# Multimodal Medical Image Retrieval: OHSU at ImageCLEF 2008

Jayashree Kalpathy-Cramer, Steven Bedrick, William Hatt, William Hersh

Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA
{kalpathy,bedricks,hattb,hersh}@ohsu.edu

**Abstract**

We present results from Oregon Health & Science University's participation in the medical image retrieval task of ImageCLEF 2008. We created a web-based retrieval system built on a full-text index of the annotations using a Ruby on Rails framework. The text-based search engine was implemented in Ruby using Ferret, a port of Lucene. In addition to this textual index of annotations, supervised machine learning techniques using visual features were used to classify the images based on image acquisition modality. All images were annotated with the purported modality. Our system provides the user with a number of search options including those for limiting the search to the desired modality, UMLS-based term expansion and Natural Language Processing based techniques. Purely textual runs as well as mixed runs using the purported modality were submitted. We also submitted interactive runs using a number of user specified search options. Latent semantic analysis of the visual features was used to reorder results. The use of the UMLS Metathesaurus increased our recall. However, our system is primarily geared towards precision. Consequently, many of our multimodal automatic runs using the custom parser as well as interactive runs had high early precision. Our runs also performed well using the bpref metric, a measure that is more robust in the case of incomplete judgments.

## Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries;

## General Terms
Measurement, Performance, Experimentation

## Keywords
Image Retrieval, Performance Evaluation, Image Classification, Medical Imaging

## Medical Image Retrieval

Advances in digital imaging technologies and the increasing prevalence of Picture Archival and Communication Systems (PACS) have led to a substantial growth in the number of digital images stored in hospitals and medical systems in recent years. In addition, on-line atlases of images have been created for many medical domains including dermatology, radiology and gastroenterology. Medical images can form an essential component of a patient's health record. Medical image retrieval systems can be important with aiding in diagnosis and treatment. They can also be highly effective in health care education, for students, instructors and patients.

### 1 Introduction

Image retrieval systems do not currently perform as well as their text counterparts [1]. Medical and other image retrieval systems have historically relied on annotations or captions associated with the images for indexing the retrieval system. The last few decades have seen numerous advancements in the area of content-based image retrieval (CBIR) [2,3]. Although CBIR systems have demonstrated success in fairly constrained medical domains including pathology, dermatology, chest radiology, and mammography, they have

demonstrated poor performance when applied to databases with a wide spectrum of imaging modalities, anatomies and pathologies [1,4,5,6].

Retrieval performance has shown demonstrable improvement by fusing the results of textual and visual techniques. This has especially been shown to improve early precision [7,8]. The medical image retrieval task within ImageCLEF (ImageCLEFmed) 2008 campaign is TREC-style [9] and provides a forum and set of test collections for the medical image retrieval community to use to benchmark their algorithms on a set of queries. The ImageCLEF campaign has, since 2003, been a part of the Cross Language Evaluation Forum (CLEF) [9,10,11] which is derived from the Text Retrieval Conference (TREC, trec.nist.gov).

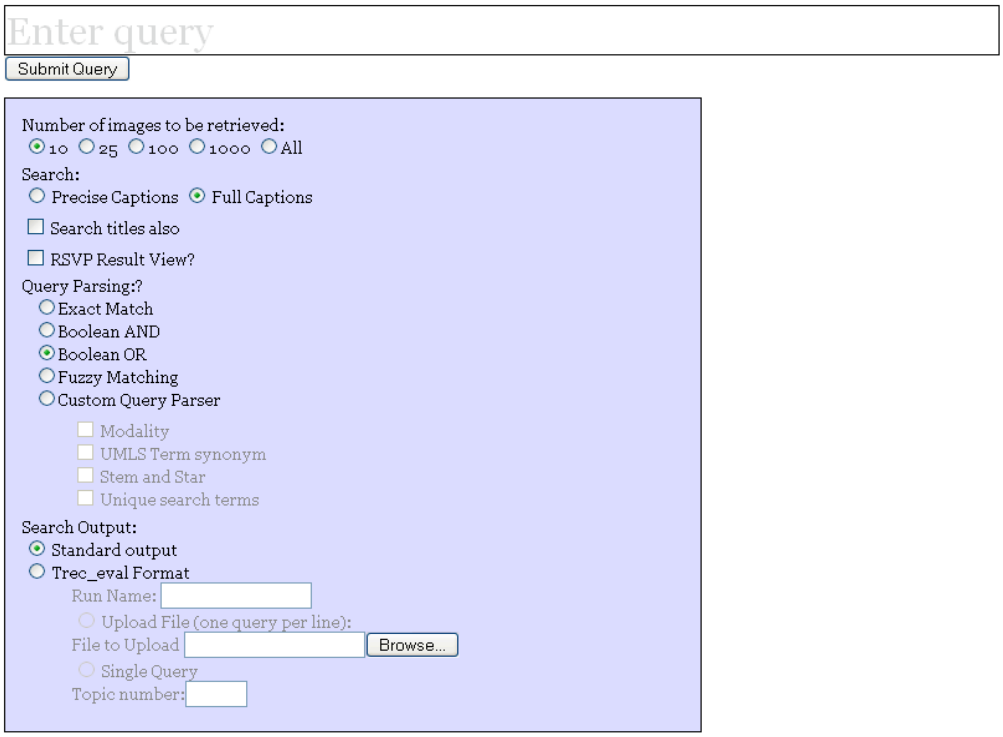
## 2. System Description of our Adaptive Medical Image Retrieval System

The ImageCLEF 2008 medical image retrieval test collection consists of about 66,000 medical images and annotations associated with them. This collection is a set of images and captions from *Radiology* and *Radiographics* , two RSNA journals. We have created a flexible database schema that allows us to easily incorporate new collections while facilitating retrieval using both text and visual techniques. The captions and titles in the collection are currently indexed and we continue to add indexable fields for incorporating visual information.

### 2.1 Database and Web Application

The data distribution included an XML file with the image ID, the captions of the images, the titles of the journal articles in which the image had appeared and the PubMed ID of the journal article. In addition, a compressed file containing the approximately 66,000 images was provided.

We used the Ruby programming language, with the open source Ruby On Rails web application framework[1, 2]. A PostgreSQL relational database was used to store mapping between the images and the various fields associated with the image. The title, full caption and precise caption, as provided in the data distribution, were indexed. The user interface for the search engine is given below in Fig.1.




[1] http://www.ruby-lang.org
[2] http://www.rubyonrails.org

**Figure 1 User Interface for OHSU search engine**

### 2.2 Image Processing and Analysis

The image itself has important visual characteristics such as color and texture that can help in the retrieval process. We created additional tables in the database to store image information that was created using a variety of image processing techniques in MATLAB[3]. These include color and intensity histograms as well as measures of texture using gray-level co-occurrence matrices and discrete cosine transforms[8]. These features can be used to find images that are visually similar to the query image. We used this in the interactive, mixed mode to reorder the images obtained from the textual search such that images that are visually similar to an image marked relevant by the user are returned at the top of the list.

Images that may have had information about the imaging modality or anatomy or view associated with them as part of the DICOM header can lose that information when the image is compressed to become a part of a teaching or on-line collection, as the image format used by these collections is usually compressed JPEG. In previous work [8], we described a modality classifier that can identify the imaging modality for medical images. We extended that work to the new dataset used for ImageCLEF 2008. Our system as previously described relied on a training set of modality-labeled images for its supervised learning. In 2008, we did not use any external database for training the modality classifier. Instead, a parser was written to extract the modality from the image caption. Images for which a single modality was parsed were used as the training set for the modality classifier. Grey scale images are classified into a set of modalities including x-rays, CT, MRI, ultrasound and nuclear medicine. Color image classes include gross pathology, microscopy, and endoscopy. The rest of the dataset (i.e., images for which zero or more than one modalities were parsed) was classified using the above classifier. We created two fields in the database for the modality that were indexed by our search engine. The first field contained the modality as extracted by the text parser, and the second contained the modality resulting from the classification process using visual features.

### 2.3 Query Parser and Search Engine

The system presents a variety of search options to the user including Boolean OR, AND, and "exact match.". There are also options to perform fuzzy searches, as well as a custom query parser. A critical aspect of our system is the query parser, written in Ruby. Ferret, a Ruby port of the popular Lucene system, was used in our system as the underlying search engine[4]. The custom query parser first performs stop word removal using a modified stop word list. The custom query parser is highly configurable, and the user has several configuration options from which to choose. The first such option is modality limitation. If the user selects this option, the query is parsed to extract the desired modality, if available. Using the modality fields described in the previous section, only those images that are of the desired modality are returned. We expect this to improve the precision as only images of the desired modality will be included within the result set. However, there might be a loss in recall if the modality extraction and classification process is not accurate.

The system is linked to the UMLS Metathesaurus. The second configuration option allows the user to perform manual or automatic query expansion using synonyms from the Metathesarus. In the manual mode, a list of synonyms is presented to the user, which the user can choose to add to the query. In the automatic mode, all synonyms of the UMLS preferred term are added to the query.

Another configuration option is the "stem and star" option, in which all the terms in the query are first stemmed. A wildcard (*) is then appended to the word to allow the search of words containing the desired root.

The last option allows the user to only send unique terms to the search engine. This can be useful when using the UMLS option, as many of the synonyms have a lot of overlap in the preferred terms.

### 2.4 Interactive mode

In addition to user-selectable search engine configuration options described above, our system provides users with other interactive features. Once a user has submitted a query using the above-described query parser, they have the option to improve the precision of their results by using an interactive re-ordering system. In this year's

[3] http://www.mathworks.com
[4] http://ferret.davebalmain.com

system, users select what they feel to be a visually representative image from their search's results. The system then attempts to re-rank the search results according to their degree of visual similarity with the "probe image" that the user selected. If the user is not satisfied with the re-ordering produced by their choice if image, they may repeat the process by selecting different probe images until they arrive at a satisfactory sorting.

To assess the visual similarity of the images within a result set, the system uses a relatively straightforward approach derived from Latent Semantic Analysis [17]. In this approach, each image in the result set is abstracted into a feature vector, which thereafter plays the same role that a document's "term vector" would play in classical LSA. We have experimented with sets of features derived from image color, texture, and frequency attributes; in our final system, the user is able to select which combinations of features they wish to use.

Once the feature vectors have been assembled for the images in a result set, they are combined into an $n \times m$ matrix. In this matrix, $n$ is equal to the number of images in the result set, and $m$ is equal to the number of features that the user has selected. Depending on the combination of features, this could be in the hundreds or low thousands. We then follow the classical LSA process, beginning by taking the Singular Value Decomposition (SVD) of our large matrix. This transforms our single matrix into three matrices that may be trivially recomposed to approximate the original matrix. The elements of one of these matrices represents the eigenvalues of the original document/term matrix; by varying the number of these elements that we use when recomposing the matrices, we may vary the fidelity of the resulting approximation.

After carrying out the SVD, we retain the first $r$ eigenvalues of the decomposed matrix, project the probe image's $m$-dimensional feature vector into the new lower-dimensional space, and, finally, compute the vector distance between the probe image's new representation and that of the images in the result set. In our system, the user is able to experiment with different values for $r$, and may pick the one that achieves the best performance for a given set of results. The user may also quickly and easily select different images to act as probe images, and can therefore evaluate many possible result sortings.

Obviously, this system's utility is variable, and depends highly on the contents of the initial result set. In the case of a set where the desired images are simultaneously visually similar to one another and distinct from the rest of the images in the set, this visual re-sorting system works quite well. However, in the case where the desired images are visually different from one another, or where all of the results (including the non-relevant ones) are visually similar, this re-sorting system is not very useful.

For example, a result set consisting entirely of ultrasound images will not be improved very much by re-sorting. In fact, in this particular case, resorting the result set may hurt its precision, as any ordering imposed by our textual search engine will be lost. On the other hand, a result set in which most of the relevant images are ultrasounds and most of the non-relevant images are x-rays could benefit from being re-ordered based on visual similarity to a user-selected probe image.

Our present system requires the user to select a combination of features to use. This is clearly sub-optimal, and our future work could include improved feature selection methods. Similarly, the user is currently able to change the number of eigenvalues used by the algorithm. While this is a powerful tool for tuning the algorithm's performance, it is also something that we would ultimately like to automate.

## 3 Runs Submitted

We submitted a total of 10 runs. The search options for the different runs are provided in table 1. These runs included textual and mixed, automatic and interactive options. Although the ImageCLEF2005-2007 collection with qrels and topics was available, we did not use any external training data.

Three automatic text-based runs were submitted with different custom parsing options including the use of UMLS term expansion. We also submitted four mixed, automatic runs. The modality classification based on the text parsing of the caption and the classification based on visual features was used to improve the precision of the search.

While the majority of our runs were automatic in nature, several of ours were interactive. In the first such run (ohsu_int_2), the user chose different combinations of options for each topic and added terms based on the list provided using the UMLS query expansion option. Two runs using the interactive result sorting system were submitted. The first such run, "ohsu_sdb_lsa", used the result sorting system on every topic. The second run, "ohsu_sdb_full_interactive", only used the result sorting system on topics where the user thought that it would be beneficial to the run's precision. This second run also featured much more intervention on the part of the user, who took full advantage of our retrieval system's interactive nature and enabled or disabled options and features as needed.

**Table 1.** OHSU runs submitted

| Run Name | Text/visual/ mixed | automatic/manual/ interactive | Data used | Parsing options |
|---|---|---|---|---|
| OHSU-text_or_1 | text | automatic | full caption | none |
| ohsu_text_3 | text | automatic | full caption, title | custom |
| ohsu_text_umls_4 | text | automatic | full caption, title | custom, umls, unique |
| ohsu_vis_mod_3 | mixed | automatic | full caption | custom, modality |
| OHSU-ohsu_mod_pars2_sp.txt | mixed | automatic | full caption | custom, modality |
| ohsu_vis_mod_5 | mixed | automatic | full caption, title | custom, modality |
| ohsu_vis_mod_umls_4 | mixed | automatic | full caption | custom, modality, umls |
| OHSU-ohsu_sdb_lsa.txt | mixed | interactive | full caption | custom, modality |
| ohsu_sdb_full_interactive.txt | mixed | interactive | full caption, title | custom, modality |
| ohsu_int_2 | mixed | interactive | precisise caption, title | custom, modality, umls |

## 4 Results and Discussion

Table 2 contains a subset of the official performance metrics for the OHSU runs. We have also included the average of these metrics for all runs, the highest measure in each category as well as data from the best run (based on MAP) in the 2008 campaign.

**Table 2.** Metrics of OHSU runs submitted

| Run Name | MAP | Bpref | P10 | P30 | Recall |
|---|---|---|---|---|---|
| OHSU-text_or_1 | 0.11 | 0.18 | 0.26 | 0.24 | 0.41 |
| ohsu_text_3 | 0.15 | 0.23 | 0.31 | 0.22 | 0.52 |
| ohsu_text_umls_4 | 0.20 | 0.30 | 0.29 | 0.25 | 0.57 |
| ohsu_vis_mod_3 | 0.15 | 0.25 | 0.32 | 0.24 | 0.53 |
| OHSU-ohsu_mod_pars2_sp.txt | 0.21 | 0.30 | **0.55** | **0.46** | 0.45 |
| ohsu_vis_mod_5 | 0.23 | 0.33 | 0.38 | 0.29 | 0.58 |
| ohsu_vis_mod_umls_4 | 0.23 | **0.35** | 0.37 | 0.28 | 0.60 |
| OHSU-ohsu_sdb_lsa.txt | 0.10 | 0.20 | 0.27 | 0.27 | 0.47 |
| ohsu_sdb_full_interactive.txt | 0.18 | 0.29 | 0.46 | 0.33 | 0.47 |
| ohsu_int_2 | 0.22 | 0.31 | 0.49 | 0.39 | 0.46 |
| average | 0.14 | 0.20 | 0.28 | 0.24 | 0.40 |
| best in category | 0.29 | 0.35 | 0.55 | 0.46 | 0.66 |
| SINAI-sinai_CT_Mesh_Fire20 | 0.29 | 0.33 | 0.43 | 0.40 | 0.62 |

OHSU performed reasonably well, especially among the runs that did not use any external training data. All but two of our runs performed better than the average for all measures. As described in the previous section, our systems have been designed to improve precision, perhaps at the expense of recall. Our custom parsing improved the mean average precision as well as the early precision, as can be seen in the text runs. The use of modality parsing and detection improved the MAP as well as the early precision. All our mixed runs performed better than the corresponding text runs. OHSU-ohsu_mod_pars2_sp.txt had the highest early precision (up to P30) of all official runs. OHSU had submitted four of the top ten mixed runs, as sorted using the precision as 10. The use of term expansion with UMLS increased the recall. We had submitted runs after the creation of the pools. This penalizes the runs as potentially fewer of the images are judged. One of these runs had the highest bpref, a measure that is robust in the case of incomplete judgments.

The performance of the first LSA run (ohsu_sdb_lsa) was unsatisfactory: as described earlier, there are many situations in which the original result sorting provided by our textual search engine was adequate, and changing

it by means of our interactive visual re-sorting system damaged a topic's precision. The second LSA run, "ohsu_sdb_full_interactive", performed much better. In fact, its p10 was greater than that of the overall competition winner's (0.46 for "ohsu_sdb_full_interactive" vs 0.43 for "SINAI-sinai_CT_Mesh_Fire20"). The third interactive run, where the parsing mode and UMLS term expansion was performed interactively also performed quite well.

### 5 Conclusions and Future Work

Our image retrieval system built using open-source tools is a flexible platform for evaluating various tools and techniques in image processing as well as natural language processing for medical image retrieval. The use of visual information to automatically extract the imaging modality is a promising approach for the ImageCLEFmed campaign. The use of UMLS term expansion, query parsing and modality detection all add value over the basic Ferret (Lucene) search engine. We will continue to improve our image retrieval system by adding more image tags using automatic visual feature extraction. Our next goal is to annotate the images with the their anatomical location and view attributes.

## Acknowledgments

We acknowledge the support of NLM Training Grant 1T15 LM009461 and NSF Grant ITR-0325160.

## References

1. Hersh, W, Muller H, et al. Advancing biomedical image retrieval: development and analysis of a test collection. J. Am. Med. Inform. Assoc. 13(5), 488-96, (2006)
2. Smeulders AWM, Worring M et al. Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349-1380, (2000)
3. Tagare HD, Jaffe C et al, Medical Image Databases: A Content-Based Retrieval Approach, J. Am. Med. Inform. Assoc. 4(3):184-198, (1997)
4. Aisen AM, Broderick LS, et al, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, Radiology, 228, 265-270, (2003)
5. Schmid-Saugeon P, Guillod J, et al, Towards a computer-aided diagnosis system for pigmented skin lesions, Computerized Medical Imaging and Graphics 27,65-78, 2003.
6. Müller H, Michoux N, Bandon D, Geissbuhler A, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, Int. J. Med. Inform., 73,1-23, (2004)
7. Hersh W, Kalpathy-Cramer J, et al. Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006, Springer Lecture Notes in Computer Science (LNCS), 660-669, (2006)
8. Kalpathy-Cramer J, Hersh W, Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval, MedInfo 2007, Brisbane, Australia, 1334-1338, (2007)
9. Braschler M, Peters C. Cross-language evaluation forum: objectives, results, achievements. Inform Retriev (7) 7–31 (2004)
10. Müller H, Deselaers T, Lehmann T, Clough P, Hersh W, Overview of the ImageCLEFmed 2006 medical retrieval annotation tasks, Evaluation of Multilingual and Multi-modal Information Retrieval, Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, LNCS , Alicante, Spain, 595-608, (2006)
11. Müller H, Clough P, et al, Evaluation Axes for Medical Image Retrieval Systems - The ImageCLEF Experience, ACM Int. Conf. on Multimedia, Singapore, November, (2005)
12. Florea F, Müller H, Rogozan A, Geissbühler A, Darmoni S. Medical image categorization with MedIC and MedGIFT. Medical Informatics Europe (MIE 2006)
13. Smith L, Rindflesch T, Wilbur W, MedPost: a part-of-speech tagger for biomedical text Bioinformatics 20(14), (2004)
14. Oliva, A, Torralba, A, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Computer Vision, 42(3): 145-175, (2001)
15. Nabney IT, Netlab: Algorithms for Pattern Recognition. London, England: Springer-Verlag. (2004)
16. Lowe DG, Distinctive image features from scale-invariant keypoints, Int. J. of Computer Vision, 60(2) :91-110, (2004)
17. Furnas G, Deerwester S, Dumais S, Landauer T, Harshman R, Streeter L, Lochbaum K. Information retrieval using a singular value decomposition model of latent semantic structure. SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (1988)