a patient with interstitial lung disease, but retrieving both these images is incorrect for any requests more specific than for a chest x-ray with pathology. This gap, often referred to as the "semantic gap" [23], is a significant hindrance to the practical use of CBIR systems in clinical medicine. However, in situations where such a limitation may not be a shortcoming, such as matching of clothing fashions, shoes, or handbags, the visual similarity technique has found a home in a number of commercial systems.

Techniques to find images semantically similar to a diagnostic image or its textual description could use textual representations of images, image features, as well as combinations of text and image features. Text used to represent images includes: passages found around the image in a

Web page, image captions or other passages of image-related text found in scientific publications, text specifically written to describe the image, and conceptual indexing of images using controlled vocabularies. The free-text representations of images can be indexed using a search engine and searched in response to a user query. In fact, most currently available image search engines (see examples listed in Section 2) implement this technique.

Conceptual indexing, such as by manually assigning Medical Subject Headings to MEDLINE citations, has been shown to improve image retrieval results [24]. However, both the manual indexing and generation of appropriate conceptual models of medical images are labor-intensive and costly tasks. For example, Bell et al. comment on difficulties in modeling chest radiography for reporting and retrieval purposes [25]. It is therefore not surprising that automatic conceptual indexing comparable in quality to manual indexing is desirable and an active research area. Woods et al. [22] have demonstrated that MetaMap finds UMLS concepts on image-related text with a high probability of being judged as exact matches to terms assigned by medical experts. Kahn et al. [26] have shown a significant improvement in the recall and precision of concept-based radiology journal figure retrieval over simple keyword matching. Kammerer et al. [27] developed a Web portal providing access to image databases for medical students and found that a navigation structure based on the UMLS semantic network offers a quick and easy-to-use learning environment.

CBIR for biomedical uses has been studied extensively in academia and at research centers. The efforts focus on identifying subtle differences between images in homogenous collections that are often acquired as a part of health surveys or longitudinal clinical studies. Examples include image retrieval of spine x-rays [28,29,30] and image analysis and retrieval of uterine cervix images for tracking prevalence and progression of cervical cancer [31]. Other efforts include the IRMA search engine that explores application of CBIR in research hospital PACS systems [32], and use of textual and image features for image classification of scientific articles [33].

Progress in CBIR and image classification based on text in image captions has motivated our research into integration of image data for semantic image retrieval. The goal of this initiative is to find successful approaches to integrate text and image features for image representation (conceptual indexing) as a means to retrieve images for clinical decision support.

6.2 Methods

We are developing three approaches to semantic image retrieval: 1) retrieval using the UMLS-based conceptual indexing of images; 2) traditional IR methods applied to image representation; and 3) classification of images as relevant to query, by supervised machine learning.

6.2.1 Conceptual indexing

Representing an image at a level of granularity suitable for a particular purpose, is a first key step in the automatic representation of images using text and eventually merging image and text features into visual keywords. We define three representation levels:

- **coarse**, which characterizes the whole image along the axes of its modality, relation to a specific clinical task (utility), body location, and teaching quality.
- **medium**, which provides a detailed description of the image content;
- **specific**, which provides very detailed descriptions of clinical entities in an image.

We hypothesized that the controlled vocabularies for the coarse and medium level can be found in the existing biomedical domain ontologies, while specific-level terms are not included in the existing ontologies and often are familiar only to clinicians specializing in a narrow area of medicine. To test these hypotheses, we developed an annotation interface that allowed our team of clinicians to select a coarse-level textual image representation from a hierarchical display of controlled vocabulary extracted from the UMLS (Figure 7A). The interface displayed medium-level textual representations of images extracted from the image-related text using MetaMap (Figure 7B, C). In addition to evaluating the automatically extracted image indexing terms for their usefulness for image retrieval, clinicians were asked to add missing terms.

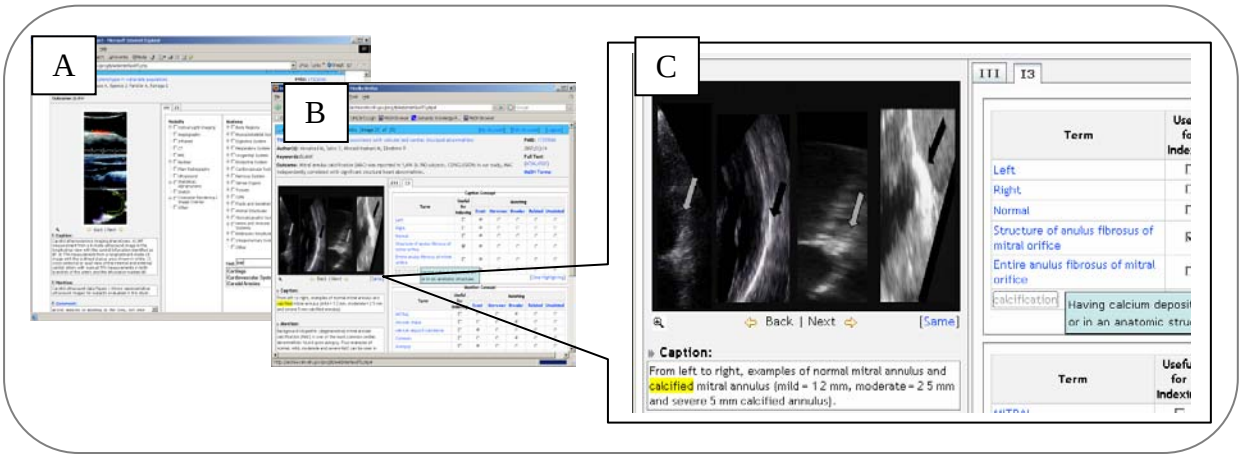


Figure 7: A Web-based application for image indexing evaluation: A. coarse-level image representation, B. medium-level image representation, C. a close –up of the UMLS concepts extracted from the caption

The evaluation interface, shown in Figure 7 was used by our team of clinicians (five physicians and one medical imaging specialist) who manually assigned missing specific terms, and evaluated the quality of medium-level indexing terms. The indexing terms were automatically extracted using MetaMap applied to captions and descriptions of 50 images randomly selected for each evaluator from all images published in the 2006 and 2007 issues of the BMC Annals of Facial and Plastic Surgery and European Journal of Cardiovascular Imaging. The judgments and additionally assigned terms were analyzed to answer the following questions:

1. Do captions and mentions of images in an article provide information beyond the indexing terms assigned by NLM indexers to the article?
2. Is the extracted text sufficient for image representation?
3. What are the coverage and accuracy of our automatic extraction method?

The first question was answered positively by intersecting the extracted terms evaluated as useful for imaging with the indexing terms assigned to the papers by NLM indexers and extracted from the bibliographic citations to the papers. There is some correlation between the MeSH terms assigned to a paper and image representation (around 30% overlap as shown in Appendix D.1.1), but only a small proportion of the MeSH terms could be used to describe an image. These terms do not describe the image completely, and additional indexing terms have to be extracted from the text.

The second question was answered by intersecting the terms additionally assigned by the evaluators with the full-text paper. Similarly to Declerck and Alcantara [34] (who identified the

title, caption, and abstract of a Web document to be the text regions possibly relevant to image representation) we found captions, mentions, abstracts, and titles of scientific publications to provide sufficient information for image representation. Only about 1% of additional terms could not be found in the text (details are provided in Appendix D.1.2).

The third question was answered positively in terms of recall, but our experiment indicated the need for selecting the terms that are most effective at describing the content of the image from the list of the potential indexing terms (details in Appendix D.1.3).

To that end, we used the 4,006 evaluated concepts (3,281 of which were unique), associated with 186 images from 109 different biomedical publications in a supervised machine learning approach aimed at reducing the number of automatically extracted ineffective indexing terms. Details on the method used to select image representation terms appear in Appendix D.2.

In parallel with our research of image annotation with text, we studied various approaches to image retrieval by enriching visual features with text concepts. There are two primary approaches to enhancing image features with semantic annotation: (i) extend the image feature vector to include text features, and (ii) assign the semantic labels to relevant regions of interest in the image in addition to the whole image. We have explored both methods [35], as described in Appendix A. Further, for each method of enriching the image index, we compared an image retrieval approach that utilizes a pipeline of text and image search engines to a supervised classification of images as relevant to a search query.

6.2.2 Supervised machine learning classification for image retrieval

The RapidMiner implementation of the SVM learner that was found to be most efficient in our earlier work on coarse-level image representation [36] was used for our current research in text-based and image-based representation. Each retrieved image was classified as to whether or not it was relevant to queries expressed using medium-level terms. The predictions of the text-based and image-based learners were combined using our own implementation of stacking [60]. Our stacking approach combines predictions from the base learners (e.g., SVM trained on text features) into a meta-level model (that combines the probabilities computed by the base learners into one prediction) using a version of the least squares linear regression adapted for classification [37].

Overall, the classification results demonstrated that the benefits of combining text and image features for the coarse-level image representation (which we observed in classifying images into six modality categories on a set of 554 images (73.66% average F-score [38])) can be extended to medium-level text representation. However, the supervised machine learning approach showed the most promising results when substantial numbers of training images are available (as shown in Section 6.3). We therefore concentrate on improving the information retrieval approaches that benefit from, but do not require hundreds of positively judged examples.

6.2.3 Information retrieval methods

In our information retrieval approach, we initially employed a pipeline approach to image retrieval. For this, we used (and compared) two open-source search engines, Lucene and Terrier[1], for indexing the set of the extracted text fields: captions, segmented captions, image mentions, article titles, abstracts and MeSH terms. We tested our approach by participating in the ImageCLEFmed 2008 contest in which each information request consisted of a text component and an image component. In the first step, we used the text component of the information request to retrieve images based on their associated text. For this we formed two types of queries: 1) information requests as provided in the ImageCLEFmed evaluation, and 2) expanded queries, in which image modality, findings, and anatomy terms were mapped to the UMLS Metathesaurus using MetaMap and supplemented with their preferred UMLS names and synonyms. For example, the expanded query for the information request *Show me MRI images of the brain with a blood clot*, included terms *Magnetic Resonance Imaging*, *MR Tomography* and other synonyms of the query term *MRI*, as well as *Thrombus* and other synonyms of the query term *blood clot*.

In the second step, the images that were retrieved using various search strategies applied to the text were re-ranked using image features. Based on features extracted from example query images, the images were automatically assigned to one of three broad categories: grayscale images (e.g., X-rays, CT, MRI, ultrasound images), color images (e.g., histopathology images, photographs), and other figures (e.g., graphs, charts, tables). This classification was done using color histogram analysis: grayscale images tend to have a simple histogram with almost no pixels that have different values for the Red, Green, and Blue channels; figure images tend to be bimodal with a greater number of white pixels than any other color; and the remainder are classified as color images that also tend to have a mixed histogram. The extracted query features and broad categories were compared to those computed for images retrieved in the first step (text-based retrieval) using the L2-norm. Retrieved images were then re-ranked according to their proximity to query images.

We use the textual image representation (described in Appendix B.4) that was developed in the above experiments and visual representations (described in Appendix A.2) to additionally research the following approaches to combining text and image features:
1. *Text-to-CBIR-query:* For each query, we first performed the textual search. We then manually selected 3–5 of the highest ranked retrieved images as relevant. We computed the mean vector of these retrieved images and used it as the query for the visual search.
2. *Text re-rank:* For each query, we first performed the textual search and then re-ranked the retrieved images based on the scores of the visual search.
3. *Interactive text-CBIR:* For each query, users manually selected relevant images from the top ten retrieved images of several text-, image-based, and combined retrieval results. We then selected additional query terms from the document representation of the relevant images, and used this expanded query as the input to the textual search. We ranked these additional images retrieved by the expanded query below the ones manually selected as relevant.

These approaches were evaluated in the ImageCLEF 2009 medical image retrieval task and compared to purely text- and image-based methods.

[1] http://terrier.org/

6.3 Experiments and Results

We evaluated image retrieval approaches using collections created in the medical image retrieval tasks in the 2008 and 2009 ImageCLEFmed contests. Retrieval results were evaluated using the *trec_eval* package, which computes Mean Average Precision (MAP), precision at different retrieval levels, and other metrics widely accepted in information retrieval research. Supervised machine learning results were evaluated using recall, precision, precision for five images classified with highest confidence as answers to a specific information request (P@5)[1], and F-score. Precision was computed as the number of images correctly annotated as relevant to the question divided by the total number of images automatically annotated as relevant. Recall was computed as the number of images correctly annotated as relevant by the classifier divided by the total number of images judged to be relevant to the question. P@5 was computed by sorting images in descending order of the classifier confidence scores, and then dividing number of images correctly annotated as relevant to the question within the five highest ranked images by 5. F-score was computed as the weighted harmonic mean of precision and recall.

Contributions of individual image-related text fields to image retrieval

Using the 2008 information requests we studied contributions of individual image-related text fields to image retrieval, and also compared the information retrieval and classification approaches to image retrieval. Table 4 presents MAP and precision at five retrieved documents (P@5) for image retrieval based on various combinations of segments of image-related text.

Table 4: Mean Average Precision (MAP) and precision at five (P@5) for 2008 medical image retrieval requests

Indexed text and query type (the request was used as supplied if query expansion is not indicated)	MAP		Precision @ 5	
	Lucene	**Terrier**	**Lucene**	**Terrier**
Short captions provided in the collection	0.151	0.045	0.347	0.200
Full captions	0.142	0.079	0.347	0.160
Segmented captions	0.149	0.081	0.353	0.167
Mentions	0.026	0.036	0.166	0.000
Captions and mentions	0.122	0.160	0287	0.386
Segmented captions + query expansion	**0.153**	0.082	**0.420**	0.200
Captions and mentions + query expansion	0.131	**0.169**	0406	**0.387**

The results of the information retrieval approach provide interesting insights into the nature and amount of text needed for a comparable performance of different information retrieval methods. Whereas the vector space model implemented in Lucene performed best on segmented captions[2], all extracted text was needed for comparable performance of the Terrier Inverse Document Frequency model with Laplace after-effect and normalization 2 (InL2), which we selected to gain early precision (boost mean precision at five retrieved documents). Although the Terrier InL2 model was not found to be sensitive to the variation in article length in several text collections [39], our results indicate that the model might not be suitable for document collections with shorter documents (averaging 66 words), and is comparable to the vector space

[1] The P@5 metric is particularly meaningful for clinical decision support since it may be assumed that a user, when presented with alternatives (as in Google search), can select the best one, but does not have time to inspect more than five – ten retrieved images.

[2] Segmented captions are sections of captions pertaining to individual image panels extracted as described in Appendix B.1.

model for the collections with longer documents (averaging 149 words). This indicates that even the best off-the-shelf search engines may not perform as well as search engines designed for a specific domain (e.g., medicine).

Notably, information contained in the descriptions of images in the body of the text is not sufficient for image retrieval and does not add value to captions when using the vector space model. The image retrieval component of our approach tends to be sensitive to the variety of features available in the image queries. Consequently, the results degraded when the example query images provided with the questions were too few in the image collection.

Comparing the information retrieval and classification approaches to image retrieval

The subset of 2008 medical image retrieval requests having 50 or more relevant images was evaluated in the supervised machine learning classification approach. The subset contained on average (per information request) 159 positive training examples, 616 negative examples, and 85 images randomly withheld for testing while still preserving the proportion of the positive and negative examples for each request. Table 5 presents average recall, precision, precision for five images classified with highest confidence as answers to a specific information request (P@5), and F-scores obtained for text-based and image-based classifiers and all possible combinations of the base classifiers. The representative stacking results are also shown here. Definitions of DWT and other features appear in Appendix A.2.

Table 5: Results of machine-learning approach to image annotation and retrieval averaged over all information requests (A), and requests with the training set containing over 180 positive examples (S)

Classifier: features	Precision		P@5		Recall		F-score	
	A	S	A	S	A	S	A	S
SVM: Segmented caption text (bag-of-words) TEXT BASELINE	0.341	0.588	0.443	0.714	**0.853**	**0.939**	0.488	0.723
SVM: DWT (Image)	0.135	0.270	0.057	0.057	0.429	0.856	0.205	0.410
SVM: Gabor filters (Image)	0.199	0.307	0.129	0,171	0.789	0.706	0.317	0.428
SVM: Color (Image)	0.202	0.315	0.171	0.343	0.817	0.778	0.324	0.449
Stacking: Text + DWT	**0.372**	**0.744**	**0.457**	**0.771**	0.424	0.848	0.396	**0.793**
Stacking: Text + Gabor filters	0.314	0.628	0.357	0.571	0.382	0.765	0.345	0.690
Stacking: Text + Color	0.344	0.688	0.457	0.714	0.426	0.852	0.380	0.761
Stacking: Color + Gabor filters	0.177	0.345	0.186	0.371	0.310	0.604	0.226	0.439
Stacking: all classifiers	0.310	0.618	0.329	0.571	0.394	0.788	0.346	0.692

The improvement in machine learning precision results for requests with more than 180 positive training examples is significant at the 0.05 level (SAS 9.1 npar1way procedure[1].) The difference in Mean Average Precision between the information requests included and excluded in machine learning experiments (shown in Table 6) is not statistically significant, which indicates there is no difference in the difficulty of the information requests (provided in the ImageCLEFmed 2008 contest) between the groups.

The difference in classification precision cannot be explained by the nature of the questions, as the better and worse performing questions were distributed evenly over question categories, complexity levels, and difficulty for retrieval measured by the average Mean Average Precision

[1] nonparametric tests for location and scale differences across a one-way classification

obtained for these information requests in the 2008 medical image retrieval evaluation [25]. The number of positive training examples could have influenced the machine learning results: all poorly performing questions had 100 or less positive training examples, whereas all better performing questions had between 189 and 288 positive examples.

Table 6: Lucene retrieval results (IR) for information requests included and excluded from machine learning (ML) experiments. The IR results for caption retrieval with query expansion (text) are shown.

Information requests	Features	IR		ML	
		MAP	P@5	Precision	P@5
Included in ML	Text	0.198	0.471	0.341	0.443
	Text + image	0.041	0.129	0.372	0.457
Best in ML (over 180 positive examples)	Text	0.202	0.400	0.588	0.714
	Text + image	0.043	0.200	**0.744**	0.771
Worst in ML	Text	0.098	0.271	0.095	0.171
	Text + image	0.019	0.029	0	0.143
Excluded from ML	Text	0.112	0.375	Information requests excluded from machine learning experiments due to lacking or insufficient positive examples	
	Text + image	0.033	0.100		
All information requests	Text	0.153	0.420		
	Text + image	0.039	0.119		

Combining text and image features for image retrieval

The multimodal relevance feedback approach (*Interactive text-CBIR* in Table 7) proved to be the best in this set of experiments, which indicates that adding a small user effort in providing feedback after an automatic initial retrieval improves image retrieval. Our result was one of the best in the 2009 medical image retrieval evaluation.

Table 7: Results of various approaches to combining image and text features for image retrieval

Approach	Recall	MAP	P@5
Interactive text-CBIR	0.65	**0.38**	**0.74**
Automatic Text Baseline	0.66	0.35	0.65
Text re-rank	0.66	0.27	0.49
Text-to-CBIR-query	0.21	0.04	0.28
Automatic Visual Baseline	0.12	0.01	0.09

The benefits of combining the text and image features are illustrated in the following example: Based on its caption alone, the image presented in Figure 8 was classified with low probability as relevant to the information request *MRI or CT of colonoscopy*. However, combining the low probability of relevance based on the textual features (0.268) with the higher probability of relevance based on the image features (0.453), the meta-classifier annotated the image as relevant with a probability of 0.891. The error in text-based annotation as well as text-based retrieval for this image can be explained by the vocabulary mismatch: none of the query terms can be found in the caption text. Even query expansion in the information retrieval approach was not helpful in this case because in the UMLS *MR* is not synonymous with *MRI*, and *colonoscopy* cannot be mapped to *colonography*.

Dark-lumen MR colonographic images in a 51-year-old man with history of colorectal carcinoma and end-to-end anastomosis. MR colonography also failed, because there was not sufficient water passing through the stenosis to permit adequate distention of prestenotic segments.

Figure 8. An image and its caption tested for relevance to the request: "MRI or CT of colonoscopy"

6.4 Summary and Next Steps

Our experiments show that titles, abstracts, captions and discussions of images in the full text of scientific publication contain enough image-related text to provide for conceptual indexing of images. The NLM resources (UMLS and MetaMap) allow extracting a substantial portion of indexing terms automatically and our filtering methods based on the elements of the clinical scenario (PICO) or supervised machine learning lead to improved precision of automatic indexing.

While we find that machine learning methods have the potential to achieve retrieval accuracy required for supporting clinical decision making, our results indicate that this accuracy level is achievable only when relatively high amounts of positive training examples are available. Therefore, in addition to seeking machine learning methods that require smaller training sets, we intend to explore the information retrieval approach. Our IR approach utilizes knowledge about useful image and text features accrued in the above experiments and focuses on ways to combine the features.

Our latest experiments show that applying knowledge about image representations gained in the earlier experiments led to significant improvements in retrieval results. We plan to test if the interactive retrieval (i.e., with user feedback) using ITSE will show improved results.

We also plan to further optimize visual feature selection. In addition, we will combine textual and visual representation by building an image ontology that will contain image features labeled with UMLS concepts. We are researching an approach to automatically generating the ontology using the ROI identified in the image related text and mapped to image regions.

We are also continuing our investigation of approaches to combining text and image features for retrieval. For example, we plan to explore the following pipeline: Start with text retrieval, identify images containing markers, identify ROIs, use image ROIs to retrieve another set of images combining local and global visual features, evaluate text related to new images and find associations between the initially retrieved text and the text retrieved through images. We will use strongly associated terms in the document collection to refine the search query.

7 Summary

Following evidence (Sandusky and Tenopir [2], and stated in the introduction) that enriching citations with relevant images can significantly improve literature retrieval for scientific research and clinical decision making, we have explored methods to combine biomedical image and text retrieval and developed an experimental search engine that combines the strengths of both. Our methods use text and image features extracted from relevant components in a document, database, or case description and create structured representations for them (the enriched citations). These enriched citations (that contain images and patient-oriented outcomes) are presented to the user as search results. Images are retrieved using image features and visual keywords developed to describe their content. The visual keywords are used to find similar images, followed by IR techniques to improve the relevance of the visually similar images retrieved. To evaluate and demonstrate our techniques, we have developed the Image and Text Search Engine (ITSE), a hybrid system that starts with a text-based search, and then refines the search using image features. Our approaches have been shown to be among the best in over a dozen teams from around the world participating in the ImageCLEFmed contests.

As next steps (beyond those mentioned in Sections 5.4 and 6.4), we continue exploring methods to improve the accuracy of retrieval of literature and images suitable for clinical decision support. The steps include:

1) Building a visual ontology by automatically detecting and recognizing pointers (arrows, text labels) and regions of interest in images and image-related text as a means to correlate image regions with UMLS concepts, and subsequently conceptually index images for retrieval.
2) Exploring methods to enable rich image queries (including the use of user-provided example images as queries).
3) Enriching short textual queries with additional information (such as the UMLS definitions of concepts identified in the queries and image features found in the visual ontology).
4) Improving extraction of the salient points from patient cases (for example, distinguishing between the findings present in the case description as part of routine examination and the chief complaints; extracting the details not covered by PICO, such as foreign travel, exposure to environmental factors, etc.)

In addition to the above informatics research steps, we are investigating scalability issues in indexing and retrieval of full-text articles and images from large publicly available collections, such as PubMedCentral.

Appendix A. Image Processing and Retrieval

In order to index an image by its content and to compute its relevance to a clinical query, it is first necessary to extract meaningful information from its visual content. This information can be extracted at several levels of detail from aggregated information gathered over the *whole* image to the features computed over *regions of interest (ROIs)* within the image, and finally down to image processing and feature extraction at *critical points* in an image. Steps in this process are illustrated in Figure A-1. Each block in this diagram is described in greater detail in the following sections. In Section A.1 we describe the process of identifying ROIs in images, first extracting subfigures from composite figures, and then finding useful "pointers" -- overlays (arrows, symbols, or text labels) that point to the ROI. Features used to represent the image content and other image indexing information is described in Section A.2 and retrieval techniques are discussed in Section A.3.

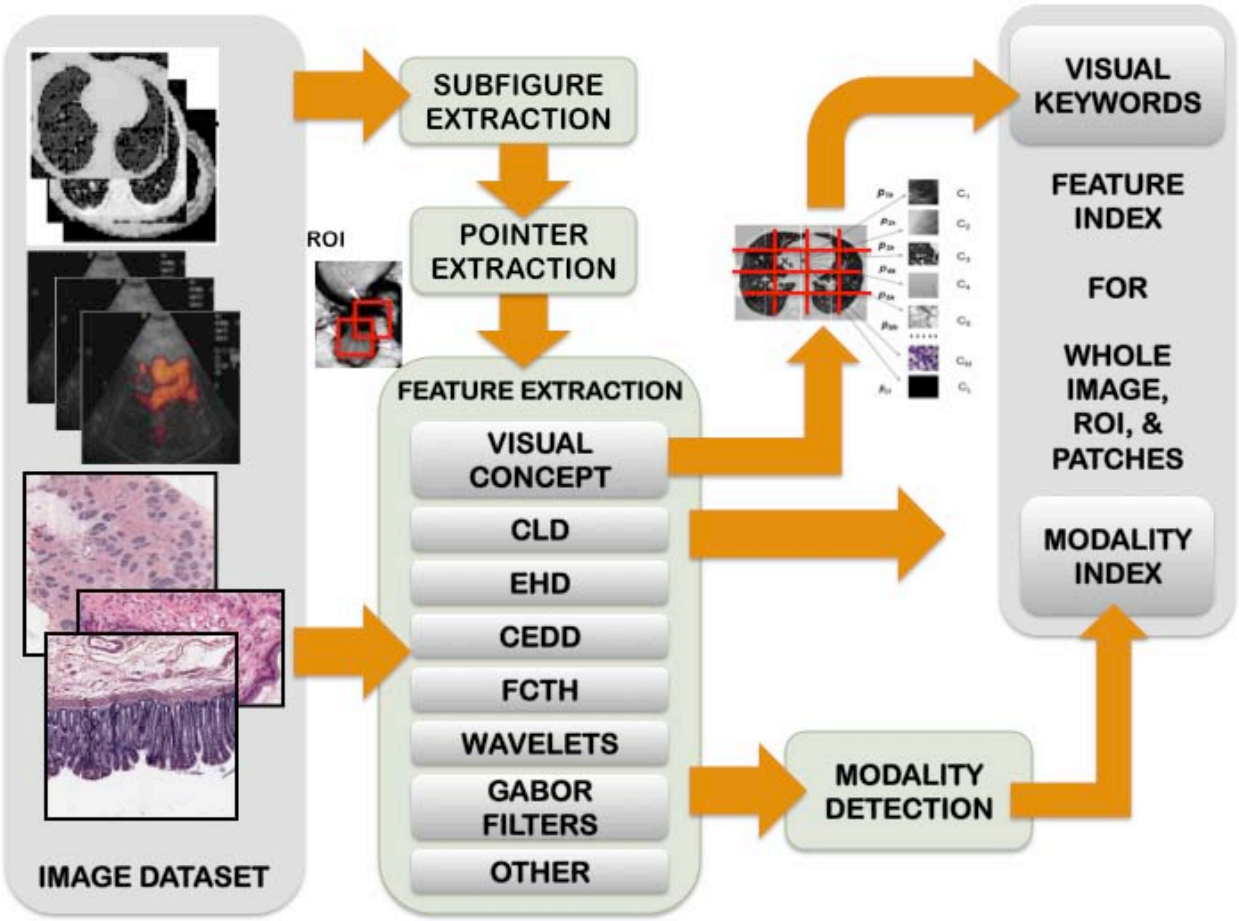
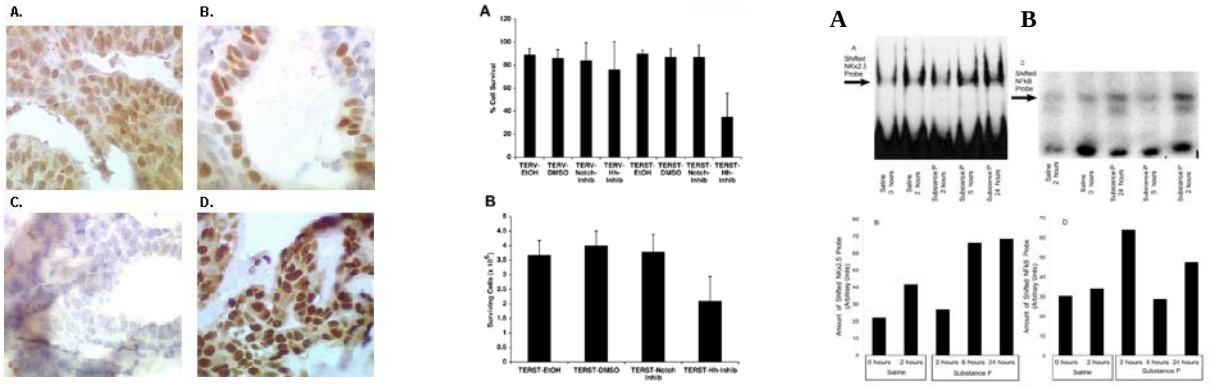


Figure A-1: Steps toward building an image feature index that supports concept-sensitive image similarity. Features include the Color Layout Descriptor (CLD), the Edge Histogram Descriptor (EHD), the Color Edge Direction Descriptor (CEDD), the Fuzzy Color Texture Histogram (FCTH), among others. Modality detection finds the imaging modality (e.g., CT, MRI, X-ray, Ultrasound, etc.) from the visual features.

A.1. Identifying Regions of Interest in Images

A.1.1. *Subfigure extraction*

In order to extract features from relevant image ROIs to implement CBIR techniques, it is often necessary to first separate figures into individual subfigure panels. As shown in Figure A-2, figures in biomedical articles often consist of several individual image panels, referred to by panel labels A, B, etc., that are combined into a single image by the author or publisher. The subfigure extraction step aims to automatically separate these panels into images that can then be used in the feature extraction step. The need to separate subfigures is clear from the example shown in Figure A-2 (c) where the subfigure panels A and B are images showing signal responses of a substance to saline for varying duration, while subfigure panels C and D show this response in a bar chart. Though the images (A,B and C,D) look alike, they are clearly different, and image features extracted to represent their content must be computed separately. As shown in the example, the author often places related images from different modalities as different subfigures, which are combined into a single image in the publication process. Correct image retrieval is only possible if the visual content expressed in the image is unimodal (e.g., all CT, MRI, or x-ray images).



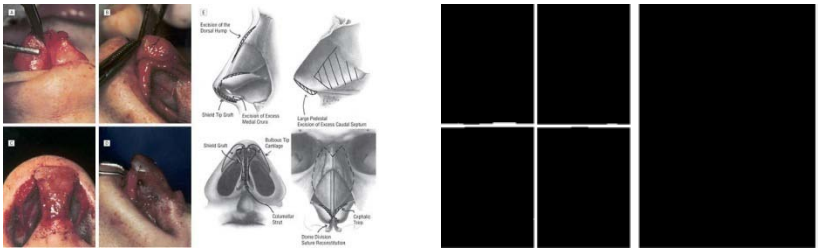
(a) (b) (c)

Figure A-2: Examples of different types of figures in articles (a) Typical biomedical images (b) bar charts (c) Mixed illustration.

To obtain unimodal images, we developed a heuristic two-phase algorithm [40] for the detection and decomposition of multi-panel figures that uses panel information predicted from figure caption text analysis as a guide. The algorithm is a heuristic decision tree that looks for strong white or black lines or a sharp transition between image panels. If these are found then the image is segmented along identified boundaries, and the algorithm is recursively applied to segmented panels until no further segmentation is possible.

Detection and decomposition of multi-panel images was tested on 516 figures extracted from the 2004 and 2005 issues of the British Journal of Oral and Maxillofacial Surgery. In this set, 427 images were single panel images and 89 were multi-panel. Overall 409 or 95.78% of the single panels and 84 or 94.38% of the multi-panel images were correctly identified. The method also corrected caption text analysis predictions for 6 of 84 multi-panel images. The method achieved 95.54% combined detection and decomposition accuracy. However, the method typically failed in cases where (i) inter-panel boundary width assumption exceeded our thresholds or (ii) there was a lack of a sharp transition between panels as is often in case of illustrations and charts. A

further limitation of the method is that it was heuristic and could not adapt to variations in the figure layouts, as illustrated in the right half of Figure A-3 (a). The left half of the figure (a) shows 4 panels of "regular" images and the right half shows 4 graphical illustrations. Figure A-3(b) presents a failure of the algorithm in separating all the subfigures. The four regular image panels are detected correctly, but the algorithm does not find strong boundaries among the four illustrations and fails to detect them as separate panels.



(a) (b)

Figure A-3: Subfigure detection algorithm example. (a) Original image. (b) Output showing detected subfigure panels.

To overcome this challenge, we have recently developed a novel Particle Swarm Optimization (PSO) [41] clustering and decision tree algorithm for the detection and decomposition of mixed multi-panel figures. The multi-panel illustrations shown in Figure A-4(a) do not have a clear inter-panel boundary but are nevertheless correctly identified by the new algorithm, as shown in Figure A-4(b).

(a) (b)

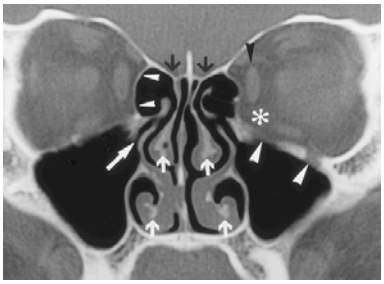
Figure A-4. Sample results from Particle Swarm Optimization for finding subfigure panels. Figure (a) shows the original illustrations. Figure (b) shows the identified bounding boxes.

The PSO clustering method uses cues extracted from the illustration, such as placement, size, and edge information to decompose individual components. The method is significantly less sensitive to the type of figures, i.e., regular images, graphical illustrations, charts, graphs, etc. Preliminary results from our ongoing evaluation of the new technique are very promising, with 99.2% accuracy for detecting and decomposing graphical illustration subfigure panel boundaries, and 93% for regular images. The method was tested on a set of 1443 figures that are roughly an equal mix of illustrations and regular images with ten-fold cross validation. Next steps include combining these methods to exploit the strengths of each.

A.1.2. *Pointer localization*

After a unimodal image is extracted, it is advantageous to extract ROIs within the image. Not only can this step help in better image understanding, but it also allows us to take advantage of

the author-annotated regions within the image that often are correlated with biomedical concepts extracted from the figure captions and mentions.



Images obtained in a 36-year-old woman with planned heart valve implantation who was examined for inflammatory paranasal disease. (a) CT scan depicts the anatomy of the infundibular (**large white arrow**) and parainfundibular complex very well. The bone structures of the inferior and middle conchae (**small white arrows**), the orbital lamina (**small white arrowheads**), the orbital floor (**large white arrowheads**), and the cribriform lamina (**black arrows**) are clearly delineated. Contrast between orbital fat (*****) and the medial rectus muscle (**black arrowhead**) is moderate. Inflammatory disease is not present.

Figure A-5: Example of an image and caption indicating presence of pointers and symbols (text marked in bold)

As illustrated in Figure A-5, annotations (markups) in the form of symbols, arrows, or text labels in images correlate with relevant text in figure captions or mentions in the article. Biomedical concepts in snippets of this text identified using the UMLS Metathesaurus and combined with image features from image regions indicated by such pointers (detected by image processing methods) can be used to improve (text and image) CBIR. The challenge in automatically localizing the pointers is to develop techniques that can recognize a large variety of symbol shapes with arbitrary sizes and locations, and that are robust against interference from the image background. The arrow-shaped markups may be coarsely categorized into three groups: (straight) arrows, curved arrows, and arrowheads. Other pointers include symbols such as asterisks or alphanumeric characters superimposed on the images.

Figure A-6. Variety of arrows (pointers) recognized by our algorithms.

As a first step toward the goal of localizing such image annotations, we have developed three alternative strategies [42,43] to detect and localize arrows in images. Examples of the variety of arrow shapes detected by our methods are shown in Figure A-6. Our first method defines an arrow as a set of edges that are organized in a particular sequence. It over-segments an image by applying a sensitive Canny edge detector and formulates the problem of recognizing arrows as a dynamic programming optimization problem. It uses Dynamic Time Warping (DTW) and is found to be more effective for straight arrows. Evaluation on 300 images resulted in an average accuracy of 75.3% in correctly identifying such arrows.

To generalize the algorithm to recognize curved arrows in any orientation, the method was coupled with a Markov random field (MRF)-based classifier framework. The classifier is trained on a large variety of arrow shapes. Further, the sequence of edge segments are modeled as a Hidden Markov Model (HMM) chain and the classifier detects pointers that have a strong contextual dependence among the edge segments. This method, though robust in most cases, is

susceptible to over-segmented (distorted) line segments that often result from highly sensitive parameters applied to image edge detectors, or when applied to images with a very noisy background, e.g., an ultrasound image. To address this problem, an Active Shape Model (ASM)-based classifier is used before rejecting a line segment classified as non-pointer by the HMM classifier. A preliminary evaluation of this multi-stage algorithm on 3000 images shows average recognition accuracy of 87% on a large variety of arrow shapes. An added advantage of these methods is that the algorithm is aware of the arrow-head and can use it to identify the ROI in the image.

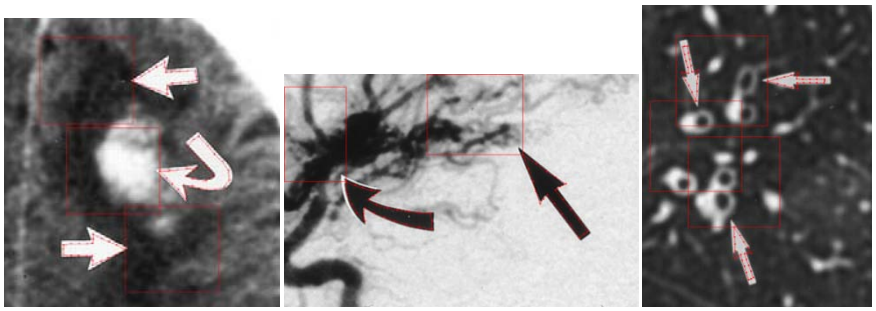


Figure A-7: Sample results from DTW-MRF-HMM-ASM pointer recognition algorithm.

Figure A-7 shows example results of our multi-stage pointer recognition algorithm. Solid red lines and points overlapped on each pointer show the best fitted pointer and its landmark points. Each red rectangle shows the ROI pointed to by each arrow.

A third strategy [44], in the early stages of development, aims to minimize background noise in the image by applying Particle Swarm Optimization-based clustering. It is very effective for managing weak pointer boundaries on complex image backgrounds. However, it does not isolate the arrowhead from the rest of the body. The initial results are very promising. The next steps include developing methods to fuse the three algorithms to further improve pointer recognition and to expand the set of recognizable pointers to include other symbols and alphanumeric characters.

A.2. Building the Image Feature Index

It has been shown that several *gaps* hinder the process of finding an image with the *best-fitting* content to a user query [45]. The challenges arise from inadequate visual features, difficulty in capturing the semantic context of the image, and multiple and ambiguous representations of the same object, to name a few. To bridge some of these gaps and improve the relevance of returned image matches we take advantage of well-defined characteristics of the biomedical imaging domain and relevant concepts from the UMLS Metathesaurus. The concepts are used together with visual features extracted from the whole image and at local regions of interest by algorithms that build an image feature index to support semantic visual similarity.

It has been well documented that low-level visual features, such as color, texture, and shape, are insufficient in capturing image semantics, though they are the primary building blocks of the visual content in an image. They can be effective if a judiciously selected feature metric is used to capture the visual content in an image, and then incorporated into a suitable machine-learning framework that supports multi-scalar and concept-sensitive visual similarity. We describe our

retrieval framework that supports these characteristics (in Section A.3) after describing various components used in the generation of a feature index.

A.2.1. *Feature Extraction*

Low-level features used to represent the visual content of the image are discussed below. These features are applied at multiple scales, i.e., on the *whole* image and at any relevant *regions-of-interest* identified by the pointer localization algorithm. These features include color, edge, texture, intensity, and other information. Various feature metrics are used for each image feature. Their contribution to the final visual similarity can be manually defined, or determined by a machine learning algorithm. The weighting of any particular feature may be further altered by user feedback on retrieved images.

Biomedical images are found in varying sizes determined by their format (DICOM CT, or MRI) or by the source, e.g., images available in GoldMiner collection tend to be very large and of different sizes. In order to obtain a uniform measure with greater computational efficiency we compute features from images reduced to a common size measuring 256 x 256 pixels. In the future, we intend to process images at a significantly higher (or full) resolution to extract meaningful local features.

Color Features: Color plays an important role in the human visual system and measuring its distribution can provide valuable discriminating data on the image. We use several color descriptors to represent the color in the image. To represent the spatial structure of images, we utilize the Color Layout Descriptor [46] (CLD) specified by MPEG-7 [47]. The CLD represents the spatial layout of the images in a compact form and can be computed by applying the discrete cosine transformation (DCT) on the 2D array of local representative colors in the YC_bC_r color space, where Y is the luminence component and Cb and Cr are the blue and red chrominence components, respectively. Each color channel is 8-bits and represented by an averaged value computed over 8 x 8 image blocks. We extract a CLD with 10 Y, 3 C_b, and 3 C_r components to form a 16-dimensional feature vector.

Another feature used is the Color Coherence Vector [48] (CCV) that captures the degree to which pixels of that color are members of large similarly colored regions. A CCV stores the number of coherent versus incoherent pixels with each color thereby providing finer distinctions than color histograms. Color moments, also computed in the perceptually linear $L^*a^*b^*$ color space, are measured using the three central color moment features: mean, standard deviation, and skewness.

Finally, 4 dominant colors in the standard RGB (Red, Green, Blue) color space and their degrees are computed using the k-means clustering algorithm.

Edge Features: Edges are not only useful in determining object outlines, but their overall layout can be useful in discriminating between images. The Edge Histogram Descriptor [46] (EHD), also specified by MPEG-7, represents a spatial distribution of edges in an image. It computes local edge distributions in an image by dividing the image into 4 x 4 sub-images and generating a coarse-orientation histogram from the edges present in each of these sub-images. Edges in the image are categorized into five types: vertical, horizontal, 45° diagonal, 135° diagonal, and other

non-directional edges. A finer-grained histogram of edge directions (72 bins of 5° each) is also constructed from the output of a Canny edge detection algorithm on the image. The feature is made invariant to image scale by normalizing it with respect to the number of edge points in the image.

Texture Features: Texture measures the degree of "smoothness" (or "roughness") in an image. We extract texture features [49,50] from the four directional gray-level co-occurrence matrices (GLCM) that are computed over an image. Normalized GLCMs are used to compute higher order features, such as energy, entropy, contrast, homogeneity and maximum probability. We also compute Gabor filters to capture image gist (coarse texture and spatial layout). The gist computation is resistant to image degradation and has been shown to be very effective for natural scene images [51]. Finally, we use the Discrete Wavelet Transform (DWT) that has been shown to be useful in multi-resolution image analysis. It captures image spatial frequency components at varying scales. We compute the mean and standard deviation of the magnitude of the vertical, horizontal, and diagonal frequencies at three scales.

Average Gray Level Feature: This feature is extracted from the low-resolution scaled images, where each image is converted to an 8-bit gray-level image and scaled down to 64 x 64 pixels regardless of the original aspect ratio. Next, this reduced image is partitioned further with a 16 x 16 grid to form small blocks of (4x4) pixels. The average gray value of each block is measured and concatenated to form a 256-dimensional feature vector.

Other Features: We extract two additional features using the Lucene image retrieval engine [52] (LIRE) library: the Color Edge Direction Descriptor (CEDD) and the Fuzzy Color Texture Histogram (FCTH) [53]. CEDD incorporates color and texture information into a single histogram and requires low computational power compared to MPEG-7 descriptors. To extract texture information, CEDD uses a fuzzy version of the five directional edge filters used in MPEG-7 EHD that are described previously. This descriptor is robust with respect to image deformation, noise, and smoothing. The FCTH uses fuzzy high frequency bands of the Haar Wavelet Transform to extract image texture.

A.2.2. *Visual Keywords*

It has been a goal of biomedical CBIR research to improve upon traditional CBIR methods that rely solely on low-level visual features to identify visually similar images. In medicine, very often we find that images that appear similar are not related at all. For example, a chest x-ray image exhibiting tuberculosis may appear similar to a chest x-ray image showing interstitial disease. Use of image features without any semantic interpretation tends to fail in distinguishing images from different semantic categories due to the limited discriminative power of the features. This shortcoming is often called the *semantic gap* in CBIR.

In an effort to minimize the semantic gap, some recent approaches have used machine learning on locally computed image features in a *bag of concepts* model. The term *concept* is used loosely and refers to a set of categorical labels. It is founded on the assumption that an image consists only of a small set of concept labels (e.g., its modality, imaged anatomy, etc.) that apply to the entire image. The bag of concepts image representation scheme is analogous to the *bag of words* representation used in retrieval of textual documents.

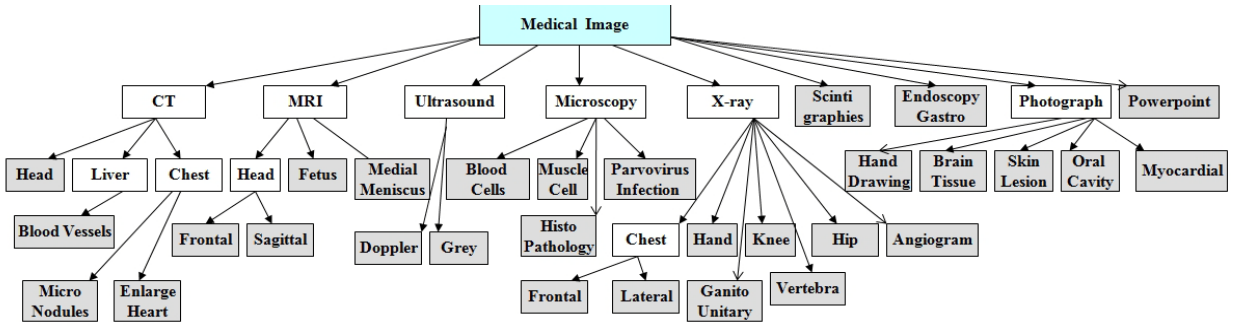
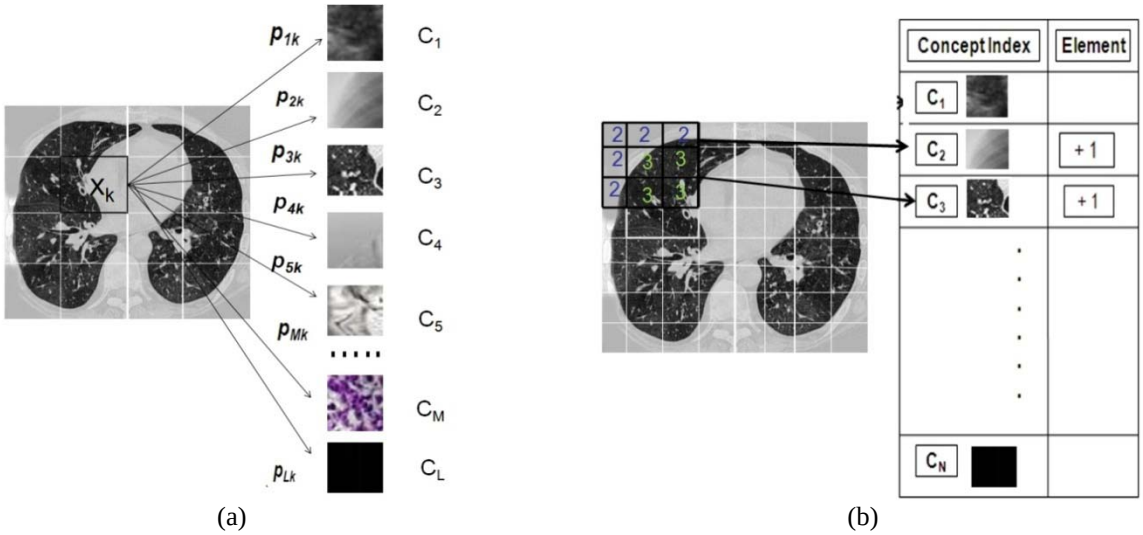


Figure A-8: Elements in an annotation hierarchy that can be used as "concepts" to annotate biomedical images. Shaded blocks are leaf nodes. A path from the root node to leaf follows the following order: image modality → anatomy → imaging direction (e.g. posterior-anterior).

We have developed a *visual keyword* hierarchical model [54] that expands the bag of concepts idea to annotate images with a set of labels that indicate the membership of local image regions in various image categories. The sets of *keywords* or *concepts* include text labels from the figure caption, the image modality, the imaged anatomy (body-part), imaging direction, and other information. A part of the hierarchical model is illustrated in Figure A-8.



(a) (b)

Figure A-9: Visual keywords associated with local regions in an image. This results in an image being associated with many classes (C_1, C_2, …, C_M) in varying amounts (p_{1k}, p_{2k}, …, p_{Lk}) that are determined as output probabilities by an SVM classifier.

The visual keywords are associated with local image "patches" that are generated by uniformly subdividing an image into an *r* x *r* grid of non-overlapping regions. The color and texture features of each patch are used in a supervised learning framework to train a SVM classifier to associate each image patch with a set of keywords, as illustrated in Figure A-9(a). The classifier is trained on feature vectors generated from a set of manually annotated images, each of which is associated with a single semantic label selected from a predetermined set of *M* labels or categories. The class probability of an image patch as belonging to one of a set of *M* labels (C_1, C_2,…, C_M) is computed. The set of these patches then characterizes the semantics of an image through confidence scores representing the weight of a category label in the overall description

of an image. With this framework, we make two advances over the generic bag of concepts model:

1. We introduce a probabilistic visual concept vector (PVCV) that models an image patch being related to all concepts in varying degrees rather than to just one concept. The degree to which it is related to a particular concept is modeled as the output probability generated by the SVM classifier. Further, this vector representation considers not only the similarity of different region vectors for different concepts, but also the visual dissimilarity of the region vectors that are mapped to the same concepts.

2. The second model is a structural feature representation scheme based on the observation that there are usually several concepts that are highly similar or correlated to the best matching one for a particular image region. For example, a region within a CT and an MRI image of the same anatomy can share not only visual similarity but also several concept labels. As a result, patches in these images are likely to have highly correlated concepts. To take advantage of this, the scheme spreads the region membership values (or confidence scores) in all local concept categories to the neighboring regions during the encoding and subsequent feature extraction process, as shown in Figure A-9(b). This has the effect of reinforcing the local concepts.

A.2.3. *Modality Detection*

Medical image retrieval from large collections can be made more effective and relevant to a query if it can be annotated with information about its imaging modality. In this work, "modality" refers to the imaging and/or representative form of the image, e.g., x-ray, CT, MRI, ultrasound, etc. Successful modality detection (or categorization) of images would enhance the performance of the CBIR system by reducing the search space to the set of relevant modalities. For example, to search for posterior-anterior (PA) chest x-rays with an enlarged heart in a radiographic collection, the database images can first be pre-filtered using automatic categorization by modality (e.g., x-ray), body part (e.g., chest), and orientation (e.g., PA) before any visual similarity between images in the database and the query image is computed. Image modality is detected by combining predictions from text-processing on image captions and mentions and image processing methods. Here we describe a method to determine the image modality using visual features.

Our method [55] uses an SVM to classify images into multiple modality categories. The degree of membership in each category can then be used to compute the image modality. In its basic formulation, the SVM is a binary classification method that constructs a decision surface and maximizes the inter-class boundary between the samples. To extend it to multi-class classification, we take the approach of combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC). The PWC method constructs binary SVMs for all possible pairs of classes. Hence, for M classes this method uses $M * (M-1) / 2$ binary classifiers, each of which provides a partial decision for classifying an image. The SVM is trained for each image feature. The class with the greatest estimated probability for each feature accumulates one vote. The class with the greatest number of votes after classifying for all features is deemed to be the winning class, and the modality category of the class is assigned to the image.

In addition to absolute voting, which is threshold dependent, we also consider four other popular classifier combination techniques derived from Bayes' theory: *product, sum, max*, and *mean*. The *posterior* probabilities of each category serve as class weights in the classifier output combination step.

A.3. Content and Concept-Based Image Retrieval

In image retrieval, the steps taken for indexing images in the multi-modal knowledge base are applied to the query image. In the simplest approach, features extracted from the query image are compared with the set of indexed features, and a list of images ranked in order of decreasing similarity to the query image is returned. However, for very large image and document collections it may be impractical to adopt a brute force approach of comparing features for all images. Instead, a reduced search space can be generated using relatively robust modality detection methods, described previously, to predict the image category. To reduce the risk of misclassification, a few highly (and closely) ranked category labels may be considered, instead of just one.

To maximize retrieval effectiveness, we consider data fusion or multiple-evidence combination strategies [56]. The similarity between a query image (I_q) and a candidate database image (I_j) is measured as a weighted linear combination of different features and expressed as

$$\text{Sim}\,(I_q, I_j) = \sum_F \omega^F \text{Sim}^F\,(I_q, I_j)$$

where $F \in \{\text{Concept, EHD, CLD, CCV, CEDD, FCTH, ... }\}$ and ω^F are feature weights. Different feature weights are assigned for different image categories. For example, a particular color feature may have greater weight for microscopic pathology and dermatology images than an edge or texture related feature that, in turn, may be emphasized more for radiographs.

To allow for user interpretation of image semantics, or correction of an erroneous category predicted by the classifier, we have explored methods for refining search results using relevance feedback. In our approach, the feature weights are updated to reflect the similarity rank for images that are marked by a user as *relevant*. Each marked image is then used as query in the new searches. The final rank for newly retrieved images is obtained through an adaptive and linearly weighted combination of individual similarities of the original query image and the images marked relevant.

Appendix B. Text Processing and Retrieval

To search a body of information (such as a collection of images or citations to the biomedical literature) for objects (scientific articles, images, case descriptions, etc.) relevant to a search query, we take the following steps: (1) automatically represent the body of information in the structured form required by our search engine (ITSE) described in Appendix C; (2) formally represent the information need submitted by a user or inferred from a patient's case using the Evidence Based Medicine framework of a well-formed question; and (3) translate the formal representation of the information need to a search query using the search engine query language. The body of information for our first research initiative (see Section 5) consists of full-text scientific publications, whereas in our second initiative (see Section 6) we process any type of free text associated with images stored in databases, or find image-related text in an article or case description that contains images.

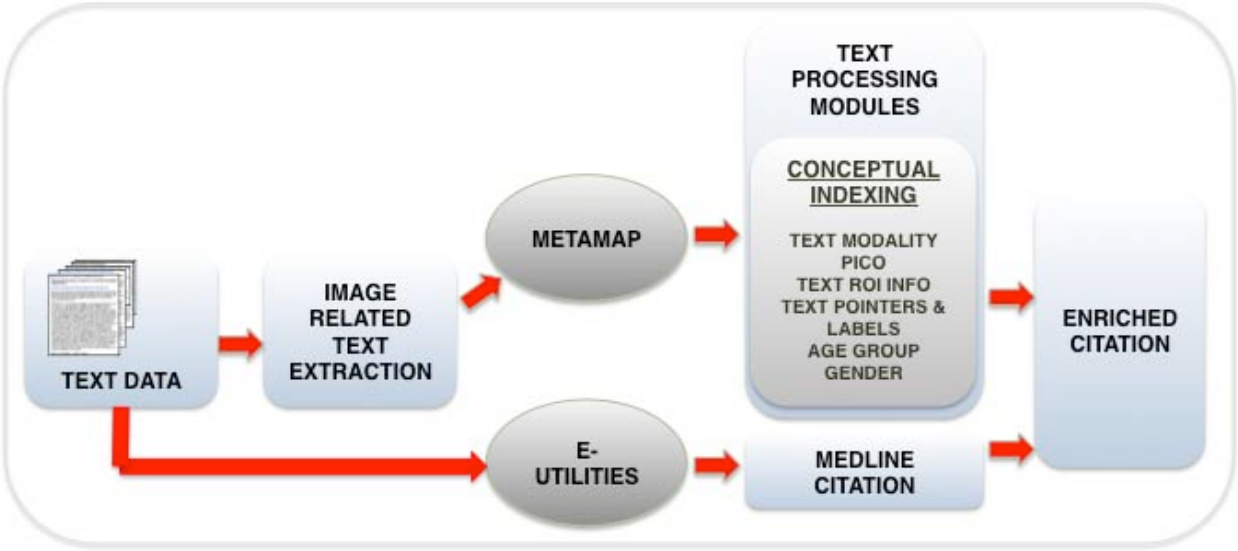


Figure B-1: Text processing pipeline.

To generate structured documents required by the ITSE search engine, we augment MEDLINE citations with image-related information extracted from the full text, creating "enriched citations". For publications that are not indexed for MEDLINE and other types of image-associated text (e.g., electronic medical records), we generate citation-like documents. The search queries range from typical short phrases provided by users to fairly long passages of clinical narrative. Figure B-1presents the text processing steps. The detailed description of each text-processing step follows.

B.1. Extracting image-related text: caption segmentation and mention extraction

The first step in generation of an enriched citation is finding image captions and mentions. In most cases, finding image captions amounts to parsing the corresponding XML tags in such documents as PubMedCentral articles or using regular expressions. For example, the pattern: `<BLOCKQUOTE>.*?(Figure\s*\d|Table\s*\d|Scheme\s*\d)(.*?)</BLOCKQUOTE>` can be used to find captions in the documents available only in HTML format. To process PDF documents, we convert the

documents to plain text using the ***pdftotext*** program and modify the regular expressions accordingly.

Despite the relative ease of finding captions, the task is somewhat more involved because the extracted captions often pertain to figures composed of several individual images (multi-panel images) or to a series of images, which requires segmenting the multi-part captions for image representation purposes and to inform the image segmentation algorithms. We developed several rule-based algorithms for caption segmentation that recognize and segment the following forms of multi-image captions: 1) an ordered list of independent meaningful descriptions of individual panels (sometimes preceded or followed by information pertaining to the whole figure), as shown in Figure B-2; 2) an ordered list of complementary information for individual panels preceded or followed by content-bearing information pertaining to the whole figure, as shown in Figure B-3; 3) a list of cross-referencing sets of descriptions, as shown in Figure B-4.

Cyclops lesion in a 45-year-old-man with 2-year history of chronic ACL tear and loss of extension.(a) Sagittal FSE proton-density image demonstrates the well-circumscribed, slightly heterogeneous cyclops nodule. (b) Coronal DESS image demonstrates a soft-tissue nodule in the anterior intercondylar notch.

Figure B-2: Type 1 multi-panel caption

Plasma concentrations (mean (SEM)) of (A) BNP, (B) NT-proBNP, (C) ANP, (D) NT-proANP, (E) CNP, (F) NT-proCNP and (G) cGMP in patients receiving BNP (nesiritide) or placebo after myocardial infarction. Infusion ran for 60 hours between measurements on days 1 and 4.

Figure B-3: Type 2 multi-panel caption

(a–c) Immediately postoperative remote cardiac MR images in 70-year-old woman with surgically repaired ventricular septal defect, which occurred as a complication of inferoseptal myocardial infarction (arrow). **(a)** Steady-state free precession cine MR image (repetition time msec/echo time msec, 3.4/1.2; flip angle, 60°) in vertical long-axis view shows focal defect. **(b)**Midchamber short-axis view of same defect in **a** shows that it had been oversewn and no residual interventricular shunt was present. **(c)** Delayed short-axis image (repetition time msec/echo time msec/inversion time msec, twice the R-R interval/4.3/280; flip angle, 30°) shows inferior wall hyperenhancement corresponding to site of transmural myocardial infarction. **(d, e)**Local cardiac MR images in 60-year-old man after ablation therapy for ventricular tachycardia.**(d)** Short-axis cardiac cine image (3.4/1.2; flip angle, 65°) demonstrates hypokinesis of lateral and inferolateral left ventricular wall. Arrow = segment of infarcted myocardium. **(e)** Short-axis delayed hyperenhancement image (twice the R-R interval/4.3/280; flip angle, 30°) shows scar (arrow) of lateral and inferolateral left ventricular myocardium. Both **d** and **e** were obtained by the same operator and have overall image quality comparable to that of **a–c**.

Figure B-4: Type 3 multi-panel caption

In these multi-panel caption types, list items are marked using sequences of upper or lower case letters (with or without parenthesis and other punctuation marks), digits or Roman numerals, and relative locations (top, bottom, right, left, middle, etc.) The markers may appear before or after the panel description. In addition, some list items contain nested lists.

Figure 2: Rash on the trunk of twin II. Parental/guardian informed consent was obtained for publication of this figure.
Mention: At 72 h of age, she developed a florid pustulo-vesicular and markedly pruritictruncal rash, with diffuse erythroderma in some areas (fig 2).

Figure B-5: Example of the figure caption and mention extracted from the text

Unlike captions, the descriptions of the figures in the body of the article (mentions) are natural parts of the text and are not indicated by tags. The description of the figure is always indicated by the word "Figure" or "Fig" (and, occasionally, within mark-up tags or punctuation) followed by a number (as shown in Figure B-5). The number allows associating the found passage to the figure. The boundaries of the extracted passage are determined as the paragraph (for the texts containing paragraph markers) or the sentence containing the indicator. If the sentence containing the indicator is shorter than five words, the preceding and the following sentences are also extracted.

B.2. Understanding image description: pointers and ROI

The extracted image-related text is further processed to identify Image Regions of Interest (ROI). ROIs are commonly described in the image caption and indicated by an overlay that facilitates locating the ROI. This is especially true for hard to interpret scientific images such as radiology images. ROIs are also described in terms of location within the image, or by the presence of a particular color. Table B-1 presents our classification of Image Markers (or pointers) and examples of Image Markers and Image Marker Referents.

Table B-1.Image Markers divided into four categories, followed by a sample caption in which Image Markers are marked in bold, Image Marker Referents are italicized.

Image Markers	Examples of Caption Text
Object Location	
front, top, bottom, left, right background, etc	Photograph of (**top**) *a polyurethane-covered nitinol stent*, (**middle**) a sheath with *inflated balloon catheter* for guiding, and (**bottom**) a *pusher catheter*
Object Color	
a distinct color that identifies a ROI.	Anterior SSD image shows an elongated *splenorenalvarix* (**blue area**). The varix travels from the splenichilar region inferiorly along the left flank, down into the pelvis, and eventually back up to the left renal vein via the left gonadal vein. The *kidney* is encoded **yellow**, *the portal system* is encoded **magenta**, and *the spleen* is encoded **tan**.
Overlay Marker	
arrows, asterisks, bounding boxes, circles, etc.	Transverse sonograms obtained with a 7.5-MHz linear transducer in the subareolar region. The **straight arrows** show *a dilated tubular structure*. The **curved arrow** indicates *an intraluminal solid mass*.
Overlay Label	
numbers, letters, abbreviations, words, etc.	Location of the calf veins. Transverse US image just above ankle demonstrates the paired *posterior tibial veins* (**V**) and *posterior tibial artery* (**A**) imaged from a posteromedial approach.

We locate the salient image region characteristics in the captions. We break down the task into two related subtasks - 1) locating and classifying textual clues for visually salient ROI features (Image Markers), and 2) locating the corresponding ROI text mentions (Image Marker Referents).

Rule-Based Approach to ROI extraction

First, we developed a two-stage rule-based, bootstrapping algorithm for locating the image markers and their referents from un-annotated data. The algorithm is based on the observation that textual image markers commonly appear in parentheses and are usually closely related semantic concepts. Thus the seed for the algorithm consists of:

1. The predominating syntactic pattern - parentheses, as in 'hooking of the soft palate (arrow)'. This pattern could easily be captured by a regular expression and does not require sentence parsing.
2. A dozen seed phrases (for example, 'left', 'circle', 'asterisk', 'blue') were identified by initially annotating a small subset of the data (20 captions). Wordnet [57] was used to look up and prepare a list of their corresponding inherited hypernyms. This hypernym list contains concepts such as 'a spatially limited location', 'a two-dimensional shape', 'a written or printed symbol', 'a visual attribute of things that results from the light they emit or transmit or reflect'. Best results were achieved when inherited hypernyms up to the third parent were used. In the first stage of the algorithm, all image captions were searched for parenthesized expressions that share the seed hypernyms. This step of the algorithm resulted in high precision, but low recall since image markers do not necessarily appear in parentheses. To increase recall, in stage 2 a full text search was performed for the stemmed versions of the expressions identified in stage 1.

This method achieves precision of 88% and recall of 70%. A baseline measure was also computed for the identification of the Image Marker Referents using a simple heuristic - the referent of the Image Marker is usually the closest Noun Phrase (NP). In the case of parenthesized image markers, it is the closest NP to the left of the Image Marker; in the case of non-parenthesized image markers, the referent is usually the complement of the verb; and in the case of passive voice, the NP preceding the verb phrase. The Stanford parser was used to parse the sentences. The accuracy of this method is 59%.

Supervised Machine Learning Approach to ROI extraction

We explored the possibility of improving the rule-based method results by applying a machine learning technique on the set of annotated data. Support Vector Machines (SVM) [58] was the approach taken because it is a state-of-the-art classification approach proven to perform well on many NLP tasks.

In our approach, each sentence was tokenized, and tokens were classified as Beginning, Inside, or Outside an Image Marker type or Image Marker Referent. Creating a classifier for relating Image Marker Referents to Image Markers is planned as future work. SVM classifiers were trained for each of these categories, and combined via 'one-vs-all' classification (the category of the classifier with the largest output was selected). The following features were used: ***token type*** (Word, Number, Symbol, Punctuation, White space); ***orthographic category*** (Upper initial, All capitals, Lower case, Mixed case); ***stem*** (extracted using the Porter stemmer); ***Wordnet superclass***; ***Wordnet hypernyms***; ***POS category*** (extracted using Brill's tagger); ***dependency path*** (the smallest sentence parse sub-tree including both the current token and the annotated image marker(s), encoded as an undirected path across POS categories.) The classifiers achieved 93.64% precision and 87.69% recall in marker identification and 61.15% accuracy in Image Marker Referent extraction.

B.3. Image representation (conceptual indexing)

To provide a structured summary of the salient image content akin to that of the MeSH indexing of the biomedical articles, we explored MetaMap-based extraction [59] of the salient indexing

terms from the image-related text as described in Section 6. Our studies show that image captions provide up to 80% of indexing terms (as judged by physicians trained in medical informatics), but additional filtering is needed for acceptable precision. While researching alternative filtering methods, we filter the extracted terms by semantic types in the groups corresponding to the elements of a well-formed clinical question (PICO) [60].

B.4. Generating structured documents (enriched citations) for retrieval

Working with a collection of structured data in XML format provides ready access to document fields, such as title, abstract, conditions, treatment, keywords, etc. Access to document structure supports differential weighting of the occurrences of search terms in different document fields. Structured documents also support faceted search and document and query frame unification (if queries are represented using the same structure). Although the Essie search engine presently supports only the differential weighting of the search terms, creating the structured documents is worthwhile, as we can adjust the field weights for different tasks. The basis of our structured document is a MEDLINE citation (usually retrieved using E-Utilities), if available. Otherwise, we create pseudo-citations using the first 250 words of a given text as abstract and obtaining automatic MeSH indexing using NLM's MTI [61]. We then enrich the citation with the above extracted free-text and entities. **Figure B-6** depicts the enriched MEDLINE citation PMID 18487544. This short paper contains no abstract and only one image. The additional information extracted from the paper consists of the citation, modality, PICO elements, and Regions of Interest.

B.5. Automatic query generation

It is possible to submit short user queries "as is" to most search engines, and we use this approach in our online image search engine. In a longer text, such as a patient's case description, this approach could be too restrictive if the presence of all terms is required (the terms are ANDed) or too noisy when all terms are ORed. We therefore explored several approaches to recognizing most salient terms and to grouping and nesting search terms.

Our term extraction and grouping approaches are based on the four components of a well-formed clinical question: **P**atient/**P**roblem, **I**ntervention, **C**omparison, and **O**utcome (PICO). To construct the PICO frames, we use Essie, MetaMap or our Clinical Term eXtractor (CTX) for clinical narrative to map the text to the UMLS Metathesaurus and extract concepts relating to problems, interventions (drugs, therapeutic and diagnostic procedures), and anatomy. We also extract age and gender using regular expressions. The problem, age, gender, and anatomy terms contribute to the Patient/Problem element of the PICO frame and the intervention terms contribute to the Intervention and Comparison elements. An additional extractor identified terms related to image modality. We use our implementation of the NegEx algorithm [62] to identify problems positively present in a patient.

```
<?xml version="1.0" encoding="utf-8" ?>
- <document>
    <meta iclef_id="239029" />
    <meta publisher="Radiology" />
    <meta journal_title="The puff of smoke sign" />
    <meta fulltext_html_url="http://radiology.rsnajnls.org/cgi/content/full/247/3/910" />
    <meta iti_id="18487544F1" />
    <meta volume="247" />
    <meta authors="Ortiz-Neira, Clara L;" />
    <meta pmid="18487544" />
    <meta issue="3" />
    <title>The puff of smoke sign</title>
    <abstract />
  - <image type="figure" id="1" src="./images/239029.jpg"
      link="http://radiology.rsnajnls.org/cgi/content/full/247/3/910/F1">
      <caption>Anteroposterior angiogram of right internal carotid artery shows abnormal hypertrophy of
        perforating arteries, which produces the puff of smoke sign (arrow) and is associated with
        narrowing (arrowheads) of the M1 and A1 segments of the distal internal carotid artery.</caption>
      <mention />
    - <pico>
        <modalityclass>xr</modalityclass>
        <modality>angiogram</modality>
        <intervention cui="C0002978" negstatus="NOT_NEGATED">angiogram</intervention>
        <anatomy cui="C0226156" negstatus="NOT_NEGATED">right internal carotid artery</anatomy>
        <problem cui="C0020564" negstatus="NOT_NEGATED">hypertrophies</problem>
        <anatomy cui="C1182750" negstatus="NOT_NEGATED">perforating arteries</anatomy>
        <anatomy cui="C0007276" negstatus="NOT_NEGATED">internal carotid artery</anatomy>
      </pico>
    - <rois>
        <roi type="arrow">smoke sign</roi>
        <roi type="arrow">narrowing</roi>
      </rois>
    </image>
  + <mesh>
  </document>
```

Figure B-6: Enriched MEDLINE citation

Figure B-7 shows a patient's case description [63] and PICO elements represented as an Extensible Markup Language (XML) document.

Our current best query generation strategy [64] (which is iterative and might take up to five iterations) can be used in asynchronous decision support, but is not yet practical for an online search engine. Therefore we are exploring two simplified approaches: type-based and concept-based. In the concept-based approach, all concepts are ORed. In the type-based approach, the concepts are grouped by type (problem, intervention, etc.), ORed within the type, and the type groups are ANDed. Until usefulness of negated terms is explored, we use only non-negated and possibly negated terms to generate search queries.

Automatic query generation is necessary for providing real-time clinical decision support and we plan to continue focusing on this problem in our next steps.

```
<case id="5802JFP">
<description>
A 70-year-old man with painful bilateral leg swelling that had gotten progressively worse over the past
week. He had no significant past medical or surgical history, took an aspirin daily, did not smoke tobacco
or drink alcohol, and had not taken any trips recently. He denied any chest pain, dyspnea, or orthopnea,
but indicated that he'd been having difficulty swallowing food for the past month. The patient had a
cachectic appearance, diminished breath sounds and dullness to percussion over the right middle and
lower lung fields, and pitting edema up to the knees bilaterally.
</description>
<sentence number="1">
<problem cui="C0030193"negstatus="NOT_NEGATED"> painful </ problem>
<problem cui="C0581394" negstatus ="NOT_NEGATED"> leg swelling</problem>
<age>70-year-old , Aged;</age>
</sentence>
<sentence number="2">
<drug cui="C0002185" negstatus ="NOT_NEGATED">aspirin</drug>
</sentence>
<sentence number="3">
<problem cui="C0008031" negstatus ="DEFINITELY_NEGATED">chest pain</problem>
<problem cui="C0013404" negstatus ="DEFINITELY_NEGATED">dyspnea</problem>
<problem cui="C0085619" negstatus ="DEFINITELY_NEGATED">orthopnea</problem>
<problem cui="C0011168" negstatus =" NOT_NEGATED">difficulty swallowing</problem>
</sentence>
<sentence number="4">
<problem cui="C0006625" negstatus =" NOT_NEGATED">cachectic</problem>
<intervention cui="C0030987" negstatus =" NOT_NEGATED">percussion </intervention>
<anatomy cui="C0934576" negstatus =" NOT_NEGATED">lower lung field</anatomy>
<problem cui="C0333243" negstatus =" NOT_NEGATED">pitting edema</problem>
<anatomy cui="C0022742" negstatus =" NOT_NEGATED">knees</anatomy>
</sentence>
</case>
```

Figure B-7: Example structured representation of a case

B.6 Essie

The Essie search engine [65], developed and used at NLM, features a number of strategies aimed at alleviating the need for sophisticated user queries. These strategies include a fine-grained tokenization algorithm that preserves punctuation, and phrase searching based on the user's query. Essie is particularly well-suited for information retrieval tasks in the medical domain since it performs concept-based indexing, automatically expands query terms using synonymy relationships in the UMLS Metathesaurus, and weights term occurrences according to their document location when computing document scores. Essie's algorithm for scoring the similarity between a document and a query can be summed up as preferring "all the right pieces in all the right places". The "right pieces" are phrases from the query, and the "right places" are the fields of a document most valuable for a retrieval task, such as image captions for image retrieval, or MeSH for literature retrieval [66]. Essie allows re-ranking search results favoring different clinical tasks, provides information about terms related to the query, and displays extracted terms in context.

Appendix C. Image Text Search Engine (ITSE)

Image Text Search Engine (ITSE) is the CEB experimental search engine for retrieval of biomedical literature and images, as well as linking evidence to patients' cases. Building upon existing tools and knowledge, ITSE combines Essie (described above) with the CEB image retrieval engine (described in Sections A.2 and A.3, respectively), and implements user interface principles developed by Hearst et al. [11]. Along with the traditional elements of search results display, such as titles and author names, ITSE provides captions of the retrieved images and short summaries of the retrieved abstracts. The summaries, which are patient-oriented outcomes extracted from abstracts, are obtained through the RIDeM[1] [67] services developed independently in another CEB project.

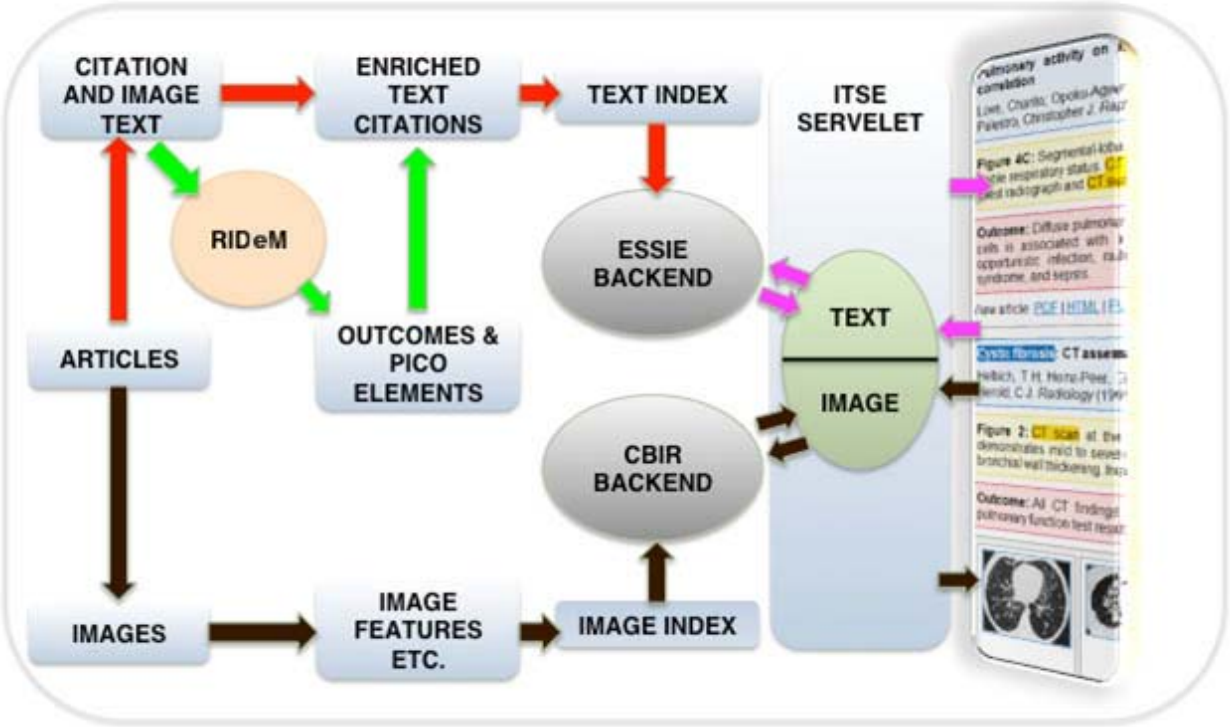


Figure C-1. ITSE search engine pipeline showing the flow of indexing, and search steps.

The ITSE search engine pipeline is shown in Figure C-1. In the diagram, the blocks to the left of the servlet are steps required for indexing text and image data. Here, red arrows indicate text indexing steps, black arrows indicate image indexing data flow, and green arrows indicate information retrieved from resources external to ITSE (e.g., RIDeM and MEDLINE using NCBI E-Utilities). Arrows in magenta indicate text query and retrieved data for the user interface while light blue arrows show the image query and retrieved data.

The ITSE user interface provides the Essie search options and displays search results in a list or grid view. Figure C-2 shows the search options.

[1] http://archive.nlm.nih.gov/ridem

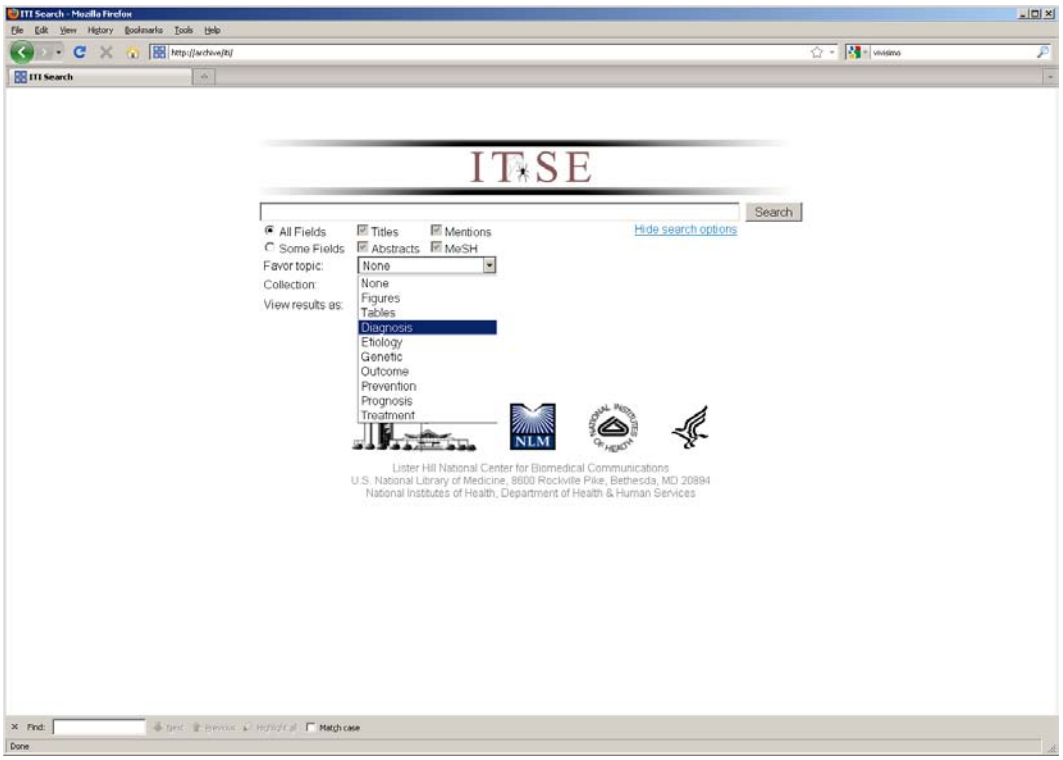


Figure C-3. ITSE search options. (Users can request ranking the search results with respect to their usefulness to a clinical task, for example, treatment.)

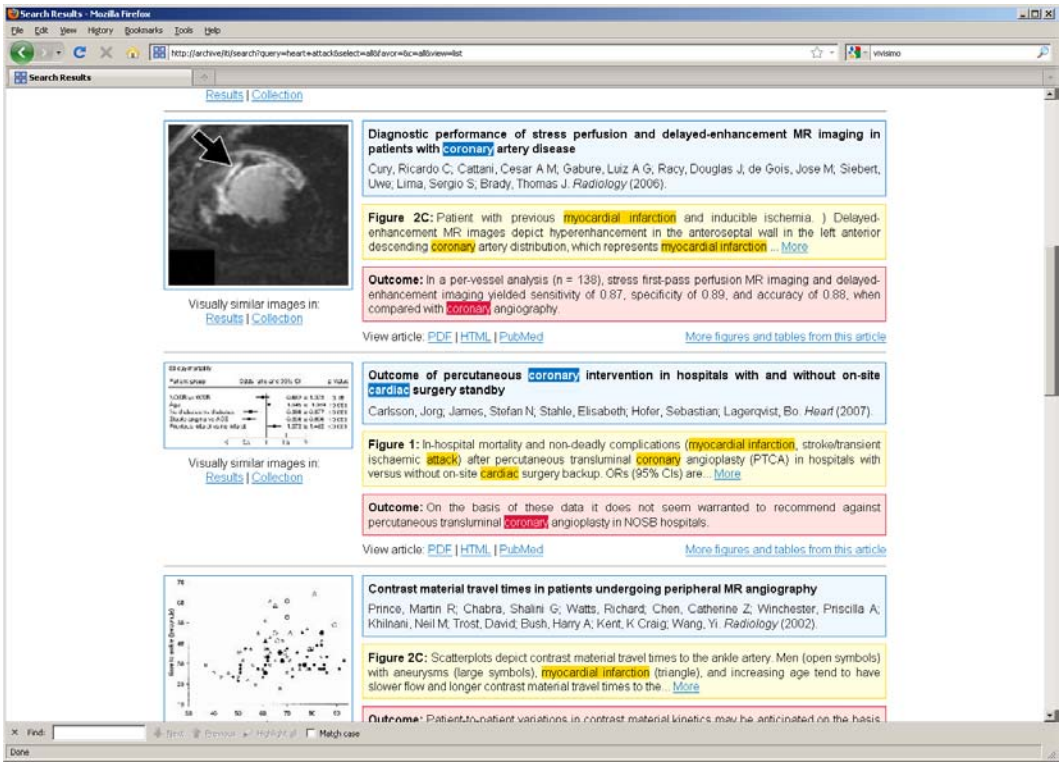


Figure C-4. Image search results in a list ("Enriched citation").

Figure C-3 shows the image search results as an "enriched citation" for the query "heart attack". The search terms and their synonyms (such as "myocardial infarction", "cardiac", and "coronary") are highlighted using Essie's hit highlighting tools. The title of the paper and other bibliographic information is displayed in the blue box, the first three lines of the image caption are displayed in the yellow box, and the bottom-line advice in the form of the patient-oriented outcome extracted from the abstract is presented in the pink box.

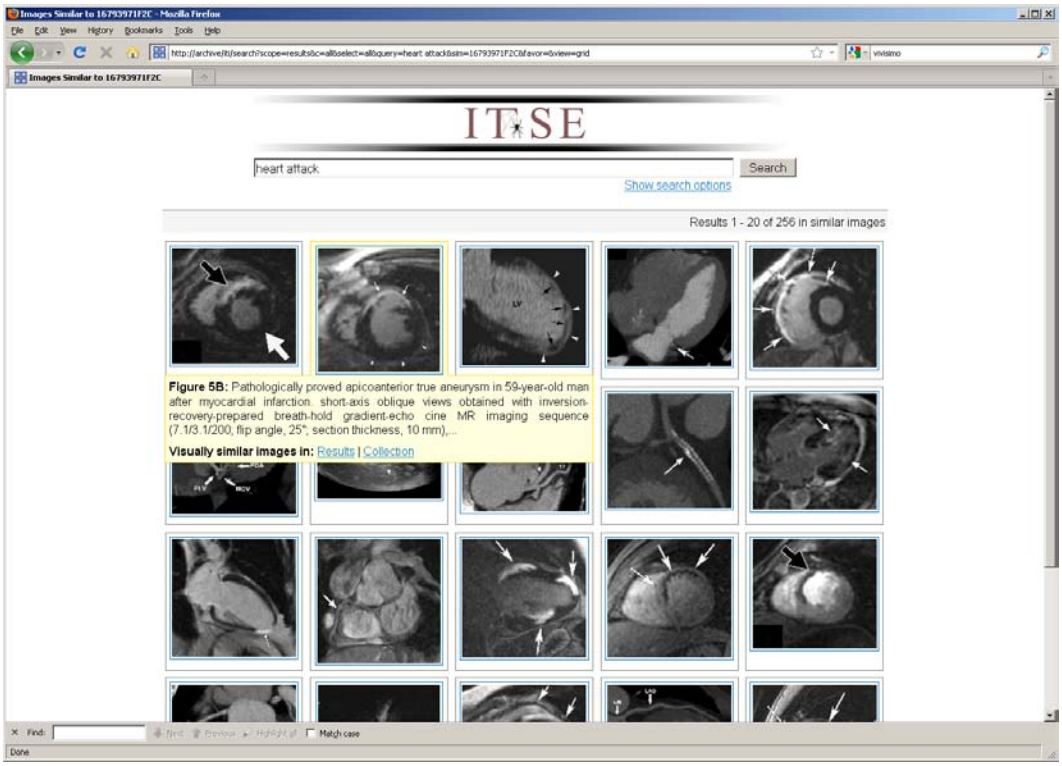


Figure C-5. Grid view of the visually similar images found within search results.

Should the user want to see other images similar to the ones retrieved in the search, he/she may re-use any of these as input to the CEB image retrieval engine. The CEB image retrieval engine uses low-level visual features, such as color, texture, and shape as the primary building blocks of the visual content in an image. The features are then transformed into *visual keywords* that represent images with a set of labels that indicate the membership of local image regions in various image categories. Similarity between a query image and database images is measured as a weighted linear combination of different features. The image retrieval engine searches for similar images within the retrieval results or in the whole collection. As shown in Figure C-4, the visually similar images within the results are displayed in a grid view that displays textual representation of the image when the mouse is placed over the image.

Appendix D. Conceptual image indexing: Methods & Evaluation

D.1. *Evaluating automatic extraction of terms for image representation*

The experiments were designed to test our questions about coverage of the conceptual image indexing terms in the image-related text and in the indexing terms assigned by NLM indexers to the papers containing those images, and our ability to extract the terms.

Overall, the evaluators rated 451 extracted terms as useful for indexing and submitted 255 additional indexing terms.Table D-1 presents the average numbers of concepts per image evaluated and found useful for indexing by each evaluator.

**Table D-1: Average number of concepts per image. ***Evaluators trained in medical informatics are marked with an asterisk.

Specialty	Indexing Terms		
	evaluated	useful	%useful
family physician*	19.26	2.38	12.4%
cardiologist*	17.80	2.02	11.4%
plastic surgeon*	17.89	1.80	10.1%
internist*	17.55	2.18	12.4%
general surgeon	19.98	1.50	7.5%
medical imaging	14.46	1.40	9.9%
Mean ± CI	17.83±2.0	1.89±0.4	10.6±2.0%

D.1.1. *Coverage of the conceptual image indexing terms in the MeSH terms assigned by NLM indexers to the papers*

Table D-2 presents the percentages of automatically extracted terms judged useful (the extracted column) and additional terms assigned by the evaluators (the added column) that match MeSH terms. The **%used** column of the table shows the overall proportion of the MeSH terms assigned to the paper that were deemed useful in annotating images.

**Table D-2: Match between indexing terms assigned to images and papers. ***Evaluators trained in medical informatics are marked with an asterisk.

Specialty	MeSH Terms		
	extracted	added	%used
family physician*	33.0%	34.9%	11.5%
cardiologist*	39.8%	48.7%	20.5%
plastic surgeon*	46.9%	41.2%	11.1%
internist*	25.0%	25.7%	11.7%
general surgeon	33.3%	––	7.1%
medical imaging	28.8%	––	5.3%
Mean ± CI (%)	34.5±8.2	25.1±21.9	11.2±5.5

D.1.2. *Coverage of the conceptual image indexing terms in the image-related text*

For three of the 255 indexing terms added by the evaluators, no image-related text was extracted. Of the remaining 252 added terms, 75 were extracted verbatim from the caption text and 11 from the discussion (mention) of the image in the text. Another 139 added terms were generated using captions and mentions through:

- Coordinating constructions, for example, extracting *Preoperative photograph* from

Preoperative and postoperative photographs;
- Paraphrasing, for example, deriving *elderly* from *89-year old*;
- Summarizing, for example, the following mention of the image: *a mobile, left-sided, nasal dorsal implant with tip ptosis, erythema, and swelling of the left nasal vestibule* as *implantation complications*;
- Generalizing based on the figure and the caption, for example, *ultrasound*; *surgical method*; or *transthoracic echocardiography*.

The remaining 27 terms were found in the paper title, abstract, and MeSH terms assigned to the paper. Of the 255 terms added by the evaluators, 103 were subsequently mapped to UMLS concepts.

D.1. 3. *Extraction accuracy*

The design of the extraction evaluation was recall oriented. All extracted terms were given to the evaluators without any filtering to have enough training examples for learning term selection in the future.

The extraction method was evaluated using recall and precision computed for each evaluator as follows: The desired index terms D for the images are the set of extracted terms evaluated as useful for indexing combined with the indexing terms added by the evaluator, A is the set of all suggested indexing terms, and within A there is a set of terms evaluated as useful for indexing C. Precision P and recall R are:

$P = |C|/|A|$

$R = |C|/|D|$

Precision and recall were computed for each evaluator, and then averaged.

Recall and precision achieved by this baseline extraction method are shown inTable D-3.

**Table D-3: Evaluation of the baseline extraction method. ***Evaluators trained in medical informatics are marked with an asterisk.

Specialty	Recall	Precision	F-score
family physician*	0.723	0.124	0.211
cardiologist*	0.447	0.114	0.181
plastic surgeon*	0.827	0.101	0.179
internist*	0.565	0.124	0.204
general surgeon	0.333	0.075	0.122
medical imaging	0.917	0.099	0.179
Average	0.635	0.106	0.182

On average, only 64% of the desirable indexing terms could be found using the existing extraction methods and ontologies. More sophisticated mapping algorithms are needed to extract another 15% of the terms, and more complex natural language processing and ontology expansion are needed to identify the remaining terms.

D.2. *Selecting image representation terms for semantic image retrieval*

To have a broad filter applicable to images in all medical specialties, we need to have training examples in all specialties, or find effective unsupervised learning methods, or develop specialty-independent techniques. We decided to explore the specialty-independent techniques by using non-lexical features to represent the conceptual indexing terms extracted by MetaMap. We use the extraction evaluation results (Table D-3) as the baseline. We use the ten-fold cross-validation results on the above set of evaluated terms as the upper bound for specialty-independent filtering, and compare the results to those obtained on an additional set of 1539 potential indexing terms

relating to 50 randomly chosen images from 31 different articles in the 2006 Archives of Dermatology journal. Table D-4shows that classifiers trained on specific examples for given specialties improve precision three-fold with 10% loss in recall and that the same classifiers improve term selection for other domains.

Table D-4: The results of image representation selection based on supervised machine learning

Annotation method	Recall	Precision	F-score
Baseline	0.635	0.106	0.182
Training	0.502	0.332	0.400
Standard	0.492	0.231	0.314

We defined the following features used to classify potential indexing terms:

1. CUI (nominal): The Concept Unique Identifier assigned to the concept in the UMLS Metathesaurus.
2. Semantic Type (nominal): The concept's UMLS semantic type.
3. Presence in Caption (nominal): true if the phrase that generated the concept is located in the image caption
4. MeSH Ratio (real): The ratio of words ci in the concept c that are also contained in theMedical Subject Headings,M, assigned to the document to the total number of words in the concept.

$$R^{(m)} = |\{ ci: ci \in M\}|/|c|$$

5. Abstract Ratio (real): The ratio of words ci in the concept c that are also in the document's abstract, A, to the total number of words in the concept.

$$R^{(a)} = |\{ ci: ci \in A\}|/|c|$$

6. Title Ratio (real): The ratio of words ci in the concept c that are also in the document's title T to the total number of words in the concept.

$$R^{(t)} = |\{ ci: ci \in T\}|/|c|$$

7. Parts of Speech Ratio (real): The ratio of words pi in the phrase p that have been tagged as having part of speech s to the total number of words in the phrase.

$$R^{(s)} = |\{pi : TAG(pi) = s\}|/|p|$$

   This feature is computed for noun, verb, adjective and adverb part-of-speech tags. We obtain POS information from the output of MetaMap.

8. Concept Ambiguity (real): The ratio of the number mappings mi of phrase p that contain concept c to the total number of mappings for the phrase:

$$A=|\{m_{p-i}:c \in m_{p-i}\}|/|m^p|$$

9. tf-idf (real): The frequency of term t_i (i.e., the phrase that generated the concept) times its inverse document frequency
10. Document Location (real): The location in the document of the phrase that generated the concept. This feature is continuous on [0; 1] with 0 representing the beginning of the document and 1 representing the end.
11. Concept Length (real): The length of the concept, measured in number of characters.

12. For the purpose of computing features 9 and 10, we indexed each collection with the Terrier information retrieval platform. Terrier was configured to use a block-indexing scheme with a tf-idf weighting model.

We explored the feature vectors constructed as described above using various classification approaches available in the RapidMiner tool. Unlike many related text and image classification problems, we were unable to achieve results with a Support Vector Machine (SVM) learner (libSVMLearner) using the Radial Base Function (RBF). Common cost and width parameters were used, yet the SVM classified all terms as ineffectual. Identical results were observed using a Naive Bayes (NB) learner. For these reasons, we chose to use the Averaged One-Dependence Estimator (AODE) learner (Webb et al., 2005) available in RapidMiner. AODE is capable of achieving highly accurate classification results with the quick training time usually associated with NB. Because this learner does not handle continuous attributes, we preprocessed our attributes with an equal frequency discretization. The AODE learner was trained in a ten-fold cross validation of our training data.

Table D-5: Feature Comparison. The information gain and chi-square statistic is shown for each feature. A higher value indicates greater influence on term effectiveness.

Feature	Information gain	chi-square
CUI	0.003	13.331
Semantic Type	0.015	68.232
Presence in Caption	0.008	35.303
MeSH Ratio	0.043	285.701
Abstract Ratio	0.023	114.373
Title Ratio	0.021	132.651
POS:		
noun	0.053	287.494
verb	0.009	26.723
adjective	0.021	96.572
adverb	0.002	5.271
Concept Ambiguity	0.008	33.824
tf-idf	0.004	21.489
Document Location	0.002	12.245
Phrase Length	0.021	102.759

The effectiveness of individual features in describing the potential specialty-independent indexing terms is shown in Table D-5. We used two measures, both of which indicate a similar trend, to calculate feature effectiveness: Information gain (Kullback-Leibler divergence) and the chi-square statistic. Under both measures, the MeSH ratio is the most effective feature. The abstract and title ratios also had a significant effect on the classification outcome. Similar to MeSH terms, these constructs are a coarse summary of the contents of an article; therefore it is not unreasonable to assume they summarize the images contained therein. Finally, the length of the UMLS concept and the nouns ratio were moderately effective. Interestingly, though, tf-idf and document location, both features computed using standard IR techniques, are among the least effective features.

References

1. Winfield H, Lain E, Horn T, Hoskyn J. Eosinophilic cellulitislike reaction to subcutaneous etanercept injection. Arch Dermatol. 2006 Feb;142(2):218-20
2. Sandusky RJ, Tenopir C. Finding and Using Journal Article Components: Impacts of Disaggregation on Teaching and Research Practice. Journal of the American Society for Information Science & Technology, 59 (6) April 2008: 970-82.
3. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281–91.
4. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(12):1349-80, Dec 2000.
5. Simpson M, Demner-Fushman D, Sneiderman C, Antani SK Thoma GR. Using Non-lexical Features to Identify Effective Indexing Terms for Biomedical Illustrations Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece, April 2009
6. Sneiderman C, Demner-Fushman D, Fung KW, Bray B. UMLS-based Automatic Image Indexing.Proceedings of the 2008 Annual Symposium of the American Medical Information Association (AMIA 2008), Washington, DC, November 2008
7. Demner-Fushman D, Antani S, Simpson M, Thoma GR. Combining Medical Domain Ontological Knowledge and Low-level Image Features for Multimedia Indexing LRE C 2008 (Sixth International Conference on Language Resources and Evaluation), OntoImage Workshop, Marrakech, Morocco, May 2008
8. Long R, Antani S, Deserno TM, Thoma GR. Content-Based Image Retrieval In Medicine: Retrospective Assessment, State of the Art, and Future Directions. International Journal of Healthcare Information Systems and Informatics. January 2009;4(1):1-16.
9. Yao J, Zhang Z, Antani S, Long R, Thoma G. Automatic Medical Image Annotation and Retrieval. J Neurocomputing. June 2008;71(10-12):2012-22.
10. Ghosh P, Antani S, Thoma GR. A Survey of Content-Based Image Retrieval Systems. CEB Internal Technical Report. August 2010. http://archive.nlm.nih.gov
11. Hearst MA, Divoli A, Buturu H, Ksikes A, Nakov P, Wooldridge MA, et al. Biotext search engine: beyond abstract search. Bioinformatics 2007;23(16):2196–7. doi: 10.1093/bioinformatics/btm301.
12. Hong Yu, Yong-gang Cao. (2008). Automatically Extracting Information Needs from Ad Hoc Clinical Questions. Proceedings of AMIA Symposium, 96-100.
13. Xu, S., McCusker, J., Krauthammer. M. (2008). Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics*, 2008
14. Antani SK, Deserno TM, Long R, Thoma GR. Geographically Distributed Complementary Content-Based Image Retrieval Systems For Biomedical Image Informatics. Proc. MEDINFO 2007. 12(1):493-7.
15. Hsu W, Antani SK, Long R. SPIRS: a Framework For Content-based Image Retrieval From Large Biomedical Databases. Proc. of MEDINFO. 12(1):188-92.
16. Hersh W. Voorhees E. 2009. TREC genomics special issue overview. Inf. Retr. 12, 1 (Feb. 2009), 1-15.
17. Regev, Y., Finkelstein-Landau, M., Feldman, R., Gorodetsky, M., Zheng, X., Levy, S., Charlab, R., Lawrence, C., Lippert, R. A., Zhang, Q., and Shatkay, H. 2002. Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1). SIGKDD Explor. Newsl. 4, 2 (Dec. 2002), 90-92.
18. Shatkay H, Chen N, Blostein D. Integrating image data into biomedical text categorization. Bioinformatics. 2006 Jul 15;22(14):e446-53.
19. Divoli A, Wooldridge MA, Hearst MA (2010) Full Text and Figure Display Improves Bioscience Literature Search. PLoS ONE 5(4): e9619.)
20. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, United Kingdom: ACM; 2004, p. 25–32.
21. Smucker MD, Allan J, Carterette B. A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal: ACM; 2007, p. 623–32.

22. Woods JW, Sneiderman CA, Hameed K, Ackerman MJ, Hatton C. Using UMLS metathesaurus concepts to describe medical images: dermatology vocabulary. Comput Biol Med. 2006 Jan;36(1):89-100.
23. Deserno TM, Antani S, Long R. Ontology of Gaps In Content-Based Image Retrieval. Journal of Digital Imaging. April 2009;22(2):202-15.
24. H. Müller, J. Kalpathy-Cramer, C.E. Kahn Jr., W. Hatt, S. Bedrick, W. Hersh, Overview of the ImageCLEFmed 2008 medical image retrieval task. Available from: http://www.clef-campaign.org/2008/working%5Fnotes/
25. Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. J Am Med Inform Assoc. 1994 May-Jun;1(3):249-62.
26. Kahn CE Jr, Rubin DL. Automated semantic indexing of figure captions to improve radiology image retrieval. J Am Med Inform Assoc. 2009 May-Jun;16(3):380-6.
27. Kammerer FJ, Frankewitsch T, Prokosch HU. Design of a web portal for interdisciplinary image retrieval from multiple online image resources. Methods Inf Med. 2009;48(4):361-70.
28. Long LR, Antani SK, Thoma GR. Image informatics at a national research center. Computerized Medical Imaging and Graphics . April 2005;29(3):171-93 .
29. Xu X, Lee DJ, Antani SK, Long LR. A Spine X-Ray Image Retrieval System using Partial Shape Matching. IEEE Transactions on Information Technology in Biomedicine. January 2008;12(1):100-8.
30. Hsu W, Antani S, Long LR, Neve L, Thoma GR. SPIRS: a Web-based Image Retrieval System For Large Biomedical Databases. Int. J. Medical Informatics. 2008. 78 Suppl 1:S13-24.
31. Xue Z, Long LR, Antani S, Jeronimo J, Thoma GR. Proc. SPIE Medical Imaging 2008. vol. 6919. San Diego, CA: Feb, 2008. A Web-accessible content-based cervicographic image retrieval system; p. 691907-1-9
32. Bueno JM, Chino FJT, Traina AJM, Traina Jr. C, Azevedo-Marques PM. How to Add Content-based Image Retrieval Capability in a PACS. Proc. IEEE International Symposium on Computer-Based Medical Systems (CBMS) 2002; 321-26
33. Deserno TM, Antani S, Long LR. Content-based Image Retrieval For Scientific Literature Access. Methods of Information in Medicine. April 2009;48(4):371-80.
34. Thierry Declerck and Manuel Alcantara. 2006. Semantic analysis of text regions surrounding images in Web documents. In OntoImage 2006 Workshop on Language Resources for Content-based Image Retrieval, pages 9–12.
35. Simpson M, Rahman MM, Demner-Fushman D, Antani S, Thoma GR. Text- and Content-based Approaches to Image Retrieval for the ImageCLEF2009 Medical Retrieval Track. CLEF2009 Working Notes. CLEF 2009 Workshop 30 September - 2 October, Corfu, Greece, in conjunction with ECDL2009.
36. Demner-Fushman D, Antani SK, Thoma GR. Automatically Finding Images for Clinical Decision Support. Proceedings of IEEE International Workshop on Data Mining in Medicine (DM-Med 2007). Omaha, NE, October 2007, pp. 139-144.
37. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res. 1999; 10: 271-89.
38. Rafkind B, Lee M, Chang SF, Yu H. Exploring text and image features to classify images in bioscience literature. Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology. 2006 Jun:73-80.
39. Amati G, Van Rijsbergen, C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 2002 Oct;20(4):357-389.
40. Antani S, Demner-Fushman D, Li J, Srinivasan BV, Thoma GR. Exploring use of images in clinical articles for decision support in Evidence-Based Medicine. Proc. SPIE-IS&T Electronic Imaging. San Jose, CA. January 2008;6815:68150Q(1-10)
41. Kennedy J, Eberhart R. Particle Swarm Optimization. Proceedings of the IEEE International Conference on Neural Networks, 1995, 1942-48.
42. You D, Apostolova E, Antani S, Demner-Fushman D, Thoma G. Figure content analysis for improved biomedical article retrieval. Proc. SPIE, Vol. 7247, 72470V (2009)
43. You D, Antani S, Demner-Fushman D, Rahman M, Govindaraju V, Thoma G. Biomedical Article Retrieval Using Multimodal Features and Image Annotations In Region-based CBIR. Document Recognition and Retrieval XVII. Edited by Likforman-Sulem, Laurence; Agam, Gady. Proc. SPIE. 2010;7534:75340V-75340V-12.
44. Cheng B, Stanley JR, Antani S, Thoma GR. A Novel Computational Intelligence-based Approach for Medical Image Artifacts Detection. Proceedings of the 2010 International Conference on Artificial Intelligence and Pattern Recognition, 2010:113-20, ISBN: 978-1-60651-015-5
45. Deserno TM, Antani S, Long R. Ontology of Gaps In Content-Based Image Retrieval. Journal of Digital Imaging. April 2009; 22(2):202-15.

46. Chatzichristos SA, Boutalis YS. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. Proc. 6th International Conference on Computer Vision Systems, Lecture Notes in Computer Science, 5008:312-22, 2008.
47. S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, 11(6):688-95, 2001.
48. G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. Proc. 4th ACM International Conference on Multimedia, pages 65-73, 1996.
49. Demner-Fushman D, Antani S, Simpson M, Thoma GR. Annotation and Retrieval of Clinically Relevant Images. International Journal of Medical Informatics: Special Issue on Mining of Clinical and Biomedical Text and Data. December 2009;78(12):e59-e67.
50. Demner-Fushman D, Antani SK, Thoma GR. Automatically Finding Images for Clinical Decision Support.Proc . IEEE International Workshop on Data Mining in Medicine (DM-Med 2007). 2007:139-44
51. Oliva A, Torralba A. Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research. Visual Perception - Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception. 155, Part 2, 2006:23-36
52. Lux M, Chatzichristos SA. LIRe: Lucene image retrival: an extensible java CBIR library. Proc. 16th ACM International Conference on Multimedia, 2008,1085-88.
53. Chatzichristos SA, Boutalis YS. FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval. Proc. 9th International Workshop on Image Analysis for Multimedia Interactive Services, 2008, 191-96.
54. Rahman MM, Antani SK, Thoma GR. A Medical Image Retrieval Framework In Correlation Enhanced Visual Concept Feature Space. 22nd IEEE International Symposium on Computer-Based Medical Symposium (CBMS). August 2009.
55. Rahman MM, Antani SK, Long LR, Demner-Fushman D, Thoma GR. Multi-Modal Query Expansion Based On Local Analysis For Medical Image Retrieval. Lecture Notes in Computer Science. First MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support (MCBR-CDS 2009); part of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention February 2010;5853/2010(doi: 10.1007/978-3-642-11769-5):110-9.
56. Rahman MM, Antani S, and Thoma GR. A Classification-Driven Similarity Matching Framework For Retrieval of Biomedical Images. 11th ACM International Conference on Multimedia Information Retrieval (MIR 2010). 2010:147-154.
57. WordNet: An Electronic Lexical Database. Christiane Fellbaum (eds.). MIT Press, Cambridge, MA, (1998)
58. Vapnik VN. (2000) The Nature of Statistical Learning Theory. Springer.
59. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
60. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007;33(1):63-103.
61. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform. 2004;107(Pt 1):268-72.
62. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
63. Baird DC, Usatine RP. Photo Rounds. Bilateral leg edema and difficulty swallowing. J Fam Pract. 2009 Feb;58(2):89-92.
64. Demner-Fushman D, Karpinski J, Thoma GR. Automatically building a repository to support evidence based practice. Proceedings of the 2nd Workshop on Building and evaluating resources for biomedical text mining (BioTxtM 2010), 7th Language Resources and Evaluation Conference (LREC 2010). May 17-23, Valetta, Malta.
65. McCray AT, Ide NC, Loane RR, Tse T. Strategies for supporting consumer health information seeking. Stud Health Technol Inform. 2004;107(Pt 2):1152-6.
66. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.
67. Demner-Fushman D, Seckman C, Fisher C, Hauser SE, Clayton J, Thoma GR. A Prototype System to Support Evidence-based Practice. Proceedings of the 2008 Annual Symposium of the American Medical Information Association (AMIA 2008), Washington, DC, November 2008.