Published in final edited form as:

Pattern Recognit Lett. 2008 November 1; 29(15): 2032–2038. doi:10.1016/j.patrec.2008.05.013.

Effectiveness of Global Features for Automatic Medical Image Classification and Retrieval – the experiences of OHSU at ImageCLEFmed

Jayashree Kalpathy-Cramer and William Hersh

Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR

Abstract

In 2006 and 2007, Oregon Health & Science University (OHSU) participated in the automatic image annotation task for medical images at ImageCLEF, an annual international benchmarking event that is part of the Cross Language Evaluation Forum (CLEF). The goal of the automatic annotation task was to classify 1000 test images based on the Image Retrieval in Medical Applications (IRMA) code, given a set of 10,000 training images. There were 116 distinct classes in 2006 and 2007. We evaluated the efficacy of a variety of primarily global features for this classification task. These included features based on histograms, gray level correlation matrices and the gist technique. A multitude of classifiers including k-nearest neighbors, two-level neural networks, support vector machines, and maximum likelihood classifiers were evaluated. Our official error rates for the 1000 test images were 26% in 2006 using the flat classification structure. The error count in 2007 was 67.8 using the hierarchical classification error computation based on the IRMA code in 2007. Confusion matrices as well as clustering experiments were used to identify visually similar classes. The use of the IRMA code did not help us in the classification task as the semantic hierarchy of the IRMA classes did not correspond well with the hierarchy based on clustering of image features that we used. Our most frequent misclassification errors were along the view axis. Subsequent experiments based on a two-stage classification system decreased our error rate to 19.8% for the 2006 dataset and our error count to 55.4 for the 2007 data.

1. Introduction

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 generated and stored in hospitals and medical systems in recent years. On-line atlases of images have been created for many medical domains including dermatology, radiology and gastroenterology [1,22,26,27]. As the number of images being generated and archived everyday increases, the ability to search and retrieve relevant images becomes a critical and challenging task. Medical and other image retrieval systems have historically relied on the annotation or captions associated with the images for indexing the retrieval system. The labor-intensive task of indexing and cataloging the images in these collections is typically performed manually, a process that can be subjective and prone to errors.

© 2008 Elsevier B.V. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Medical images typically store information including the imaging modality and anatomical location in the DICOM header [28]. This information can be useful for image retrieval when the query provides information about the desired imaging modality or anatomical location. However, this information is often lost when the images are compressed as JPEG images or stored in teaching and on-line collections. There have also been reported errors [8] that can occur in the DICOM headers.

Thus, the ability to automatically categorize and annotate medical images with information can be very useful for the purposes of image management and retrieval. Imaging modality, anatomical location, biological system and view are four possible dimensions along which medical images can be categorized [6,14,17]. Content-based image retrieval (CBIR), where the image itself is used in the query to find similar images, has been an area of interest to researchers in the last decade [23,25]. The algorithms developed in the field of computer vision and CBIR can also be applied to the task of image categorization and annotation. In recent years, researchers have been investigating methods to classify images based on visual appearance and annotate the image with the purported category [7,16]. This information can be used in addition to other textual annotations or captions to improve the performance of retrieval engines [11,12,13].

1.1 ImageCLEFmed and the automatic annotation task

The ImageCLEF medical image retrieval and automatic annotation tasks (ImageCLEFmed) [29] provide test collections for the medical image retrieval community to benchmark their algorithms. Since 2003, the ImageCLEF campaign has been a part of the Cross Language Evaluation Forum (CLEF) [2,10,30] which is an offshoot from the Text Retrieval Conference (TREC, trec.nist.gov).

In 2006, the goal of the automatic annotation task within ImageCLEFmed was to correctly classify 1000 radiographic medical images into 116 categories [18,29]. These anonymized images, selected from the image collection of the Department of Diagnostic Radiology, Aachen University of Technology (RWTH), Aachen, Germany, differed in the “modality, body orientation, body region, and biological system examined,” according to the track Web site [31]. The task organizers provided a set of 9,000 training images for which the class membership was provided.

In addition, a set of 1000 classified images 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.

In 2007, although the 116 classes remained the same, the goal was to correctly classify a similar set of 1000 radiographic medical images using the hierarchical IRMA code [5,6,19]. This code classifies the image along the modality, body orientation, body region, and biological system axes. 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 within the hierarchy. The task organizers provided a set of 10,000 training images and 1000 development images.

2. System Description

Our overall goal in the long term is to annotate all images in our medical image retrieval system [12,13] with the anatomical location and view of the images. The ImageCLEF automatic annotation task offered us the opportunity to create a fairly simple but extensible automatic image annotation system as it provided us with a large training dataset of well-annotated

images. Since the images in our image retrieval system are extremely varied in imaging modality, anatomical location and quality, we were primarily interested in exploring the ability of low level, global features for automatic annotation and categorization.

We followed the steps described below in creating the classifier for the ImageCLEF med automatic annotation task using MATLAB [32] as our primary development environment:

1. A variety of feature vectors were created for all images in the training, development and test images.
2. A 116-class classifier was developed using primarily open source toolboxes.
3. The classifier was optimized using the development set of images.
4. This optimal classifier was applied to the set of test images.
5. In 2007, the class code was converted to the IRMA code required for submission.

2.1 Features vectors

The images provided were resized such that at least one dimension was 512 pixels. We further padded the images to create a 512×512 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 could provide useful information for classification. For instance, the aspect ratios of most images of arm and leg bones are significantly different than the aspect ratio of lungs. We chose to pad the images instead of either extracting features directly from the resized image or from images resized to square dimensions as early experiments had shown that maintaining an indirect measure of the aspect ratio helped in the classification.

All images were then resized to 256×256 pixels using the bilinear interpolation algorithm provided in the MATLAB image processing toolbox [32].

For our first efforts in the medical image automatic annotation domain in 2006, 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 by tiling the image and creating feature vectors for the overlapping tiles.

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 16×16 pixel 'icon' of the image was created by resizing the image using bilinear interpolation. This resulted in a 256-dimensional feature vector.
- GLCM: Four gray level co-occurrence matrices (GLCM) [9] matrices with offsets of 1 pixel, 0, 45, 90 and 135 degrees were created for the image after rescaling the image to 16 levels of intensity. 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 (256×256) was partitioned into 5 squares of size 128×128 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 create

feature vector of dimension 100. The difference between the padded and unpadded images is most apparent in this feature vector.

- Hist: A 32-bin intensity histogram was created for each image and counts in each bin were used as a 32-dimensional feature vector.
- DCT: A global discrete cosine transform was created for each image. The upper left (10×10) vectors were concatenated and used as a 100-dimensional feature vector.

Concatenating one or more of the above vectors created additional feature vectors.

In 2007, we primarily evaluated the ‘gist’ features developed by Oliva and Torralba [21]. All images were convolved with a set of 32 multiscale-oriented Gabor filters. We created a 512-dimensional vector using statistics from these filters. The use of Principal Component Analysis (PCA) was evaluated as a means of dimension reduction. The 512-dimensional vector as well as the reduced dimension vectors were evaluated using an assortment of classifiers.

2.2 Classifiers

We evaluated a variety of classifiers based on primarily open source toolboxes. A neural network classifier based on the Netlab toolbox [20] was used to create the official runs in 2006 and 2007. We used a multilayer perceptron (MLP) architecture to create the 116-class classifier. It had 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, primarily the number of hidden nodes and iteration used for training, was optimized using the training and development sets provided. In 2007, in addition to the neural network based classifier, we evaluated k-nearest neighbors using Euclidean distances and maximum likelihood classifiers using the Classification toolbox as well as the MATLAB statistics toolbox [24,32].

Subsequent to the official runs, we experimented with a two-stage classifier. The first stage consisted of the neural network as described above. Image classes that were visually most similar based on the confusion matrix of the development set were combined and a second stage classifier was applied to them. We created a second layer of classification built around a support vector machine (SVM) using scale-invariant feature transform (SIFT) features [15] as inputs. This new binary classifier was used to determine the final class assignments for images in the two most frequently misclassified.

2.3 Clustering algorithms

We evaluated the visual similarity between image classes by performing clustering experiments on the dataset. Clustering techniques attempt to partition the data into groups or clusters such that images within a group are similar but images in different clusters are different. Mean-shift, hierarchical and k-means clustering techniques were used to evaluate correspondence between hierarchies based on visual features with the IRMA hierarchy which is based on semantic or conceptual categories.

K-means clustering is a popular, unsupervised algorithm that clusters data into the specified number (k) of partitions such that the sum of the squared error between the data points and their cluster centers is minimized.

The mean-shift algorithm [3,4] is a hill-climbing algorithm that can be used for nonparametric clustering. One advantage of mean-shift clustering over k-means clustering is that the number of clusters does not need to be defined a-priori. It has been shown that “mean-shift converges from almost any starting point and that, in general, its convergence is of linear order” [3].

Hierarchical clustering is another unsupervised clustering algorithm where the data is grouped “over a variety of scales by creating a cluster tree or dendrogram”[32]. In this case, the number of clusters does not necessarily need to be specified a-priori but one can investigate the cluster membership over a variety of scales or number of clusters. The height of the tree can be used to determine natural partitions of our image classes.

3. Experiments Performed

3.1 Official Runs

Our primary classifier was a neural network in 2006 and 2007. To avoid over-fitting, we optimized the network architecture heuristically using the number of hidden nodes as well as number of iterations as variables. 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 close to 100% classification accuracy on the training set. However, there were issues with over-fitting 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 in 2006 and 2007.

Figure 1 presents the classification accuracy on the development set in 2006 for two different feature sets as a function of the number of hidden node. In 2006, to optimize performance on the development set, we chose a network consisting of 200 hidden nodes for the iconHist features while we chose 300 hidden nodes for vectors consisting of iconGLCM features.

Similarly, figure 2 shows the impact of iterations during the training in 2007 for a network consisting of gist features that were dimension-reduced to 80 using PCA. Too many iterations can lead to overfitting where the classification accuracy on the training set continues to improve with more iterations, but the accuracy on an unseen development set drops beyond a certain maxima. We chose the number of iterations that resulted in the best performance on the development set, which for this set of features was 500.

We submitted four runs as our official runs for the automatic annotation task in 2006; iconGLCM2 using just the training set for creating the net, iconGLCM2 using the development and training set for creating the net, iconHist, and iconHistGLCM.

We submitted just two runs for in 2007. The first run used ‘gist’ feature vectors to train the multi-layer perceptron with one hidden layer containing 250–500 nodes. A neural network was used to create a multi-class classifier consisting of 116 classes. The final configuration of the classifier used 300 hidden nodes. We used 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.

PCA was used for dimension reduction. As seen in figure 3 below, the classification accuracy on the development set increased significantly with the number of dimensions up to about 100. However, the improvement in accuracy using additional dimensions was limited. We had used an 80-dimensional feature vector for our official run.

3.2 Post-workshop runs

We performed additional experiments after our official submissions in 2006 and 2007. We expanded our work with the two stage classifier, evaluated other classifiers, performed visual clustering experiments to identify visually similar classes, and tried to establish correspondence between the IRMA code and visual similarity using our features.

4. Results and Analysis

4.1 Official Runs

The best results for the development set in 2006 were obtained using a 356-dimensional normalized input vector consisting of the icon (16×16) concatenated with the GLCM2. The classification rate on the development set was 80%.

The next best result was obtained using a 288-dimensional normalized input vector consisting of the icon (16×16) 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 1. However, the results obtained on the test set were lower than those of the development set.

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 2). Class 108 saw the most significant differences, 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.

The relationship between semantic and visual hierarchy remains an open area of research. In 2007, the ImageCLEF error counting was based on the hierarchical IRMA code, with greater penalty for misclassification than for less precise classification. However, based on our experiments on this collection, the use of the IRMA hierarchy for the image classes did not improve the quality of our automatic annotations.

Classes 108 and 111 were most frequently interchanged by our classification methods. To try to improve the performance, we created a second layer of classifier for these two classes. Images classified as 108 and 111 from the first stage were combined and reclassified using a two-class SVM classifier. SIFT features were used instead of the gist features. However, the error count for both our runs was 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 correspond to a 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.

Figure 4 displays the official results for the error count for all 68 runs. Our runs have an error count that is about three times as high as the best group and about one seventh the error count of the worst run. The best groups used more local features in addition to purely global features. Many groups that tried to use the hierarchical IRMA code performed poorly [19].

4.2 Post Workshop Experiments

4.2.1 Two stage classifier in 2006—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 using the iconGLCM2 features and a neural network for the second stage. We also evaluated the use of SIFT features and an SVM for the

second stage of the hierarchical classifier. We saw a small further improvement to 80.2% using the gist features. This seems like a promising approach to improve the classification ability of our system.

4.2.2 Other classifiers in 2007—We compared the results using the MLP classifier with those obtained using a simple K-nearest neighbor classifier using Euclidean distances as well as a Maximum Likelihood classifier [24]. Using the flat classification scheme of 116 classes, our results are shown in table 3.

As seen in table 3 above, although there is a significant improvement with the use of neural network architecture, even a simple 1-nearest neighbor classifier can provide sufficient accuracy to be used in the context of an image retrieval system. There is no architecture optimization required for the 1-NN classifier unlike the neural network where the number of hidden nodes as well as the number of iterations needs to be optimized empirically.

In addition, the actual time required for classification significantly slower for the neural net depending on the number of hidden nodes and iterations, as shown in table 4. The times in the table below were for training on 10,000 images and testing on 1,000 images on a MacBook with 2GB of RAM running OSX (10.5).

However, the best results, using the features describe above without any feature selection or optimizations, were found with the two-stage classifier described above. In 2007, using the two-stage classifier, we reduced our error count from 67.8 to 55.4

4.2.3 Misclassification Analysis—We further evaluated those classes that had four or more images misclassified, especially with those that were confused with a single other class. As seen in table 5, classes 2,3 and 56 are often confused. Classes 108, 109 and 111 are also confused. These results correlate well with the results of the clustering experiments, as seen in the next section.

4.2.4 Clustering experiments—We performed hierarchical and mean shift clustering images using an “average” feature vector per class. These were then related to the misclassified classes. There were 4 main clusters of visually similar images that had sufficient images per class to perform further analysis. Classes 2, 3, 4 and 56 were visually similar images of hands/fingers. Classes 13, 14 and 15 were images of fingers and bones. Classes 44 and 45 were pictures of knees while classes 108,109 and 111 were chest images.

4.2.5 Misclassification and IRMA code—Table 6 lists the most commonly misclassified classes and their IRMA codes. These classes are listed by clusters. For instance, classes 2, 3 and 56 frequently occur as a single cluster, as do classes 14 and 15, classes 44 and 45, and classes 108, 109, 111.

From our experiments, the main cause for the misclassification was due to global nature of our features, such that the distances between the features vectors of images with spatial symmetry were small. Hence, the view axis caused us the most problems. We were unable to distinguish between the posteroanterior and anteroposterior views of chest x-rays or coronal and sagittal views of fingers. Similarly, right and left knees were similar for our classifiers.

4.2.5 Challenges—One of the issues with the database is that the distribution of the number of training images in each of the classes is quite varied. A histogram of the class sizes can be seen in figure 5.

Of the 10,000 images in the training set and the 1000 images in the development set, just three of the 116 classes together contain about 40% of the total images in the training set while 36 classes have less than 20 images. Although, this can be typical of collections in the real world, this variation in class membership size leads to difficulties for the classifiers. Another issue is that other classes that have lot of intra-class variation while other classes have significant inter class similarity.

A consequence of hierarchical nature of the codes and the penalty of misclassification is that the cost of misclassification is not constant. If the flat classification is used, as in 2006, the penalty of misclassification is 1 irrespective of the nature of the misclassification. However, for the hierarchical classification computation, the cost depends on the distance in the tree between the correct class and the estimated class [5]. The penalty for errors between classes 108 (1123-110-500-000) and 111 (1123-127-500-000) is due to error in the second digit of the view axis. However, errors between classes 2 (1121-110-411-700) and 56 (1121-200-411-700) occur in the first digit of the view axis and are more costly.

5. Conclusions and Future Work

Global image features can be used to classify medical radiographic images with moderate success at speeds to enable real-time applications. The image features as well as the classifiers tested can significantly affect the performance of the classification. In 2006, our best results were obtained by using a feature vector consisting of a 16×16 icon and grey-level co-occurrence features. A multi-layer perceptron architecture was used for the neural network. Our classification accuracy was 74%. In 2007, we obtained an error rate of 67 out of 100 images using the hierarchical error computation algorithm or an equivalent 77% accuracy using the 2006 computation. Both times, we improved our performance during the post-workshop experiments by creating a two-stage classifier. In 2006, we had created a two-stage classifier for classes 108/111 and classes 2/56. Recent experiments using a two-stage classifier for all classes are showing promise. Our best results so far have been with a two-layer classifier using gist features.

We need to address the issue of symmetry in creating the feature vectors to be able to distinguish between images that are close to mirror images. However, we believe that local features, including SIFT or saliency points and their distribution, may provide additional performance improvements. We would like to explore them in future work.

We would like to further investigate the mapping between the semantic and visual hierarchy of images in the IRMA collection. We primarily used a flat classifier in this work. Our classifiers used a constant cost for all classes and misclassifications. However, it might be possible to improve the performance using the IRMA hierarchy by the use of a cost function that depends on the hierarchy of the IRMA classes.

We have demonstrated the utility of automatic annotation of image modality for medical image retrieval at the ImageCLEF med retrieval task [12,13]. We are currently expanding our system to include automatic image annotation with anatomy and view, based on the work that was presented in this paper. We plan on annotating the entire collection of 66,000 images with the anatomy and view, in addition to the modality. Preliminary results indicate an improvement in performance using this multi-modal approach to retrieval.

Acknowledgments

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

References

1. Aisen AM, Broderick LS, et al. Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment. *Radiology* 2003;228:265–270. [PubMed: 12832587]
2. Braschler M, Peters C. Cross-language evaluation forum: objectives, results, achievements. *Inform Retrieval* 2004;7:7–31.
3. Carreira-Perpinan MA. Acceleration Strategies for Gaussian Mean-Shift Image Segmentation. *CVPR* 2006:1160–1167.
4. Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2002;24(5):603.
5. Deselaers T, Kalpathy-Cramer J, Müller H, Deserno TM. Hierarchical classification for ImageCLEF 2007 medical image annotation. 2007
6. Deselaers T, Müller H, Clough P, Ney H, Lehmann TM. The CLEF 2005 automatic medical image annotation task. *Int. J. Computer Vision* 2007 August;74(1):51–58.
7. Florea F, Müller H, Rogozan A, Geissbühler A, Darmoni S. Medical image categorization with MedIC and MedGIFT. *Medical Informatics Europe (MIE)*. 2006
8. Güld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM. Quality of DICOM header information for image categorization. *Proceedings SPIE* 2002;4685:208–287.
9. Haralick RM, Shanmugan K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* 1973;Vol. SMC-3:610–621.
10. Hersh W, Muller H, et al. Advancing biomedical image retrieval: development and analysis of a test collection. *J. Am. Med. Inform. Assoc* 2006;13(5):488–496. [PubMed: 16799124]
11. Hersh, W.; Kalpathy-Cramer, J.; Jensen, J. Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006. In: Peters, C.; Clough, P.; Gey, F.; Karlgren, J.; Magnini, B.; Oard, DW.; de Rijke, M.; Stempfhuber, M., editors. *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Springer Lecture Notes in Computer Science. Vol. Vol. 4730. p. 660-669.
12. Hersh W, Kalpathy-Cramer J, et al. Medical image retrieval and automated annotation: OHSU at ImageCLEF 2007. *Springer Lecture Notes in Computer Science (LNCS)*. (to be published)
13. Kalpathy-Cramer, J.; Hersh, W. *MedInfo 2007*. Brisbane, Australia: 2007. Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval.
14. Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. *Proceedings SPIE* 2003;5033:109–117.
15. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision* 2004;60(2):91–110.
16. 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* 2004;73:1–23. [PubMed: 15036075]
17. Müller, H.; Clough, P., et al. Evaluation Axes for Medical Image Retrieval Systems - The ImageCLEF Experience; *ACM Int. Conf. on Multimedia*; Singapore. 2005 Nov.
18. Müller, H.; Deselaers, T.; Lehmann; Clough, P.; Kim, E.; Hersh, W. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Peters, C.; Clough, P.; Gey, F.; Karlgren, J.; Magnini, B.; Oard, DW.; de Rijke, M.; Stempfhuber, M., editors. *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Springer Lecture Notes in Computer Science. Vol. Vol. 4730. p. 595-608.
19. Müller, H.; Deselaers, T.; Kim, E.; Kalpathy-Cramer, J.; Deserno, TM.; Clough, P.; Hersh, W. *Springer Lecture Notes in Computer Science (LNCS)*. Budapest, Hungary: 2008. Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. to appear
20. Nabney, IT. *Netlab: Algorithms for Pattern Recognition*. London, England: Springer-Verlag; 2004.
21. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Computer Vision* 2001;42(3):145–175.
22. Schmid-Saugeon P, Guillod J, et al. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics* 2003;27:65–78. [PubMed: 12573891]

23. 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* 2000;22(12):1349–1380.
24. Stork, DG.; Yom-Tov, E. *Computer Manual in MATLAB to Accompany Pattern Classification*. Vol. 2nd Edition. Wiley; 2004 [accessed 3.20.2008]. <http://www.yom-tov.info/toolbox.html>
25. Tagare HD, Jaffe C, et al. Medical Image Databases: A Content-Based Retrieval Approach. *J. Am. Med. Inform. Assoc. (JAMIA)* 1997;4(3):184–198.
26. [accessed 12.20.2007]. <http://mypacs.net>
27. [accessed 12.20.2007]. <http://www.visualdx.com>
28. [accessed 12.20.2007]. <http://medical.nema.org/>
29. [accessed 12.20.2007]. <http://www.imageclef.org>
30. [accessed 12.20.2007]. <http://www.clef-campaign.org>
31. [accessed 12.20.2007].
<http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/mediclaat.html>
32. [accessed 12.20.2007]. <http://www.mathworks.com>

Figure 1.
Number of development images classified correctly as a function of number of hidden nodes of MLP

Figure 2.
Number of development images classified correctly as a function of number of iterations of MLP

Figure 3.
Classification accuracy as a function of the number of PCA features used

Figure 4.
Official results for 2007

Figure 5.
Histogram of class sizes

Table 1

Classification rates for OHSU automatic annotation runs

Feature vector	Classification rate (%)	
	Development	Test
DCT ¹	71	
Hist ¹	72	
Icon ¹	74	
GLCM ¹	75	
GLCM2 ¹	76	
iconDCT ¹	75	
iconHist ² (official run)	78	69
iconGLCM (pre workshop)	78	
iconGLCMHist ²	78	72
iconGLCM ²	80	74
Gist ³		77

¹ pre workshop² official run³ post workshop

Table 2

Differences for selected classes between development and test set of images

Class	Number of images in class		Diff. in error count	
	Development set count	Test set count	# correct	
108	93	78	92	54
61	21	21	20	16
44	10	7	10	2
12	23	21	22	16

Table 3

Results using gist features and k-nearest neighbor classifier

Feature Size (PCA)	MLP	1-NN	5-NN	ML
20	68.0%	67.7%	66.3%	60.9%
40	74.0%	71.1%	69.7%	68.2%
60	76.4%	71.4%	70.9%	69.1%
80	76.5%	72.4%	70.4%	70.4%
100	77.5%	70.3%	70.3%	-
120	77.4%	73.0%	71.0%	-
200	77.3%	72.8%	71.0%	-
512 (no PCA)	78.1%	72.8%	71.0%	-

Table 4

Classification times in MATLAB

	Classification times (s)		
nodes	MLP	1-NN	5-NN
20	3000	21	22
80	3321	59	60
120	3537	81	85
200	4314	126	128
512	6197	290	304

Table 5

Commonly misclassified categories

Class	Class size	Accuracy	Number misclassified	Confused class#	# in confused class
2	14	0.71	4	56	3
3	36	0.83	6	56	3
15	4	0	4	14	1
44	10	0.6	4	45	2
56	13	0.46	7	2	4
108	92	0.58	39	111	37
109	5	0.2	4	108,111	2 each
111	195	0.92	15	108	12

Table 6

IRMA code for commonly misclassified categories

Class	IRMA code	Technique	Direction	Anatomy	Bio System
2	1121-110-411-700	krav, plain radiograph, analog, overview	coronal, posteroanterior (PA), unspecified	Upper extremity, hand, finger	musculoskeletal system, unspecified
3	1121-110-414-700	krav, plain radiograph, analog, overview	coronal, posteroanterior (PA), unspecified	Upper extremity, hand, left hand	musculoskeletal system, unspecified
56	1121-200-411-700	krav, plain radiograph, analog, overview	sagittal, unspecified	Upper extremity, hand, finger	musculoskeletal system, unspecified
14	1121-120-413-700	krav, plain radiograph, analog, overview	coronal, anteroposterior (AP), unspecified	Upper extremity, hand, finger	musculoskeletal system, unspecified
15	1121-120-421-700	krav, plain radiograph, analog, overview	coronal, anteroposterior (AP), unspecified	Upper extremity, hand, carpal bones	musculoskeletal system, unspecified
44	1121-120-942-700	krav, plain radiograph, analog, overview	coronal, anteroposterior (AP), unspecified	left knee	musculoskeletal system, unspecified
45	1121-120-943-700	krav, plain radiograph, analog, overview	coronal, anteroposterior (AP), unspecified	right knee	musculoskeletal system, unspecified
108	1123-110-500-000	krav, plain radiograph, analog, high beam	coronal, posteroanterior (PA), unspecified	chest, unspecified	unspecified
109	1123-112-500-000	krav, plain radiograph, analog, high beam	coronal, posteroanterior (PA), expiration	chest, unspecified	unspecified
111	1123-127-500-000	krav, plain radiograph, analog, high beam	coronal, anteroposterior (AP), supine	chest, unspecified	unspecified