MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval

Jayashree Kalpathy-Cramer[a], William Hersh[a]

[a] *Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon, USA*

Abstract

Medical image retrieval can play an important role for diagnostic and teaching purposes in medicine. Image modality is an important visual characteristic that can be used to improve retrieval performance. Many test and on-line collections do not contain information about the image modality. We have created an automatic image classifier for both grey-scale and colour medical images. We evaluated the performance of the two modality classifiers, one for grey-scale images and the other for colour images on the CISMeF and the ImageCLEFmed 2006 databases. Both classifiers were created using a neural network architecture for learning. Low level colour and texture based feature vectors were extracted to train the network. Both classifiers achieved an accuracy of > 95% on the test collections that they were tested on. We also evaluated the performance of these classifiers on a selection of queries from the ImageCLEFmed 2006. The precision of the results was improved by using the modality classifier to re-sort the results of a textual query.

Keywords:

medical imaging, neural networks, image annotation, content-based image retrieval

Introduction

Medical images form a vital component of a patient's health record. Effective medical image retrieval systems can play an important role in aiding in diagnosis and treatment; they can also be effective in the education domain for healthcare students, instructors and patients alike. As a result of advances in digital imaging technologies, there has been a large growth in the number of digital images stored in recent years. In addition to the Picture Archival and Communication Systems (PACS) that are becoming omnipresent in hospital and clinics, there are numerous on-line collections of medical images. On-line atlases of images can be found for many medical domains including dermatology, radiology and gastroenterology. The sheer volume of medical image data provides for numerous challenges and opportunities in the arena of medical image retrieval.

Historically, the task of indexing and cataloging these collections has been performed manually. This is an arduous and painstaking task, and is prone to errors. Consequently, there is a desire to be able to automate the task of indexing these collections with a goal to improve the ability to search and retrieve relevant documents.

Medical image retrieval systems have traditionally been text- based, relying on the annotation or captions associated with the images as the input to the retrieval system. The last few decades have offered advancements in the area of content-based image retrieval (CBIR) [1]. CBIR systems have had some success in fairly constrained medical domains, including pathology, head MRIs, lung CTs, and mammograms [2]. However, purely content-based image retrieval systems currently have limitations in more general medical image retrieval situations, especially when the query includes information about pathology [3, 4]. Mixed systems (using both textual and visual techniques) have demonstrated improved retrieval performance, especially with regards to precision at the top of the list [4].

Medical image databases used for image retrieval or for teaching purposes often contain images of many different modalities, taken under varied conditions with variable accuracy of annotation. This can be true for images found in various on-line resources, including those that access the on-line content of journals[1].

Image modality is an important, fundamental visual characteristic of an image that can be used to aid in the retrieval process. However, the annotations or captions associated with images often do not capture information about the modality. Images that may have had modality 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. There have also been reported errors in the accuracy of DICOM headings [5].

The medical image retrieval task within ImageCLEF has provided both a forum as well as test collections to benchmark image retrieval techniques. The ImageCLEF campaign has been a part of the Cross Language Evaluation Forum since 2003 [3]. CLEF itself is an offshoot from the Text REtrieval Conference (TREC). In 2004, ImageCLEFmed, a domain–specific task, was added to evaluate medical image retrieval algorithms and techniques.

1 http://gm.arrs.org/ (accessed 3/26/2007)

Approaches combining both visual and textual techniques for retrieval have shown some promise at medical image retrieval tasks [3]. In 2005, a medical image annotation task was added to ImageCLEF. The goal of this task was to correctly classify 1000 test images into 116 classes given a set of 10,000 training images. The classes differed primarily in anatomy and view of the image. It should be noted, however, that these images were primarily of a single modality (X-rays).The goal of the ImageCLEF medical image retrieval task of 2006 was to retrieve relevant images for thirty topics from a test collection of about 50,000 annotated images of different modalities. These tasks were divided by the organizers into those expected to be amenable to textual, visual, or mixed retrieval techniques.

We participated in both the medical image retrieval and the automatic medical image annotation tasks at ImageCLEF 2006 [6, 7]. The techniques developed for those tasks have been extended for the more general task of medical image modality classification and annotation.

Using medical image modality for image annotation and retrieval has recently been studied. Florea et al [8] have compared the efficacy of two different systems (MedIC and MedGIFT) in classifying the modality of a database with six standard modalities for radiology and nuclear medicine images.

In this paper, we compare the results obtained on our system with those described in previous publications [8] for the six modalities of the CISMeF database. We will also extend this technique to classify colour images from the ImageCLEF medical retrieval task collection [6] into six categories. We will finally report on the improvement in precision that we observed for a selected number of tasks of the ImageCLEF medical retrieval task for 2006 by incorporating the modality classifier in series with a text-based retrieval system.

Methods

We employed a supervised machine learning approach to problem of medical image modality classification using a hierarchical classification scheme as seen in figure 1. There were two primary databases that were used to create and test the classifiers. We worked with a small subset of the CISMeF database as the primary target for our grey-scale (radiographic and nuclear medicine) image classifier [9]. This database had a set of 1332 images classified into one of six classes based on modality. These include angiography, computerized tomography scans (CT), X-ray, Magnetic resonance (MRI), ultrasound, and scintigraphy. The images in this database had been acquired under differing conditions over a long period of time. Consequently, there was considerable intra-class variation in quality, size, contrast, illumination and background.

The imageCLEFmed database contains 50,000 images of differing modalities, including radiography and nuclear medicine, as well as microscopic and histopathological images, photographs and gross pathology images, power point slides, electroencephalograhical images (EEGs) and electrocardiograms (ECGs), as well as a few miscellaneous images.

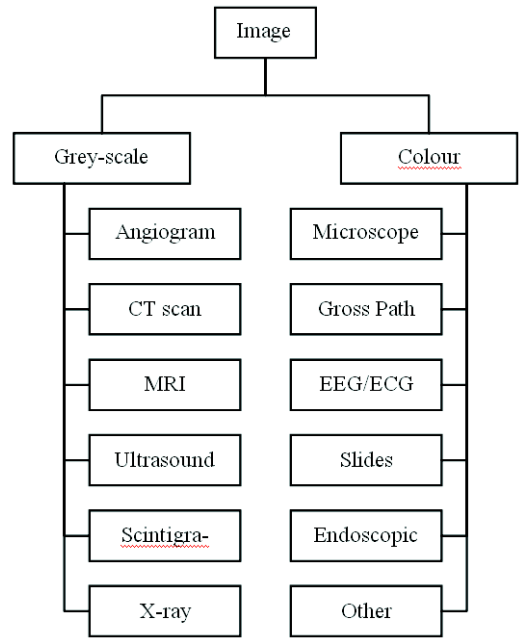


Figure 1 - Hierarchical classification scheme for images

A neural network-based scheme using a variety of low level, primarily global image features was used to create a six-class classification system for the grayscale images. The multilayer perceptron architecture used a hidden layer of approximately 50-150 nodes. The classification system was created in MATLAB[2], in part using several routines modified from the Netlab toolbox[3].

We experimented with a variety of feature vectors as inputs to the network. A combination of texture and intensity histogram features provided the best classification [10, 11]. All images were first resized while maintaining the aspect ratio such that the smaller dimension was 256 pixels. The image was divided into five overlapping blocks. Grey level correlation matrices were computed for each block using four angles and an offset of 1 pixel. Contrast, correlation, energy, homogeneity and entropy were calculated for each matrix. A quantized grey scale histogram was then appended resulting in a 132-dimension feature vector for each image for the texture. 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 1332 images in the database were randomly split into a training set of 1000 images and a test set of 332 images. A small random subset of the training images was initially used to create the classifier (200 images). The classifier

2 www.mathworks.com (accessed 3/26/2007)
3 http://www.ncrg.aston.ac.uk/netlab/index.php (accessed 3/26/2007)

was then applied to the entire training set and images that were misclassified were then added to the images used to refine the classifier. The classifier was finally tested on the test images.

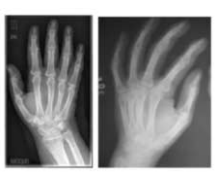
A similar scheme was used to create the classifier for colour images. We believe this novel idea can improve the retrieval performance of purely textual systems or for images for which the associated modalities are not known. Although modality detectors for grey-scale medical images have been reported [9], we are unaware of a similar effort for classification of other categories of medical images like those produced by histo-pathology and endoscopy. The images used for this classification task were taken from the test collection of images used in the ImageCLEFmed retrieval task. 2250 colour images in this collection were broadly categorized into six categories as microscopic, gross pathology, EEG/ECG or other charts, powerpoint slides, endoscopic images and other. There was considerable intra-class variability in this dataset. These 2250 images were again split randomly into training (1750) and test images (500). A similar training methodology to that described above was used to incrementally improve the classifier, starting with a smaller subset of the training database.

A two-layer architecture with 25-150 hidden nodes was used for the neural network. The feature vector in this case consisted of colour histogram features, as well as texture features obtained using the grey level correlation matrices. The image was split into 9 uneven blocks. Colour histogram properties of image after conversion into the L*A*B* colour space were calculated, while texture features were calculated after converting the image to grey-scale

These neural network classifiers can be created to further classify images within a given modality. For instance, x-ray images could now be classified to account for anatomy. Anatomical classifiers were used in the automatic annotation task at ImageCLEFmed.

The tasks had been stated in English, German and French, and had and provided example images. All but three of the tasks stated the desired modality of the image to be retrieved. Two examples of the tasks are shown in figure 2.

Show me images of a hand x-ray.
Zeige mir Röntgenbilder einer Hand.
Montre-moi des radiographies de la main.



Show me blood smears that include polymorphonuclear neutrophils. Zeige mir Blutabstriche mit polymorphonuklearer Neutrophils. Montre-moi des échantillons de sang incluant des neutrophiles polymorphonucléaires.

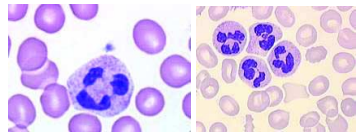


Figure 2 - Sample textual and visual queries at ImageCLEFmed 2006

Once our classifiers had been trained to achieve >95% classification accuracy, they were tested on a random subset of the ImageCLEFmed topics.

The schematic of our modified retrieval system is shown below. The query was initially fed to our Lucene[4] based text retrieval system. The queries were manually edited by one of the authors. The resulting images were subsequently classified by the hierarchical classifier for modality. Images of the desired modality (as stated in the query or as discerned by the automatic classifier based on the sample images) were moved to the top of the list while maintaining the ranking of the textual system within a class.

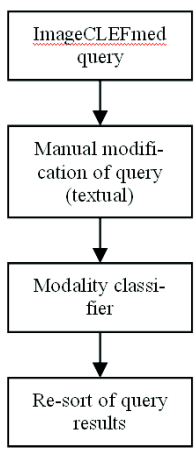


Figure 3 - Image retrieval system used for the ImageCLEFmed 2006 test collection

We compared the results of our purely textual system with that including the addition of the modality classifier.

Results

A classification accuracy of 96.4% was achieved on the CISMeF database. The confusion matrix suggests that the primary misclassification occur between the MRI and CT scan classes. This is not surprising as these classes are visually quite similar. Florea et al [8] have reported similar results both in terms of accuracy and inter-class misclassification patterns. The classification of grey-scale medical images into commonly occurring modalities using low level image features and machine learning techniques appears to be a tractable task. We expect to achieve over 98% accuracy with further refinement of our machine

4 http://lucene.apache.org/ (accessed 3/26/2007)

learning approach by the use of more advanced cross-validation, bootstrapping, boosting or bagging techniques.

Preliminary testing of the classifiers on 2250 colour images of the imageCLEFmed test collection resulted in a modality classification accuracy of 98.6%. Most of the misclassifications involved the "other" class with contained a set of miscellaneous images not belonging to the other five specific categories

The colour modality classifier was tested on a small random subset of the ImageCLEFmed 2006 topics. The topics for imageCLEFmed 2006 fell into three categories (visual, mixed, semantic) consisting of 10 tasks each. Although visual techniques had, in general, performed extremely poorly at the semantic tasks, use of some visual information (primarily modality) was shown to increase the precision [4].

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. A compelling example is given in figure 4 and table 1. Only 90 of the top 2000 images returned by the textual query were of the desired modality.

Task 1 - Show me images of the oral cavity including teeth and gum tissue

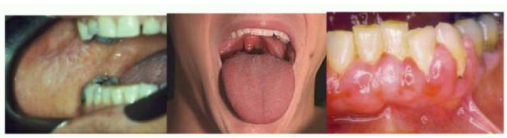


Image type	Number of images
Total returned by textual query	2000
Grey-scale	1831
Photograph/gross pathology	90
Microscope	71
Other	8

Figure 4- Sample query suitable for visual retrieval at ImageCLEFmed 2006

These images were then classified using our modality classifier. The ranked list of retrieved images was resorted taking into account the desired modality based on the query.

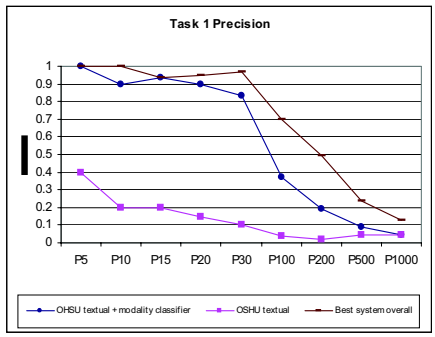


Figure 5 - Improvement in precision resulting from modality classification.

Figure 5 plots the precision for varying number of documents retrieved for the purely textual system, the improvement with the use of the modality classifier and the overall best system (mixed visual and textual based) that participated in ImageCLEFmed 2006. This increased the precision of the query as seen in figure 5. The improvement in precision at the top of the ranked list (P5 – P200) is better with the use of the modality detector compared to a purely textual search. We should note that a perfect modality classifier will only improve the precision of the search and not the recall if it is applied in the serial manner described above. The mean average precision (MAP) would still be limited by the number of relevant images that are retrieved by the textual search (recall of the textual search).

Even in searches that are expected to be semantic, we see an improvement in precision by using the modality classifier as seen in figure 6 and 7.

Task 2 - Show me microscopic images of tissue from the cerebellum (semantic query)

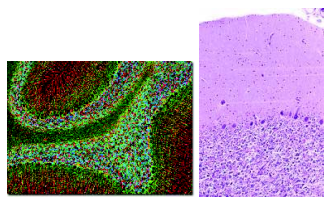


Image type	Number of images
Total returned by textual query	2000
Greyscale	1476
Photograph/gross pathology	408
Microscope	116

Figure 6 - Sample query suitable for visual retrieval at ImageCLEFmed 2006

The precision of this search was similarly improved by the use of the modality detector as seen in figure 7.

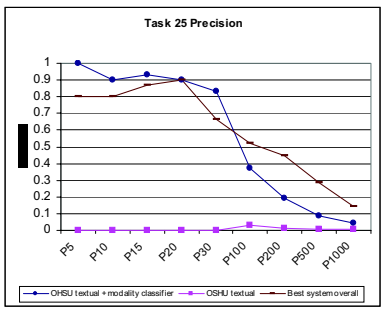


Figure 7 - Improvement in precision resulting from modality classification

Four of the six tasks tested showed improvement in precision by the use of the modality detector for colour images. There were two tasks amenable to textual methods for which there very little change in precision with the addition of the modality information.

We plan on testing the performance of the modality detector on the complete set of tasks for ImageCLEFmed 2005 and 2006. We also intend to index the entire collection of 50,000 images used in the ImageCLEFmed test collection using the modality classifier. Information about the class membership of an image will be added to the metadata. This should improve the performance of the retrieval in two ways. Clustering of the data by modality and perhaps anatomy will speed up the search process as fewer documents will have to be compared to the query image/text. Secondly, we expect that the overall precision of the search will improve by considering the modality of the image that is desired by the user. However, we can expect a small degradation in the recall due to potentially misclassified images not being searched.

Conclusion

We have developed a neural network based, hierarchical classifier for the modality classification of medical images. This system can classify colour images including histo-pathological and endoscopic images, and photographs as well as grey-scale (radiological and nuclear medicine). The classifier uses a histogram and texture properties as inputs to the two level neural network. This classifier results in a classification accuracy of greater than 95% for the grey-scale images of the CISMeF database as well as a selection of colour and grey-scale images from the ImageCLEFmed database. The use of this classifier increases the precision of retrieval of our primarily text based retrieval system by moving images of the desired modality to the top of the ranked list.

Acknowledgments

The authors would like to thank Henning Müller for his valuable input in discussions as well as providing access to the CISMeF images. We would also like to thank Jean-Nicholas Dacher at the Rouen Hospital, France, for proving the CISMeF images used in this paper. We would like to thank Dr. T. Lehmann at RWTH Aachen University, Germany, for access to the IRMA image database. We acknowledge the support of NLM Training grant 5T15 LM07088-15 and NSF Grant ITR-0325160.

References

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(12), 1349-80.

[2] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics 2004; **73**, 1-23.

[3] 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, editors: Peters C, Clough P, Gey F, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M, LNCS 2006,Alicante, Spain, to appear.

[4] Hersh W, Kalpathy-Cramer J, Jensen, J. Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006. Working Notes for the CLEF 2006 Workshop, Alicante, Spain. http://www.clef-campaign.org/2006/

[5] Güld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM, Quality of DICOM header information for image categorization, Proceedings SPIE, 4685, 280-287, 2002.

[6] Hersh W, Müller H, Jensen J, Yang J, Gorman P, Ruch P. Advancing biomedical image retrieval: development and analysis of a test collection. J Amer Med Inform Assoc 2006; 13(5).

[7] Lehmann TM, Güld MO, Thies C, Fischer B, Keysers D, Kohnen M, Schubert H, Wein BB. Content-based image retrieval in medical applications for picture archiving and communication systems, in Medical Imaging, SPIE Proceedings, 5033, 440-451, San Diego, California, 2003.

[8] Florea F, Müller H, Rogozan A, Geissbühler A, Darmoni S, Medical image categorization with MedIC andMedGIFT. Medical Informatics Europe (MIE 2006).

[9] Douyere M, Soualmia LF, Neveol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004; 21(4):253-61.

[10] Haralick R. Statistical and structural approaches to texture. Proceedings of the IEEE 67**,** 786-804, 1979.

[11] Howarth P, Ruger S. Evaluation of texture features for content-based imageretrieval. In: Proceedings of the International Conference on Image and Video Retrieval, Springer-Verlag, pp. 326-324, 2004.

Address for correspondence

Jayashree Kalpathy-Cramer
5th Floor, Biomedical Information Communication Center
3181 S.W. Sam Jackson Park Rd.
Portland, Oregon 97239-3098

Email: kalpathy@ohsu.edu