Medical image categorization with MedIC and MedGIFT

Filip Florea[a,c,1], Henning Müller[b], Alexandrina Rogozan[a,c], Antoine Geissbuhler[b] and Stèfan Darmoni[a,c]
[a] *INSA de ROUEN, LITIS Laboratory, Rouen, France*
[b] *University and Hospitals of Geneva, Service of Medical Informatics, Geneva, Switzerland*
[c] *Rouen University Hospital & GCSIS Medical School of Rouen, France*

Abstract. Images play an important role in medical diagnostics and treatment planning. Whereas medical text retrieval, analysis, and reuse have been practiced for many years, much less is known on the storage and reuse of images in a context other than a single patient, although several medical teaching files with images exist. The aim of automatic image indexing/retrieval is to provide efficient and fast access to image collections to reuse stored information. When indexing medical images, the automatic categorization provides the means of extracting otherwise unavailable information about the images. Image categorization is usually applied in a context where a large number of images needs to be treated automatically and where no or only little text is available. This study will focus on a database extracted from online resources of the CISMeF health-catalogue. The aim is to assess the performance of two medical image categorization architectures in a complex environment (10322 images, 33 classes, multiple modalities, anatomical regions, and view-angles). The image database was extracted and annotated in collaboration with an experienced radiologist. The two compared systems have very different architectures. MedIC is an architecture based on machine learning applied to several sets of image texture and grey-level statistics. It achieves a maximum accuracy of 97.24% on 32 classes (representing different modalities, anatomical regions and view angles) and 98.47% on 6 classes (representing only the medical modality), when using an SVM classifier. MedGIFT on the other hand is a visual retrieval system where the visually most similar images are used to classify new images, currently without any learning strategy. This approach reaches an accuracy of 95.25% for the 32 classes and 97.15% for detecting the modality. By including a proper learning strategy these results can be expected to increase. The results show that image categorization has reached an accuracy that can be regarded as sufficient for many automatic image categorization tasks to get information on poorly annotated image collections.

Keywords: Medical imaging, Content-based image retrieval, image categorization, visual information retrieval

1. Introduction

Medical imaging has grown over the last decade to become an essential component of diagnosis, medical education (i.e. teaching of medical students, health professionals), and information of the general public. The development of the Internet has made medical images available in large numbers in online repositories, atlases, and other heath-related resources. These images are representing a valuable source of knowledge and are of significant importance for medical information retrieval. Unfortunately, the shear amount of medical data available online makes it very difficult for users to find exactly the images that they are searching for. The cost of manually annotating these images is prohibitively high as it is time-consuming and requires domain-related knowledge. Although many online databases have some form of annotation, it is often hard to automatically associate the text with a single image. Information, for example on the modality, is often simply missing. To address this problem two systems are presented to assess their capability to automatically extract visual information for image indexing and retrieval.

Content-based image-retrieval has already been proposed for general images as well as for medical images many times [1,2]. Image categorization has also been proposed several times to automatically extract information from large medical datasets [3]. However, most current systems concentrate on a single modality and/or anatomic region and are rarely confronted to the significant variability of the images extracted from various online resources. Some well known examples are KMeD [4] that treats MRI images of the head, ASSERT [5] dealing with lung CT images, and the National Library of Medicine (NLM) project [6] employing shape-analysis on spine x-rays

The most general image characteristic for automatic detection is the modality (CT, MRI, XRay, etc.), a problem already treated in [7]. The modality is often not well annotated in medical teaching files [8] so the automatic extraction of the modality seems beneficial even when annotation is available. Second target for automatic extraction is the anatomic region, also described several times in the literature, for example by the IRMA group [3]. With an automatic extraction of anatomic region and modality, a significant knowledge can already be extracted from the images. It is of course desirable to extract the pathology as well but this has proven difficult from the image alone.

[1] Corresponding Author: Filip Florea, 1060, Avenue de l'Université, F-76801 St Etienne du Rouvray Cedex, France. filip.florea@insa-rouen.fr

On the other hand this information is usually well documented in the annotation of the images [8]. This article will concentrate on the visual aspects, only and not treat medical ontologies for images or text-based retrieval.

With DICOM (Digital Imaging and Communication in Medicine) a standard for image production has become available that stores information on the production modality and anatomic regions. Still, these fields often contain a significant amount of errors, for example in the field anatomic region (~16%, [9]). In the context of the Internet the situation becomes even more complex. To be posted on web pages, the images are often converted to formats such as JPEG and GIF. These formats are used because they are easy to handle (i.e. file size, integration with Internet technologies) but also for anonymization purposes (when converted the images are loosing the DICOM header information). Unfortunately, with anonymization, the images are loosing not only confidential patient or hospital information, but also the technical information residing in the DICOM headers (e.g. acquisition modality and parameters, anatomical region, …). In this case the automatic extraction of modality and anatomic region using content-based methods permits the search at least along these two axes.

On the Internet, a variety of medical search engines and information repositories exist such as HON[2] (Health On the Net), HEAL[3] (Health Education Assets Library), or CISMeF[4] (French acronym for Catalog and Index of French-language health resources). In this paper we will use a database extracted from the CISMeF resources to compare two approaches for automatic image categorization. CISMeF is a quality-controlled subject gateway initiated by the Rouen University Hospital (RUH) in 1995 [10]. Its role is to provide online searching capabilities for health resources, by describing and indexing the most important documents of institutional health information in French.

The rest of this paper is organized as follows. Section 2 will describe the dataset used, Sections 3 and 4 the two applied systems. Section 5 presents the results and the last sections discuss the results critically and show ways for future improvements of our strategy.

2. Methods

2.1. CISMeF image database

The dataset used in this paper consists of 10322 anonymous images from 32 classes representing the medical modality, anatomic region and acquisition view angle. The images were extracted from the Rouen University Hospital's clinical dataset and from web documents indexed in CISMeF. We considered the six most frequent modalities: angiography, ultrasonography (US), magnetic resonance imaging (MRI), standard radiography (RX), computer tomography (CT), and scintigraphy.

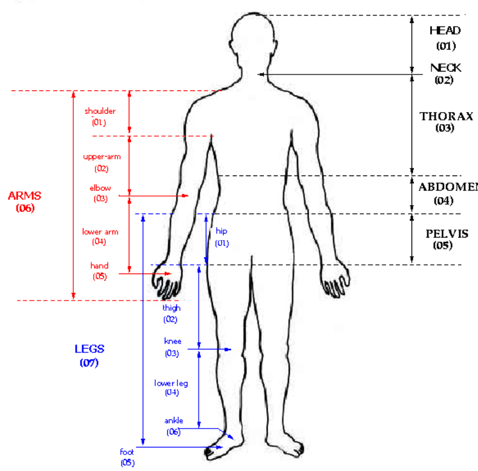

Figure 1: The anatomic regions differentiated in our database.

For each modality, a trained radiologist divided the corpus by anatomical region (e.g. head, thorax, legs), sub-anatomical regions (e.g. hip, knee, ankle), and acquisition views (coronal, axial, sagittal) (see Figure 1). This hierarchical organization seems to be appropriate in this context as it allows to partition images by acquisition technique and by regional criteria. The frequency of the images in the various classes can be examined in Figure 2.

[2] http://www.hon.ch/
[3] http://www.healcentral.org/
[4] http://www.cismef.org/

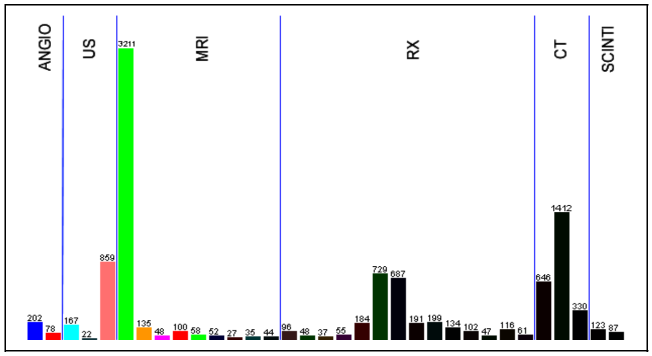

Figure 2 – Distribution of the numbers of images in the classes.

The images were not all stored using the same parameters. We noted strong variations in size, compression ratio, contrast, background, and text marked directly in the image. This can be explained with the fact that images were acquired with varying digital or analogical equipment in different hospitals in a time span of several years. Images published on the Internet are usually further modified (e.g. resizing, cropping, high-compression, addition of superposed didactical annotations and drawings). When converting tomographic images into JPEG, the level/window settings can also vary, depending on the system under observation. Thus, the intra-class variability is high. Categorization is also complicated by a strong inter-class similarity (representing different regions or modalities):

Table 1: Modality and anatomic region for Figure 3.

(MRI-head-***-axial)	(ANGIO-thorax-***-***)	(ANGIO-***-***-***)
(CT-head-***-axial)	(RX-thorax-***-***)	(MRI-MRA-***-***)

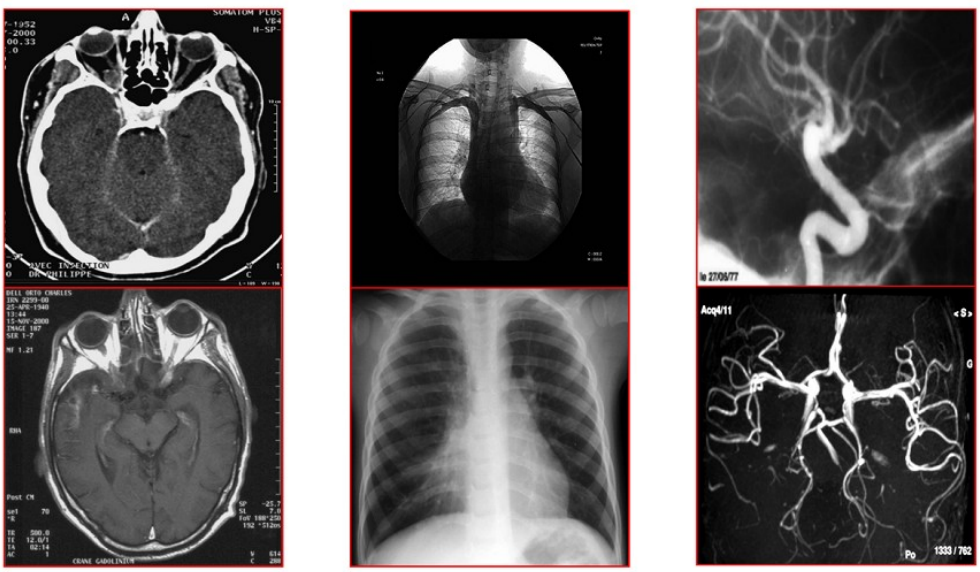

Figure 3 – Examples for inter-class similarity

3. The MedGIFT approach

The MedGIFT[5] system is an image retrieval engine [11]. It is based on the open source system GIFT[6], outcome of the Viper[7] project of the University of Geneva. This system offers components for content-based indexing and retrieval of images such as feature extraction algorithms, feature indexing structures and a communication interface called MRML (Multimedia Retrieval Mark-up Language). GIFT uses techniques from text retrieval such as frequency-based feature weights, inverted file indexing structures, and relevance feedback mechanisms. As visual features to represent images, four feature groups are chosen:

[5] http://www.sim.hcuge.ch/medgift/
[6] http://www.gnu.org/software/gift/
[7] http://viper.unige.ch/

- local and global texture features based on responses of Gabor filters;
- color/grey scale characteristics on a global image scale and locally within image regions.

Gabor filters measure the change in the image in a certain direction and scale. This means that it describes a texture with respect to its directions as well as with respect to the size and strength. Small or slow changes can easily be distinguished from quick and large changes. Local features are obtained by successively dividing the image into four regions of the same size. The mode color of each region at each scale is taken as visual feature. This creates a multi-scale representation of the image. Local Gabor filters allow determining in which region which shapes or textures occur. The potential feature space is very large (>85,000 possible visual features). Each image contains roughly 1'000-2'000 features. The feature weighting is based on the following two principles:

- Features that are frequent in an image describe this image well (described by *df*, document frequency);
- Features that are frequent in the entire collection do no distinguish images well from each other and are thus less important (*cf*, collection frequency).

In this paper, the standard configuration of GIFT is taken into account using the Gabor filter responses in four directions and at three different scales. The color features are extracted in the HSV (Hue, Saturation, Value) space taking into account a total of 166 colors among them only four grey levels. A small number of grey levels has shown to have a good performance for retrieval despite a significant loss of information. They are more invariant with respect to image acquisition parameters. Although a general learning architecture of feature weights exists for GIFT [12], this could not be implemented for this paper due to time constraints.

To use a retrieval system for image categorization, a very simple approach was used. We simply took the first image retrieved from the database (the most visually similar) and used its modality and anatomic region for categorization. A more complicated approach could take into account more than a single image and can thus improve the categorization quality.

4. The MedIC approach

The MedIC (Medical Image Categorization) approach is developed inside the CISMeF team. The goal is to add a "document retrieval by image" functionality to the Doc'CISMeF search engine, by providing image information (modality, anatomical region, view angle) when indexing CISMeF resources (i.e. documents) containing images. This will allow users to perform image-oriented queries using image-related keywords (*e.g.* "find me all the resources/documents containing CT images").

This approach can be viewed as a two-stage-process: a) first a representation phase where image features are extracted to describe the visual content and b) a training/classification phase where machine learning techniques are used to classify the image description vectors into defined classes.

As for MedGIFT, to take into consideration the spatial disposition of information inside the images, the features were extracted locally. The original images are downscaled to 256x256 and split in 16 equal blocks. Thus, each image is represented by a vector of 16 blocks, and from each block features will be extracted to describe its content.

The specificity of the medical image database presented in this paper is ruling out features like color (successfully used for non-medical image representation) and shape (the shape features are both difficult to extract – e.g. ultrasonography – and much too variable for some classes – e.g. angiography). MedIC uses a combination of texture-based features and statistical grey level measures that proved to be a well-suited descriptor for medical images [13]. Other than the responses of Gabor filters, MedIC uses texture features extracted from the grey-level co-occurrence matrices, fractal dimensions [14], and the Galloway run-lengths [15]. Features derived from statistical grey-level measures are various estimations of the first order (mean, median and mode), second order (variance and L2 norm), third, and forth order moments (skewness and kurtosis).

For training and classifying this representation, two well known supervised classifiers were chosen, a simple and fast K Nearest Neighbors (i.e. KNN – similar to the approach used by MedGIFT) and a complex and more accurate Support Vector Machine (i.e. SVM). Upon testing, the best parameters were: 1NN (i.e. first neighbor, for KNN) and a polynomial kernel, second degree with a C=100 penalty coefficient (for SVM). For the classification a 10-fold cross-validation scheme was used.

5. Experiments and Results

Table 1 presents the results in terms of global categorization accuracy.

The MedGIFT system was run in a single configuration and with an extremely simple categorization algorithm. It simply took the nearest neighbor of a new image and took its class for categorization. No learning was used. A first run was done comparing only the modality. This run led to an error rate of 2.85% (97,15% accuracy) for a six-class problem. The 32-class problem led to an error rate of 4.75% (95,25% accuracy), having 9831 images correctly classified and 491 images incorrectly.

Table 1. Comparison of the accuracy of MedGIFT and MedIC.

	MedGIFT		MedIC			
	accuracy	no. missclassified	accuracy	no. missclassified	accuracy	no. missclassified
32 class	95,25%	491	95,28%	487	97,24%	285
6 class	97,15%	295	97,88%	219	98,47%	158
	1-NN		**1-NN**		**SVM**	

Using a similar nearest neighbor architecture MedIC obtains similar results. The differences are most likely caused by the fact that MedGIFT is not applying any learning. A proper learning scheme as described in [12] can improve results significantly. MedIC uses a 10-fold cross validation scheme (the dataset is split in 10 bins, on which 10 runs are made, each time using 90% as annotated examples and 10% for test). The SVM classifier gains 1-2% of accuracy, misclassifying 158 images, when only the modality is needed. This proves that using a learning architecture is more accurate. However, a learning architecture is time-consuming, SVM being several times slower than KNN, even when highly optimized.

Usually, the classification technique assumes that training examples are evenly distributed among classes. Our dataset is unbalanced, meaning that some classes are representing a large portion of all the examples, while others have are representing only a small percentage. Unbalanced datasets can influence the classification performance. The "hardest" and "easiest" five classes of the MedIC system (based on the resulted F-measure – a harmonic average of precision and recall) are presented in Table 2.

Table 2: The "hardest" and "easiest" classes of the MedIC system.

No.	MedIC F-measure	Class	No of images
1	0.69	RX-LEGS-ANKLE-SAGITAL	181
2	0.74	MRI-LEGS-***-CORONAL	135
3	0.74	RX-HEAD-***-CORONAL	729
4	0.74	MRI-HEAD-***-CORONAL	135
5	0.80	RX-LEGS-KNEE-SAGITAL	47
…	…	…	…
28	0.97	RX-NECK-***-SAGITAL	202
29	0.97	CT-HEAD-***-AXIAL	646
30	0.98	RX-THORAX-***-CORONAL	37
31	0.98	MRI-HEAD-***-AXIAL	3211
32	0.99	RX-THORAX-MAMMO-SAGITAL	686

6. Discussion

The results show that image classification is mature for automatic information extraction on large, poorly annotated databases. Particularly when using a proper learning approach such as the SVM technique of MedIC, error rates even for large imbalanced datasets can be below 2%. Creating a balanced dataset for learning is often used for classifying images but this does not correspond to the reality, so imbalanced datasets and their consequences for classification need to be taken into account. The MedGIFT system does not have as good results for the classification as the SVM technique but is as good as the MedIC KNN system. This means that the feature space must be representing the image well. Applying a learning strategy [12] and small changes feature space are expected to improve results significantly. Another advantage is its free availability and versatility. It is easy to adapt to new problems as new features can easily be used. It can be applied out-of-the-box without need for coding. Quick response times are another advantage of MedGIFT (below 0.5 seconds for classification with a simple desktop PC). Whereas a KNN approach can be roughly as quick and accurate, SVMs are significantly slower but better. Nevertheless, the categorization and indexing of medical images is generally done offline (before the images are searched). Therefore, in this particular context, the classification speed is less important.

The experiments described in this article show that image categorization and retrieval have reached a quality that seems sufficient for simple information extraction from large repositories such as those available on the Internet.

7. Conclusion

Image indexing and retrieval is an important research domain as much less analysis has been done with medical images than with medical texts. Although new imaging modalities include information on the modality and also anatomic region in their DICOM headers, some of the fields are known to contain significant error rates [9]. Another problem is that many images are converted to jpeg to anonymize them and also to show them in simple web interfaces to avoid the burden of dealing with large DICOM files. Some of these collections contain annotations but not all of them. Image retrieval benchmarks of annotated medical image databases such as ImageCLEFmed show that the annotation is often not sufficient to determine modality or anatomic region sufficiently well [8]. This is a domain where visual image categorization algorithms can have a significant impact as they can quickly classify very large numbers of images, for example from web repositories.

The presented algorithms show that fairly simple approaches are sufficient to classify a large number of images quickly and with small error rates. More optimized approaches such as SVMs can improve the results but require more time for training. Using MedGIFT with proper learning algorithms included still needs to be tested and promises good results with a reasonable effort.

Acknowledgements

The authors want to thank Jean-Nicolas Dacher, MD, PhD, and head of the Radiology Service at Rouen University Hospital, for his support in gathering and organizing the image database used.

References

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

[2] AWM Smeulders, M Worring, S Santini, A Gupta, R Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12):1349-1380, 2000.

[3] TM Lehmann, MO Güld, C Thies, B Fischer, D Keysers, M Kohnen, H Schubert, BB Wein. 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.

[4] W Chu, C Hsu, A Cardenas, R Taira. Knowledge-based image retrieval with spatial and temporal constructs. *IEEE Transactions on Knowledge and Data Engineering*, **10**(6):872-888, 1998.

[5] CR Shyu, CE Brodley, AC Kak, A Kosaka, AM Aisen, LS Broderick. ASSERT: A physician-in-the-loop content based retrieval system for HRCT image databases, *Computer Vision and Image Understanding*, **75**(1/2):111-132, 1999.

[6] L Long, S Antani, D Leeb, D Krainakc, G Thoma. Biomedical information from a national collection of spine x-rays film to content-based retrieval. *In Proceedings SPIE*, **5033**, San Diego, California, 2003

[7] A Mojsilovic, J Gomes. Semantic-based categorisation, browsing and retrieval in medical image databases. *Proceedings International Conference on Image Processing(ICIP2002)*, 2002.

[8] P Clough, H Müller, T Deselaers, M Grubinger, T Lehmann, J Jensen, W Hersh, The CLEF 2005 Cross-Language Image Retrieval Track, *Springer Lecture Notes in Computer Science*, 2006 – to appear.

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

[10] SJ Darmoni, JP Leroy, B Thirion, F Baudic, M Douyère, J Piot. Cismef: a structured health resource guide, *Methods of Information in Medicine*, **39**(1):30-35, 2000.

[11] H Müller, A Rosset, JP Vallée, A Geissbuhler, Integrating Content-Based Access Methods into a Medical Case Database, *Proceedings of Medical Informatics Europe (MIE 2003)*, 480-485, St. Malo, France, 2003.

[12] H Müller, DM. Squire, T Pun, Learning From User Behavior in Image Retrieval: Application of the Market Basket Analysis, *International Journal on Computer Vision*, **56**(1-2):65-77, 2004. (Special Issue on Content-Based Visual Information Retrieval)

[13] F Florea, A Rogozan, A Bensrhair, S Darmoni. Medical image retrieval by content and keyword in an on-line health-catalogue context, *Proceedings of Mirage 2005 (Computer Vision / Computer Graphics CollaborationTechniques and Applications*, 229-236, France, 2005

[14] B Jahne, H Haussecker, P Geissler. Handbook of Computer Vision and Applications, **2**, *Academic Press,* 1999

[15] MM Galloway. Texture Analysis using Graylevel Runlengths Computer, *Graphics and Image Processing,* **4**:172-179,1975.