Medical Image Retrieval and Automatic Annotation: OHSU at ImageCLEF 2007

Jayashree Kalpathy-Cramer
William Hersh
Department of Medical Informatics & Clinical Epidemiology
Oregon Health and Science University
Portland, OR, USA
{kalpathy, hersh}@ohsu.edu

Abstract

Oregon Health & Science University participated in the medical retrieval and medical annotation tasks of ImageCLEF 2007. In the medical retrieval task, we created a web-based retrieval system for the collection built on a full-text index of both image and case annotations. The text-based search engine was implemented in Ruby using Ferret, a port of Lucene, and a custom query parser. 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. Purely textual runs as well as mixed runs using the purported modality were submitted. Our runs performed moderately well using the MAP metric and better for the early precision (P10) metric.

In the automatic annotation task, we used the 'gist' technique to create the feature vectors. Using statistics derived from a set of multi-scale oriented filters, we created a 512 dimensional vector. PCA was then used to create a 100-dimensional vector. This feature vector was fed into a two layer neural network. Our error rate on the 1000 test images was 67.8 using the hierarchical error calculations.

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

Performance, Image annotation, Experimentation, Algorithms

Keywords

Query parsing, Text retrieval, Image modality classification, Neural networks

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 alike.

a. 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 labor-intensive task of indexing and cataloging the images in these collections has traditionally been performed manually, a process that can be subjective and prone to errors.

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) 2007 campaign is a 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 an offshoot from the Text Retrieval Conference (TREC, trec.nist.gov).

b. System description of our adaptive medical image retrieval system

The ImageCLEF medical retrieval collection consists of about 66,000 medical images and annotations associated with them. We wanted to create a flexible database schema that could incorporate new collections easily while facilitating retrieval using both text and visual techniques. The text annotations in the collection are currently indexed and we continue to add indexable fields for incorporating visual information.

Database and web application

We used the Ruby programming language, with the open source Ruby On Rails web application framework [http://www.ruby-lang.org, http://www.rubyonrails.org]. A PostgreSQL relational database was used to store the images and annotations.

The database has images from the four different collections that were part of the ImageCLEFmed 2006 image retrieval challenge as well as two new collections for 2007. The approximately 66,000 images in these collections resided in cases, with annotations in English, German and/or French. The collections themselves were substantially heterogeneous in their architectures. Some collections had only one image per case while others had many images per case. Annotation fields were also quite different among the collections. Some collections had case-based annotations while others had image-based annotations. This difference is especially significant for text based retrieval as images of different modalities or anatomies or pathologies could be linked to the same case annotation. In this situation, even though only one image from a case containing many images might be relevant to a query (based on the annotation), all images for the case would be retrieved in a purely text based system, reducing the precision of the search.

We used the relational database to maintain the mappings between the collections, the cases in the collections, the cased-based annotations, the images associated with a collection, and the image based annotations.

Image Processing and Analysis

The image itself has important visual characteristics such as color and texture that can help in the retrieval process. 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.

We created additional tables in the database to store image information that was created using a variety of image processing techniques in MATLAB (www.mathworks.com). For instance, the images in the collection typically do not contain explicit details about the imaging modality. In previous work[8], we have described our modality classifier that can identify the imaging modality for medical images with a high level of confidence (>95% accuracy on the database used for the validation). 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.

Each image was annotated in the database with the purported image modality and a confidence value. This can be extremely useful for queries where the user has specified a desired image modality. An example query from ImageCLEF 2006 was "*Show me microscopic images of tissue from the cerebellum*".

The precision of the result of such a query can be improved significantly by restricting the images returned to those of the modality desired [8]. This is especially useful in eliminating images of the incorrect modality that may be part of a case containing a relevant image from the returned list of images. However, this increase in precision may result in a loss in recall if the classification algorithm incorrectly classifies the image modality.

We continue to experiment with a variety of image clustering and classification algorithms and adding the numerical data and labels to the database. Clustering images that look visually similar can be again used to improve the precision of the image retrieval process and speed up the system searching of images in the same cluster as the query image (if available).

Query parser and Search Engine

The system presents search options to the user including Boolean OR, AND and exact match. There are also options to perform fuzzy searches and custom query parsing.

The cornerstone 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 [http://ferret.davebalmain.com].

Queries were first analyzed using MedPost, a Parts-of-Speech (POS) Tagger created using the Medline corpus, and distributed by the National Library of Medicine [ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost] [14].

A simple Bayesian classifier [classifier.rubyforge.org] was trained to discern the desired image modality from the query, if available. The classifier performed extremely well within the constrained vocabulary of imaging modalities. Stop words were then removed from the query. These include Standard English stop words as well as a small set of stop words determined by analyzing queries from the last three years, including 'finding', 'showing', 'images', 'including' and 'containing'.

The system is also linked to the UMLS metathesarus. The user can choose to perform automatic query expansion using synonyms from the metathesarus.

A sample query "Show me CT images with a brain infarction" is automatically parsed and the following information is extracted from it: CT-> imaging modality, brain -> anatomic location, infarction -> finding. This information can be used to combine the results of the textual and visual systems more effectively.

c. Runs submitted

We submitted a total of 10 runs. These runs included textual and mixed, automatic and manual options. We also submitted runs using different weighted combinations of the FIRE baseline (published by the organizers) with our baseline textual runs. The run labels and a short description is given below.

d. Results and Discussion

Table 2 presents the mean average precision (MAP) and early precision of the OHSU runs as well that of the best overall run.

Our baseline textual system performed quite well, with a MAP of 0.3453. There was only a minimal improvement in performance with the use of the image modality. Our system is tuned to improve precision, at the cost of recall as we believe that most real users are interested in early precision. Consequently, our system demonstrated good early precision, with the highest P10 and high P30 results.

Our baseline textual system performed quite well, with a MAP of 0.3453. There was only a minimal improvement in performance with the use of the image modality. Our system is tuned to improve precision, at the cost of recall as we believe that most real users are interested in early precision. Consequently, our system demonstrated good early precision, with the highest P10 and high P30 results.

Hide Options ↑ Search
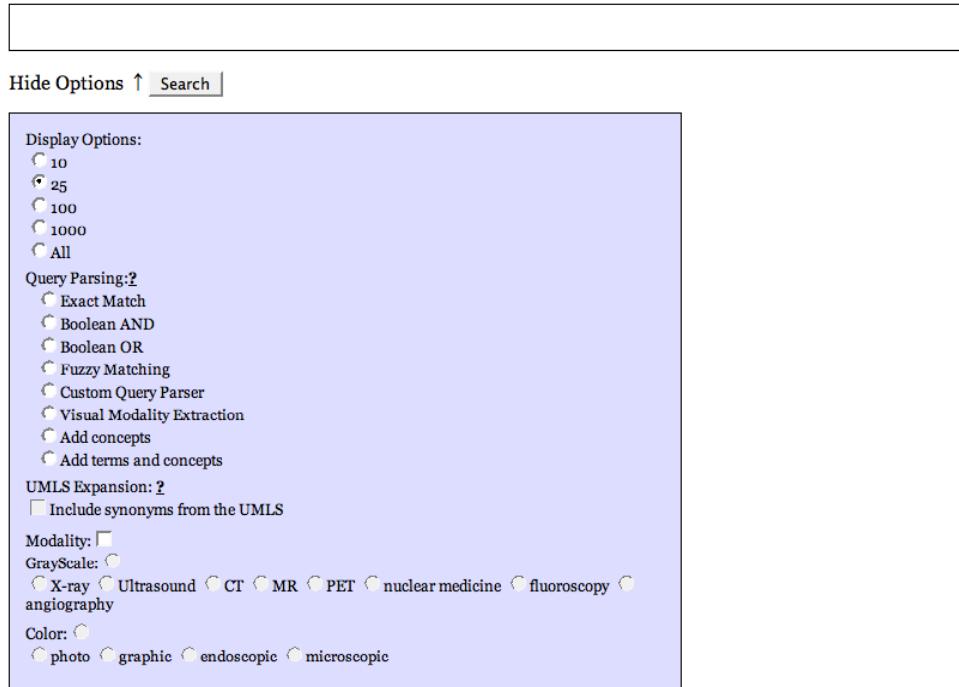
Display Options:
- ○ 10
- ◉ 25
- ○ 100
- ○ 1000
- ○ All

Query Parsing:**?**
- ○ Exact Match
- ○ Boolean AND
- ○ Boolean OR
- ○ Fuzzy Matching
- ○ Custom Query Parser
- ○ Visual Modality Extraction
- ○ Add concepts
- ○ Add terms and concepts

UMLS Expansion: **?**
- ☐ Include synonyms from the UMLS

Modality: ☐
GrayScale: ○
○ X-ray ○ Ultrasound ○ CT ○ MR ○ PET ○ nuclear medicine ○ fluoroscopy ○ angiography

Color: ○
○ photo ○ graphic ○ endoscopic ○ microscopic

Figure 1: Screen shot of our system displaying user options

run#	Label	Description
1	OHSU-iclefmed2007_as_out_1000rev1_c.txt.eval	Baseline textual run using the custom query parser (automatic)
2	OHSU-ohsu_m1.eval	Mixed run using modality to resort the results of the standard textual baseline, only 50 images returned per topic (automatic)
3	OHSU-OHSU_txt_exp2.eval	Textual run using UMLS term expansion (automatic)
4	OHSU-oshu_c_e_f_q.eval	Weighted combination of FIRE and OHSU textual runs (automatic)
5	OHSU-oshu_man2.eval	Manual query modification, purely textual (manual)
6	ohsu_comb3_ef_wt1_rev1_c.txt.eval	Weighted combination of FIRE and OHSU textual runs (automatic)
7	ohsu_fire_ef_wt2_rev1_c.txt.eval	Weighted combination of FIRE and OHSU textual runs (automatic)
8	ohsu_m2_rev1_c.txt.eval	Mixed run using modality to resort the results of the standard textual baseline (automatic)
9	ohsu_text_e4_out_rev1.txt.eval	Textual run using modified query parser (automatic)

Table 1: OHSU run labels and descriptions

Label	MAP	P10	p30
OHSU-iclefmed2007_as_out_1000rev1_c.txt.eval	0.3453	0.53	0.4433
OHSU-ohsu_m1.eval	0.2117	0.52	0.4578
OHSU-OHSU_txt_exp2.eval	0.3135	0.5867	0.4878
OHSU-oshu_c_e_f_q.eval	0.1129	0.2	0.1544
OHSU-oshu_man2.eval	0.3428	0.54	0.44
ohsu_comb3_ef_wt1_rev1_c.txt.eval	0.1134	0.3	0.1833
ohsu_fire_ef_wt2_rev1_c.txt.eval	0.0586	0.2	0.1211
ohsu_m2_rev1_c.txt.eval	0.3461	0.5567	0.4622
ohsu_text_e4_out_rev1.txt.eval	0.3321	0.5867	0.4878
LIG-MRIM-LIG_MU_A.eval	0.3962	0.5067	0.46

Table 2: Results of OHSU runs

Run Type	MAP
Visual	0.23
Mixed	0.35
Textual	0.47

Table 3: Results of OHSU best run by topic category

Although there was significant inter-category variation, our textual runs performed well on textual queries and more poorly on visual queries.

Future Work

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.

2. Automatic Image Annotation

The goal of this task was to correctly classify 1000 radiographic medical images using the hierarchical IRMA code. This code classifies the image along the modality, body orientation, body region, and biological system axes. There were 116 unique classes. The task organizers provided a set of 9,000 *training* images and 1000 *development* images. The goal of the task was to classify the images to the most precise level possible, with a greater penalty applied for incorrect classification than for a less specific classification in the hierarchy.

a. Introduction

A supervised machine learning approach using global gist features and neural network architecture was employed for the task of automatic annotation of medical images with the IRMA code.

b. System Description

The automatic image annotation was based on a neural network classifier using Gist features [14]. The classifiers were created in MATLAB using the Netlab toolbox [15].

All images were convolved with a set of 32 multiscale-oriented Gabor filters. We created a 512-dimensional vector using statistics from these filters. Principal component analysis was then used to reduce the dimensionality of the vector to 100.

A multilayer perceptron with one hidden layer containing 250-500 nodes was used to create and train a multi-class classifier. The training data set of 10,000 images was used to optimize performance of the development set of 1000 images. The final configuration of the classifier used 300 hidden nodes.

Images originally classified as classes 108 and 111 in the old code were commonly misclassified by the neural network classifier described above. To handle this special case, we created a second layer of classification built around a support vector machine (SVM) using scale-invariant feature transform (SIFT) features [16] as inputs. This new binary classify was used to determine the final class assignments for images in classes 108 and 111

c. Runs submitted

OHSU submitted two runs for the automatic annotation task. The first run used gist feature vectors to train the multi-layer perceptron. A neural network was used to create a multi-class classifier consisting of 116 classes. These were the original classes from 2006 and did not use the hierarchical nature of the IRMA code. These classes were then converted to the IRMA code, as required for the submission in 2007.

The second run used a hierarchical classifier architecture, with the first layer as described above and the second classifier using SIFT features and an SVM.

d. Results and Analysis

The relationship between semantic and visual hierarchy remains an open area of research. Based on our experiments using this collection of images used for automatic annotation, the use of hierarchy of the semantic classes did not improve our automatic annotations.

The error count for both our runs were quite similar at 67.8 and 67.97 for 1000 images, compared to the best count of 26.84 and worst count of 505.61. There was only a very slight improvement in using the two-layer classifier. There were 227 errors using the 2006 classes, which corresponds to an classification accuracy of 77.3%. However, of these 227 errors, only 15 were wrong along all 4 axes. 76 were misclassified along two axes (primarily view and anatomy) while 12 were misclassified along 3 axes. 77 of our single misclassifications were along the view axis. A significant portion of these occurred where class 111 was misclassified as 108, an error due to confusion between posterior-anterior and anterior-posterior views of the chest.

e. Future Work

We would like to further investigate the mapping between the semantic and visual hierarchy of images in the IRMA collection. We also plan to further explore specifying the image classification in the semantic hierarchy only to the level

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. (JAMIA), 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).
8. Kalpathy-Cramer J, Hersh W, *Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval*, accepted to MedInfo 2007, Brisbane, Australia, 2007.

9. Braschler M, Peters C. Cross-language evaluation forum: objectives, results, achievements. Inform Retriev 2004 (7) 7–31.
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 2006,Alicante, Spain, to appear.
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*. 2004, London, England: Springer-Verlag
16. Lowe DG, *Distinctive image features from scale-invariant keypoints*, Int. J. of Computer Vision, 60(2) :91-110, 2004