Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006

William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen

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

Abstract. Oregon Health & Science University participated in both the medical retrieval and medical annotation tasks of ImageCLEF 2006. Our efforts in the retrieval task focused on manual modification of query statements and fusion of results from textual and visual retrieval techniques. Our results showed that manual modification of queries does improve retrieval performance, while data fusion of textual and visual techniques improves precision but lowers recall. However, since image retrieval may be a precision-oriented task, these data fusion techniques could be of value for many users. In the annotation task, we assessed a variety of learning techniques and obtained classification accuracy of up to 74% with test data.

1 Introduction

Oregon Health & Science University (OHSU) participated in the medical image retrieval task and automatic annotation tasks of ImageCLEF 2006 (Müller, Deselaers et al., 2006). Similar to our approach for the medical retrieval tasks in 2005, we focused on two primary areas: the manual modification of queries and the use of visual information in series with the results of the textual searches.

The automatic annotation task provided a forum to evaluate our algorithms for medical image classification and annotation. We evaluated the performance of a variety of low level image features using a two-layer neural network architecture for the classifier. This classifier was then used in conjunction with textual results of the image retrieval system to improve the precision of the search for the medical retrieval tasks.

2 Image Retrieval

The goal of the ImageCLEF medical image retrieval task is to retrieve relevant images from a test collection of about 50,000 images that are annotated in a variety of formats and languages (Müller, Deselaers et al., 2006). Thirty topics were developed, evenly divided as amenable to textual, visual, or mixed retrieval techniques. The top-ranking images from runs by all participating groups were judged as definitely, possibly, or not relevant by relevance judges.

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 660 – 669, 2007.

The mission of information retrieval research at Oregon Health & Science University (OHSU) is to better understand the needs and optimal implementation of systems for users in biomedical tasks, including research, education, and clinical care. The goals of the OHSU experiments in the medical image retrieval task of ImageCLEF were to assess manual modification of topics with and without visual retrieval techniques. We manually modified the topics to generate queries, and then used what we thought would be the best run (which in retrospect was not) for combination with visual techniques, similar to the approach we took in ImageCLEF 2005 (Jensen and Hersh, 2005).

2.1 System Description

Our retrieval system was based on the open-source search engine, Lucene (Gospodnetic and Hatcher, 2005), which is part of the Apache Jakarta distribution. We have used Lucene in other retrieval evaluation forums, such as the Text Retrieval Conference (TREC) Genomics Track (Cohen, Bhuptiraju et al., 2004; Cohen, Yang et al., 2005). Documents in Lucene are indexed by parsing of individual words and weighting of those words with an algorithm that sums for each query term in each document the product of the term frequency (TF), the inverse document frequency (IDF), the boost factor of the term, the normalization of the document, the fraction of query terms in the document, and the normalization of the weight of the query terms, for each term in the query. The score of document d for query q consisting of terms t is calculated as follows:

$$score(q,d) = \sum_{t\ in\ q} tf(t,d) * idf(t) * boost(t,d) * norm(d,t) * frac(t,d) * norm(q)$$

where:

tf(t,d) = term frequency of term t in document d
idf(t) = inverse document frequency of term t
boost(t,d) = boost for term t in document d
norm(t,d) = normalization of d with respect to t
frac(t,d) = fraction of t contained in d
norm(q) = normalization of query q

As Lucene is a code library and set of routines for IR functionality, it does not have a standard user interface. We have therefore also created a search interface for Lucene that is tailored to the ImageCLEF medical retrieval test collection structure (Hersh, Müller et al., 2006) and the ability to use the MedGIFT search engine for visual retrieval on single images (Müller, Geissbühler et al., 2005). We did not use the user interface for these experiments, though we plan to undertake interactive user experiments in the future.

2.2 Runs Submitted

We submitted three general categories of runs:

- Automatic textual - submitting the topics as phrased in the official topics file directly into Lucene. We submitted each of the three languages in separate runs,

along with a run that combined all three languages into a single query string and another run that included the output from the Babelfish translator[1]

- Manual textual - manually editing of the official topic files by one of the authors (WRH). The editing mostly consisted of removing function and other common words. Similar to the automatic runs, we constructed query files in each of the three languages, along with a run that combined all three languages into a single query string and a final run that included the output from the Babelfish translator.
- Interactive mixed - a combination of textual and visual techniques, described in greater detail below.

The mixed textual and visual run was implemented as a serial process, where the results of what we thought would be our best textual run were passed through a set of visual retrieval steps. This run started by using the top 2000 retrieved images of the OHSU_all textual run. These results were combined with the top 1000 results distributed from the medGIFT (visual) system. Only those images that were in both lists were chosen. These were ordered by the textual ranking, with typically 8 to 300 images in common.

A neural network-based scheme using a variety of low level, global image features was used to create the visual part of the retrieval system. The retrieval system was created in MATLAB[2] using Netlab[3] (Bishop, 1995; Nabney, 2004). We used a multilayer perceptron architecture to create the the two-class classifiers to determine if a color image was a 'microscopic' image or 'gross pathology.' It was a two layer structure, with a hidden layer of approximately 50-150 nodes. A variety of combinations of the image features were used as inputs. All inputs to the neural network (the image feature vectors) were normalized using the training set to have a mean of zero and variance of 1.

Our visual system then analyzed the sample images associated with each sub-task. If the query image was deemed to be a color image by the system, the set of top 2000 textual images was processed and those that were deemed to be color were moved to the top of the list. Within that, the ranking was based on the ranking of the textual results.

A neural network was created to process color images to determine if they were microscopic or gross pathology/photograph. The top 2000 textual results were processed through this network and the appropriate type of image (based on the query image) received a higher score. Relevance feedback was used to improve the training for the network (Crestani, 1994; Han and Huang, 2005; Wang and Ma, 2005). Low level texture features based on grey-level co-occurrence matrices (GLCM) were used as input to the neural network (Haralick, 1979; Rahman, Desai et al., 2005). We also created neural networks for a few classes of radiographic images, based on the system that we had used for the automatic annotation class (described in detail in the next section). Images identified as being of the correct class received a higher score.

The primary goal of these visual techniques was to move the relevant images higher on the ordered list of retrieved images, thus leading to higher precision.

[1] http://babelfish.altavista.com
[2] http://www.mathworks.com
[3] http://www.ncrg.aston.ac.uk/netlab/index.php

However, we would be limited in the recall to only those images that had already been retrieved by the textual search. Thus, even in the ideal case, where all the relevant images were moved to the top of the list, the MAP would be limited by the number of relevant images that were retrieved by the textual search (recall of the textual search).

We improved our modality classifier after submitting our official runs to ImageCLEF. We can now classify images into 6 categories for color images (stained histology (microscopic) images, photographs and gross pathology images, electroencephalograhical images (EEGs) and electrocardiograms (ECGs), Powerpoint slides, as well as a few miscellaneous images. Grey-scale images are also classified into modalities including angiography, computerized tomography scans (CT), X-ray, Magnetic resonance (MRI), ultrasound, and scintigraphy. These classifiers, achieving >95% accuracy, were tested on a random subset of the ImageCLEFmed topics.

2.3 Results and Analysis

The characteristics of the submitted OHSU runs are listed in Table 1, with various results shown in Figure 1. The automatic textual runs were our lowest scoring runs. The best of these runs was the English-only run passed through the Babelfish translator, which obtained a MAP of 0.1264. The remaining runs all performed poorly, with all MAP results under 0.08. The manual textual runs performed somewhat better. Somewhat surprising to us, the best of these runs was the English-only run (OHSUeng). This was our best run of all, with a MAP of 0.2132. It outperformed an English-only run with terms from automatic translation added (OHSUeng_trans, with a MAP of 0.1906) as well as a run with queries of topic statements from all languages (OHSUall, with a MAP of 0.1673).

The MAP for our interactive-mixed run, OHSU_m1, was 0.1563. As noted above, this run was based on modification of OHSUall, which had a MAP of 0.1673. At a first glance, it appears that performance was worsened with the addition of visual techniques, due to the lower MAP. However, as seen in Figure 1, and similar to our results from 2005, the average precision at various numbers of images retrieved was higher, especially at the top of the retrieval list. This confirmed our finding from 2005 that visual techniques used to modify textual runs diminish recall-oriented measures like MAP but improve precision at the very top of output list, which may be useful to real users. There was a considerable variation in performance on different topics. For most topics, the addition of visual techniques improved early precision, but for some, the reverse was true.

We also looked at MAP for the tasks separated by their perceived nature of the question (one favoring visual, semantic, or mixed techniques, see Table 2). For the visual and mixed queries, the incorporation of visual techniques improved MAP. However, for semantic queries, there was a serious degradation in MAP by the addition of the visual steps in the retrieval process. This, however, is driven by only one query, number 27, where MAP for OHSU_all was 0.955, while for OHSU_m1 was 0.024. Excluding this query, MAP for OHSU_m1 was 0.161 while that of OHSU_all was 0.140, indicating a slight improvement as a result of the addition of visual techniques.

Table 1. Characteristics of OHSU runs

Run_ID	Type	Description
OHSU_baseline_trans	Auto-Text	Baseline queries in English translated automatically
OHSU_english	Auto-Text	Baseline queries in English only
OHSU_baseline_notrans	Auto-Text	Baseline queries in all languages
OHSU_german	Auto-Text	Baseline queries in German only
OHSU_french	Auto-Text	Baseline queries in French only
OHSUeng	Manual-Text	Manually modified queries in English
OHSUeng_trans	Manual-Text	Manually modified queries in English translated automatically
OHSU-OHSUall	Manual-Text	Manually modified queries in all three languages
OHSUall	Manual-Text	Manually modified queries in all three languages
OHSUger	Manual-Text	Manually modified queries in German
OHSUfre	Manual-Text	Manually modified queries in French
OHSU-OHSU_m1	Interactive-Mixed	Manually modified queries filtered with visual methods

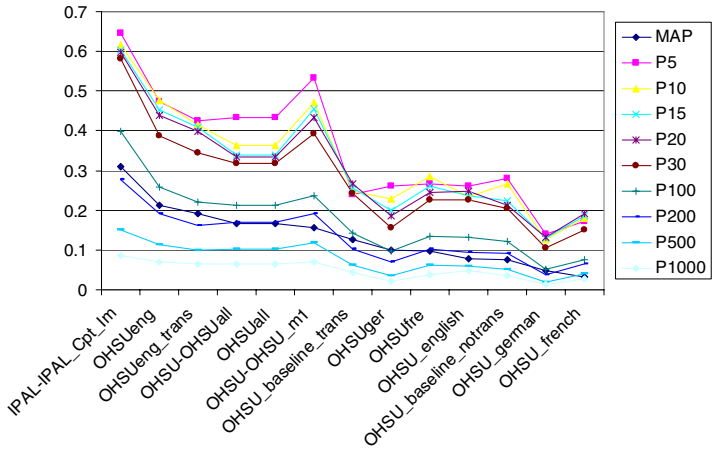

Fig. 1. MAP and precision at various retrieval levels for all OHSU runs and the run with the best overall MAP from ImageCLEFmed 2006, IPAL-IPAL_Cpt_Im

Table 2. MAP by query type for mixed and textual queries

Query Type	MAP	
	OHSU_m1	OHSUall
Visual	0.139	0.128
Mixed	0.182	0.148
Semantic	0.149	0.226

Table 3. Modality of images returned using textual query for topic 25, *Show me microscopic images of tissue from the cerebellum*

Image type	Number of images
Total returned by textual query	2000
Grey-scale	1476
Photograph/gross pathology	408
Microscope	116

The new color modality classifier was tested on a small random subset of the ImageCLEFmed 2006 topics. Analysis of our textual results indicated that in many queries, especially those of a visual or mixed nature, up to 75% of the top 1000 results were not of the correct modality. An example of this is shown in Table 3. Only 90 of the top 2000 images returned by the textual query alone were of the desired modality. The precision of this search was improved by the use of the modality detector as seen in Figure 2.

Our runs demonstrated that manual modification of topic statements makes a large performance difference, although our results are not as good as some groups that did automatic processing of the text of topics. Our results also showed that visual retrieval techniques provide benefit at the top of the retrieval output, as demonstrated by higher precision at various output levels, but are detrimental to recall, as shown by lower MAP.

3 Automated Image Annotation

The goal of this task was to correctly classify 1000 radiographic medical images into 116 categories (Müller, Deselaers et al., 2006). The images differed in the "modality, body orientation, body region, and biological system examined," according to the track Web site. The task organizers provided a set of 9,000 *training* images that were classified into these 116 classes. In addition, another set of classified images (numbering 1000) was provided as a *development* set. The development set could be used to evaluate the effectiveness of the classifier prior to the final evaluation.

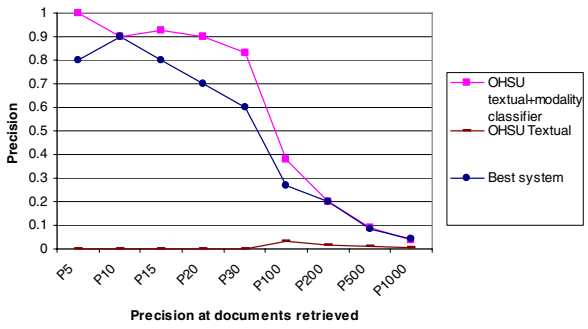

Fig.2. Improvement of precision for topic 25 using the image modality classifier in series with textual results

We used a combination of low-level image features and a neural network based classifier for the automated image annotation task. Our results were in the middle of the range of results obtained for all groups, indicating to us the potential capabilities of these techniques as well as some areas of improvement for further experiments.

3.1 System Description

A neural network-based scheme using a variety of low-level, largely global, texture and histogram image features was used to create the classifier. The system was implemented in MATLAB using the Netlab toolbox. All images were padded to create a square image and then resized to 256x256 pixels. A variety of features described below were tested on the development set. These features were combined in different ways to try to improve the classification ability of the system, with the final submissions were based on the three best combinations of image features. The features included:

- *Icon*: A 16x16 pixel 'icon' of the image was created by resizing the image using bilinear extrapolation.
- *GLCM*: Four gray level co-occurrence matrices (GLCM) (Haralick, 1979) matrices with offsets of 1 pixel, 0, 45, 90 and 135 degrees were created for the image after rescaling the image to 16 levels. GLCM statistics of contrast, correlation, energy, homogeneity and entropy were calculated for each matrix. A 20 dimensional vector was created for each image by concatenating the 5 dimensional vector obtained by each of the four offset matrices.
- *GLCM2*: In order to capture the spatial variation of the images in a coarse manner, the resized image (256x256) was partitioned into 5 squares of size 128x128 pixels (top left, top right, bottom left, bottom right, centre). A gray level correlation matrix was created for each partition. The five 20-dimensional vectors from each of the partitions were concatenated to create a feature vector of dimension 100.
- *Hist:* A 32-bin histogram was created for each image and counts were used as the input

We used a multilayer perceptron architecture to create the multi-class classifier(Bishop, 1995; Nabney, 2004). It was a two layer structure, with a hidden layer of approximately 200-400 nodes. A variety of combinations of the above image features were used as inputs. All inputs to the neural network (the image feature vectors) were normalized using the training set to have a mean of zero and variance of 1. The architecture was optimized using the training and development sets provided.

The network architecture, primarily the number of hidden nodes, needed to be optimized for each set of input feature vectors, since the length of the feature vectors varied from 32 to 356. The training set was used to create the classifier, typically with the accuracy increasing with an increase in the number of hidden nodes. It was relatively easy to achieve 100% classification accuracy on the training set. However, there were issues with overfitting if too many hidden nodes were used. We used empirical methods to optimize the network for each set of feature vectors by using a

network architecture that resulted in the highest classification accuracy for the development set.

3.2 Runs Submitted

We submitted four runs, iconGLCM2 using just the training set for creating the net, iconGLCM2 using the development and training set for creating the net, iconHist, and iconHistGLCM.

3.3 Results and Analysis

The best results for the development set were obtained using a 356 dimensional normalized input vector consisting of the icon (16x16) concatenated with the GLCM vectors of the partitioned image. The classification rate on the training set was 80%. The next best result was obtained using a 288 dimensional normalized input vector consisting of the icon (16x16) concatenated with Hist. The classification rate on the development set was 78%. Most other runs including just the icon or GLCM2 gave about 70-75% classification accuracy, as seen in Table 4. However, the results obtained on the test set were lower than those of the development set.

A few classes were primarily responsible for the differences seen between the development set and test set. Class 108 saw the most significant difference. Most of the misclassification of class 108 was into class 111, visually a very similar class. Observing the confusion matrices in general for all the runs, the most misclassifications occurred between classes 108 and 111, and 2 and 56.

Table 4. Classification rates for OHSU automatic annotation runs

Feature vector	Classification rate	
	Development	Test
iconHist	78	69
iconGLCMHist	78	72
iconGLCM2	80	74

Following our availability of the results, we performed additional experiments to improve the classification between these sets of visually similar classes. We created two new classifiers to distinguish between class 2 and 56, and between class 108 and 111. We merged images labeled by the original classifier as class 2 and 56, and class 108 and 111 and then applied the new classifiers on these newly merged classes. Using this hierarchical classification, we improved our classification accuracy by about 4% (to 79%) overall for the test set. This seems like a promising approach to improve the classification ability of our system.

One of the issues with the database is that the number of training images in each of the classes is quite varied. Another issue is that there are some classes that are visually quite similar while other classes that have quite a bit of within class variation. These issues were proved to be a little challenging for our system.

4 Conclusions

Manual modification of topic statements improved our performance for the image retrieval task. Inclusion of visual techniques in series with our textual result increased precision, but was detrimental to recall, depending on the techniques used. However, for most image retrieval tasks, precision may be more important than recall, so visual techniques may be of value in real-world image retrieval systems. Additional research on how real users query image retrieval systems could shed light on which system-oriented evaluation measures are most important.

Also suggested by our runs is that system performance is dependent upon the topic type. In particular, visual retrieval techniques degrade the performance of topics that are most amenable to textual retrieval techniques. This indicates that systems that can determine the query type may be able to improve performance with that information. However, use of image-based modality classifier can improve the precision of the retrieval, even for tasks that are amenable to textual means.

We obtained moderate results in the ImageCLEFmed automatic annotation task using a neural network approach and primarily low level global features, The best results were obtained by using a feature vector consisting of a 16x16 icon and grey-level co-occurrence features. A multi-layer perceptron architecture was used for the neural network. In the future, we plan to explore using a hierarchical set of classifiers to improve the classification between visually similar classes (for instance, different views of the same anatomical organ). This might also work well with the IRMA classification system.

Acknowledgements

This work was funded in part by NSF Grant ITR-0325160 and NLM Training Grant 1T15 LM009461.

References

Bishop, C.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)

Cohen, A., Bhuptiraju, R., Hersh, W.: Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In: The Thirteenth Text Retrieval Conference: TREC 2004, Gaithersburg, MD, National Institute of Standards and Technology (2004)

Cohen, A., Yang, J., Hersh, W.: A comparison of techniques for classification and ad hoc retrieval of biomedical documents. In: The Fourteenth Text Retrieval Conference - TREC 2005, Gaithersburg, MD, National Institute for Standards & Technology (2005)

Crestani, F.: Comparing probabilistic and neural relevance feedback in an interactive information retrieval system. In: Proceedings of the 1994 IEEE International Conference on Neural Networks, Orlando, Florida, pp. 3426–3430 (1994)

Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning Publications, Greenwich, CT (2005)

Han, J., Huang, D.: A novel BP-based image retrieval system. In: IEEE International Symposium on Circuits and Systems, Kobe, Japan, 2005, pp. 1557–1560. IEEE, Los Alamitos (2005)

Haralick, R.: Statistical and structural approaches to texture. Proceedings of the IEEE 67, 786–804 (1979)

Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: Advancing biomedical image retrieval: development and analysis of a test collection. Journal of the American Medical Informatics Association 13, 488–496 (2006)

Jensen, J., Hersh, W.: Manual query modification and data fusion for medical image retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 673–679. Springer, Heidelberg (2006)

Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval - Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. Springer Lecture Notes in Computer Science, Springer, Heidelberg (in press, 2006)

Müller, H., Geissbühler, A., Marty, J., Lovis, C., Ruch, P.: The use of MedGIFT and EasyIR for Image CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 724–732. Springer, Heidelberg (2006)

Nabney, I.: Netlab: Algorithms for Pattern Recognition. Springer, London, England (2004)

Rahman, M., Desai, B., Bhattacharya, P.: Supervised machine learning based medical image annotation and retrieval in Image CLEFmed 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 602–701. Springer, Heidelberg (2006)

Wang, D., Ma, X.: A hybird image retrieval system with user's relevance feedback using neurocomputing. Informatica 29, 271–279 (2005)