

# Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006

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

## Abstract

Oregon Health & Science University participated in both the medical retrieval and medical annotation tasks of ImageCLEF 2005. 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.

## 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

Image retrieval, Performance, Image annotation, Experimentation

## Keywords

Manual query modification, Data fusion, Classification, Neural networks

## 1. 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. 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.

### a. Introduction

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 [1].

### b. System Description

Our retrieval system was based on the open-source search engine, Lucene [2], 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 [3, 4]. 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:

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

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 [5] and the ability to use the MedGIFT search engine for visual retrieval on single images [6]. We did not use the user interface for these experiments, though we plan to undertake interactive user experiments in the future.

### c. 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 (<http://babelfish.altavista.com/>).
- 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. The manually modified query strings are listed in Table 1.
- 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 using Netlab [7, 8]. 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.

Table 1 - Manually modified queries for OHSU manual textual runs.

Topic	English	German	French
1	oral cavity including teeth and gum tissue	Mundhöhle mit Zähnen und Zahnfleisch	cavité buccale incluant des dents et du tissu des gencives
2	frontal head MRI	MR Frontalaufnahmen des Kopfes	IRM frontal du crâne
3	knee x-ray	Röntgenbilder des Knies	radiographies du genou
4	x-ray of a tibia with a fracture	Röntgenbilder einer gebrochenen Tibia	radiographies du tibia avec fracture
5	x-ray of a hip joint with prosthesis	Röntgenbilder eines Hüftgelenks mit Prothese	radiographies d'articulation de la hanche avec une prothèse
6	hand x-ray	Röntgenbilder einer Hand	des radiographies de la main
7	ultrasound with a triangular result	Ultraschallbilder mit dreieckigem Ergebnis	des échographies de résultats triangulaires
8	PowerPoint slides	von Powerpoint Folien	des images de diapositives PowerPoint
9	EEG or ECG	EEG oder EKG	EEG ou ECG
10	chest CT with nodules	CT der Lunge mit Knötchen	CTs du thorax avec nodules
11	ultrasound with gallstones	Ultraschallbilder mit Gallensteinen	échographies de calculs biliaires
12	chest x-ray with tuberculosis	Röntgenbilder der Lunge mit Tuberkulose	radiographies de la poitrine avec une tuberculose
13	CT with a brain infarction	CT eines Gehirnschlages	CT avec un infarctus cérébral
14	MRI of the brain with a blood clot	MR des Gehirns mit Blutgerinnsel	IRM du cerveau avec un caillot sanguin
15	x-ray of vertebral osteophytes	Röntgenbilder von vertebrealen Osteophyten	radiographies d'ostéophytes vertébraux
16	ultrasound of a foetus or fetus	Ultraschallbilder eines Fötus	échographies d'un foetus
17	abdominal CT of an aortic aneurysm	CT des Abdomens mit einem Aneurismus der Aorta	CTs abdominaux d'un anévrisme aortique
18	blood smears that include polymorphonuclear neutrophils	Blutabstriche mit polymorphonuklearer Neutrophils	échantillons de sang incluant des neutrophiles polymorphonucléaires
19	multinucleated giant cells	mehrkernige riesenzellen	cellules géantes multinucléées
20	lung tissue	Lungengewebe lung	tissu pulmonaire
21	infected wound	infizierten Wunde wound	plaie infectée wound
22	tumours or tumors	Tumoren	tumeurs
23	CT or x-ray of heart	CT oder Röntgenbilder des Herzens	CT ou des radiographies qui montrent le coeur
24	muscle cells	Muskelzellen	cellules musculaires
25	tissue from the cerebellum	Kleinhirngewebe kleinhirn	tissu du cervelet
26	x-ray of bone cysts	Röntgenbilder von Knochenzysten	radiographies de kystes d'os
27	Budd-Chiari malformation	Budd-Chiari Verformung	malformation de Budd-Chiari
28	parvovirus infection	Parvovirusinfektion parvovirus infection	infection parvovirale
29	bacterial meningitis	bakteriellen Hirnhautentzündung meningitis	méningite bactérienne meningitis bacterial
30	findings with Alzheimer's Disease	Fällen mit einer Alzheimer Diagnose	observations avec la maladie d'Alzeimer

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 [9-11]. Low level texture features based on grey-level co-occurrence matrices (GLCM) were used as input to the neural network [12, 13]. 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. 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).

#### d. Results and Analysis

The characteristics of the submitted OHSU runs are listed in Table 2, 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.

Table 2 - 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 only
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 only
OHSUfre	Manual-Text	Manually modified queries in French only
OHSU-OHSU_m1	Interactive-Mixed	Manually modified queries filtered with visual methods

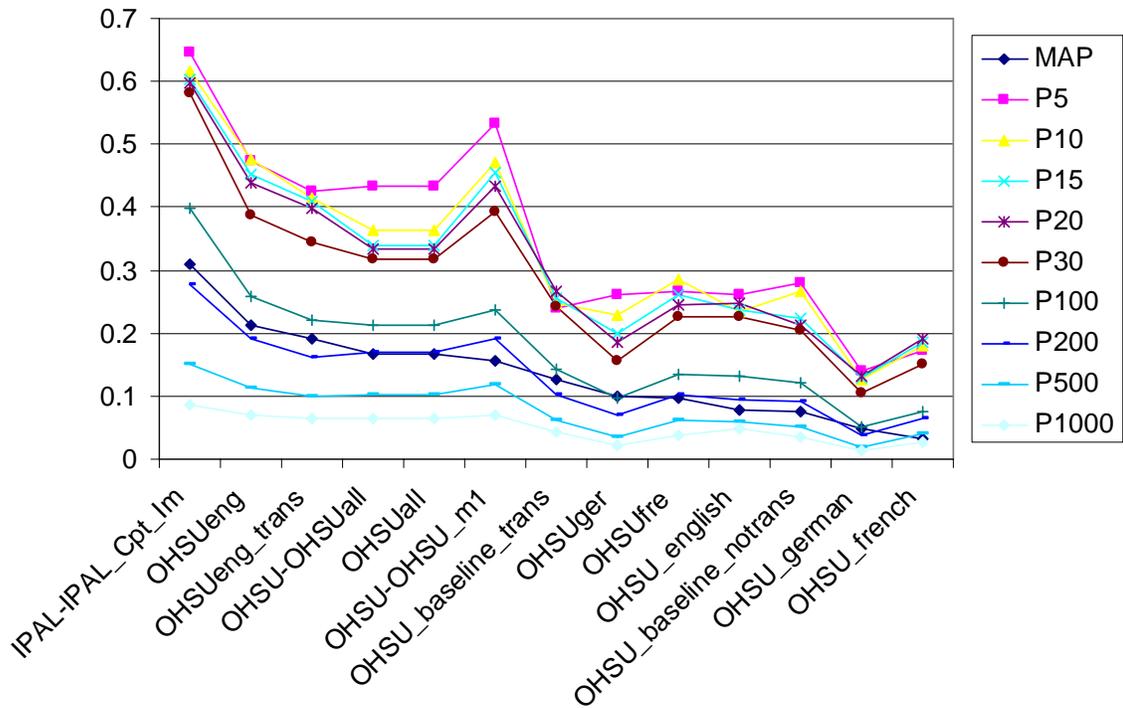


Figure 1 - MAP and precision at various retrieval levels for all OHSU runs and the run with the best overall MAP from ImageCLEFmed 2006, IPAL-IPAL\_Cpt\_Im.

We also looked at MAP for the tasks separated by their perceived nature of the question (one favoring visual, semantic, or mixed techniques). 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 for the addition of visual techniques.

### e. Conclusions

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. 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.

Table 3 - MAP by query type for mixed and textual runs.

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

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.

## 2. Automated Image Annotation

The goal of this task was to correctly classify 1000 radiographic medical images into 116 categories. 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 suggested procedure was to create a classifier based on the training images. The development set could then be used to test the effectiveness of the classifier. One could then combine the training and development tests to create a larger database to create the final classifier for the *test* images.

### a. Introduction

For the automated image annotation task, we used a combination of low-level image features and a neural network based classifier. Our results (error rate of 26.3% for our best run) 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.

### b. System Description

A neural network-based scheme using a variety of low-level, largely global image features was used to create the classifier, which was implemented in MATLAB using Netlab. A variety of feature vectors were then tested with the results. For our first efforts in the medical image automatic annotation domain, we started with low-level, commonly used, global, texture and histogram features. In addition, we tried to capture a sense of spatial differences between images classes.

Images were first padded to create a 512x512 image, with the original image centered within this new image. White (255) and black (0) pixels were tested for the padding. This was done since we had noted that the aspect ratio of the image can provide information useful for classification. All images were resized to 256x256 pixels using bilinear extrapolation.

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. This vector of dimension 256 was fed directly into the input of the neural network
- *GLCM*: Four gray level co-occurrence matrices (GLCM) [ Haralick] 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 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. A 20 dimensional vector was created for each partition. Subsequently, the 5 vectors from each of the partitions were concatenated to created feature vector of dimension 100.
- *Hist*: A 32-bin histogram was created for each image and counts were used as the input
- *DCT*: A global discrete cosine transform was created for each image. The upper left (10x10) vectors were concatenated and used as inputs

We used a multilayer perceptron architecture to create the multi-class classifier[7, 8]. 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 (see Figure 2). 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. For instance, for the feature vectors consisting of iconHist features, we would use 300 hidden nodes, while for iconGLCM, we would use a network consisting of 200 hidden nodes.

### c. 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.

### d. 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. 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 DCT 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.

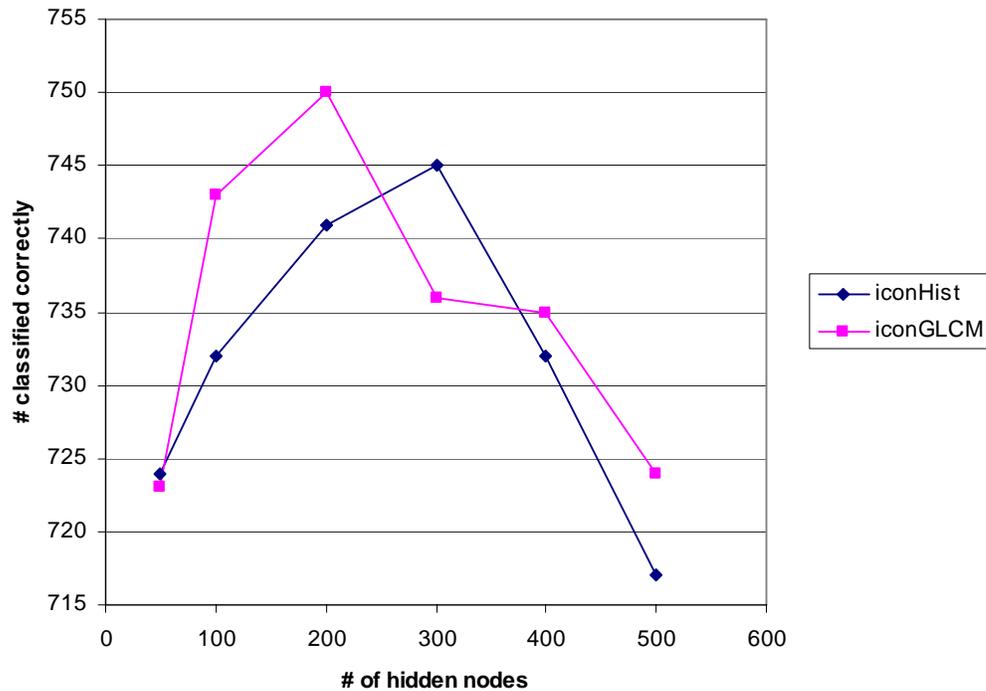


Figure 2 - Images classified correctly vs. number of hidden nodes.

Table 4 - Classification rates for OHSU automatic annotation runs.

Feature vector	Classification rate	
	Development	Test
DCT	71	-
icon	74	-
iconDCT	75	-
iconHist	78	69
iconGLCM	78	-
iconGLCMHist	78	72
iconGLCM2	80	74

Analyzing the data, it appeared that a few classes were primarily responsible for the differences seen between the development set and test set (see Table 5). Class 108 had the most significant difference seen, which was about 2.4% of the 6% difference seen in iconGLCM2. 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 were between classes 108/111 and 2/56.

Following our availability of the results, we performed additional experiments aiming to improve the classification between these sets of visually similar classes. We created two new additional 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.

#### e. Conclusions

Using a neural network approach and primarily low level global features, we obtained moderate results in the ImageCLEFmed automatic annotation task. 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.

Table 5 - Differences for select classes.

Class	Development set		Test set		Difference in error count
	Count	# correct	Count2	#correct2	
108	93	78	92	54	23
61	21	21	20	16	4
44	10	7	10	2	5
12	23	21	22	16	4

## Acknowledgements

This work was funded in part by Grant ITR-0325160 of the US National Science Foundation.

## References

1. Jensen J and Hersh W. *Manual query modification and data fusion for medical image retrieval*. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science. 2005. Vienna, Austria. in press. <http://medir.ohsu.edu/~hersh/imageclef-OHSU-05.pdf>.
2. Gospodnetic O and Hatcher E, *Lucene in Action*. 2005, Greenwich, CT: Manning Publications.
3. Cohen AM, Bhuptiraju RT, and Hersh W. *Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage*. *The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. <http://trec.nist.gov/pubs/trec13/papers/ohsu-hersh.geo.pdf>.
4. Cohen AM, Yang J, and Hersh WR. *A comparison of techniques for classification and ad hoc retrieval of biomedical documents*. *The Fourteenth Text Retrieval Conference - TREC 2005*. 2005. Gaithersburg, MD: National Institute for Standards & Technology. <http://trec.nist.gov/pubs/trec14/papers/ohsu-geo.pdf>.
5. Hersh WR, et al., *Advancing biomedical image retrieval: development and analysis of a test collection*. Journal of the American Medical Informatics Association, 2006: Epub ahead of print. <http://www.jamia.org/cgi/reprint/M2082v1>.
6. Müller H, et al. *The use of MedGIFT and EasyIR for ImageCLEF 2005*. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science. 2005. Vienna, Austria: Springer-Verlag. in press.
7. Bishop CM, *Neural Networks for Pattern Recognition*. 1995, Oxford: Clarendon Press.
8. Nabney IT, *Netlab: Algorithms for Pattern Recognition*. 2004, London, England: Springer-Verlag.
9. Crestani F. *Comparing probabilistic and neural relevance feedback in an interactive information retrieval system*. *Proceedings of the 1994 IEEE International Conference on Neural Networks*. 1994. Orlando, Florida. 3426-3430.
10. Han JH and Huang DS. *A novel BP-based image retrieval system*. *IEEE International Symposium on Circuits and Systems, 2005*. 2005. Kobe, Japan: IEEE. 1557- 1560.
11. Wang D and Ma X, *A hybrid image retrieval system with user's relevance feedback using neurocomputing*. *Informatica*, 2005. 29: 271-279.
12. Haralick RM, *Statistical and structural approaches to texture*. *Proceedings of the IEEE*, 1979. 67: 786-804.
13. Rahman MM, Desai BC, and Bhattacharya P. *Supervised machine learning based medical image annotation and retrieval in ImageCLEFmed 2005*. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science. 2005. Vienna, Austria: Springer-Verlag. in press.