Enriching content-based medical image retrieval with automatically extracted MeSH-terms

Müller H, Ruch P, Geissbuhler A

Service of Medical Informatics, University Hospitals of Geneva, 24 Rue Micheli-du-Crest, 1211 Geneva 14, Switzerland, henning.mueller@sim.hcuge.ch, http://www.sim.hcuge.ch/medgift/

Introduction

The rising amount of digitally produced images and other visual documents such as signal curves and videos in medical departments creates a need to develop new tools to manage these visual databases. The radiology department of the University hospitals of Geneva alone produced, for example, more than 13,000 images per day in 2003. Current access methods to these visual data are most often limited to access by numerical patient identification. Sometimes, the search by textual key words from the radiology report [1] or the electronic patient record is possible. Content-based image retrieval on the other hand allows to browse and search in large image collections based on visual features that are automatically extracted from the images and consequently cheap to produce.

Scenarios for content-based visual data access

Not only within teaching and research exists a need to browse large image databases by their visual content to find interesting or important cases and to compare visually similar images and their diagnoses. Such functionalities are needed to use the images up to their full potential. Also for fields such as case-based reasoning or evidence-based medicine, there is a need for finding similar medical cases. When only one or several image(s) are available for diagnostics, an image retrieval system can deliver pointers that might lead to a correct diagnosis. In pathology, dermatology, or for high resolution CTs of the lung the diagnosis depends strongly on texture and colour/grey level characteristics of the images. Thus, the medically similar cases are often visually similar cases that can be found by content-based image retrieval.

Content-based access to medical images has been proposed several times [2,3,4,5] and a few research projects exist such as IRMA (Image Retrieval in Medical Applications, [6]) or ASSERT (Automatic Search and Selection Engine with Retrieval Tools, [7]). In a first test as a tool for diagnostic aid, an image retrieval system has shown to improve the diagnostic quality [8].

Although the retrieval quality is sufficient for some tasks and the automatic extraction of visual features is rather convenient, there is still a semantic gap between the low-level visual features (textures, colours) automatically extracted and the high-level concepts that users normally search for (tumour, abnormal tissue).

Extracting MeSH term in addition to visual features

Proposed solutions to bridge this semantic gap are the connection of visual features to known textual labels of the images [3] or the training of a classifier based on known class labels and the use of the classifier on unknown cases [6]. Combinations

of textual and visual features for medical image retrieval have as of yet rarely been applied, although medical images in the electronic patient record or case databases basically always do have text attached to them. The complementary nature of text and visual image features for retrieval promises to lead to good retrieval results.
Radiology reports that come with the images, on the other hand, have other problems. Often, the quality of the text is not extremely good. Spelling errors, various and differing abbreviations and non-standardized coding hinder the efficient retrieval [9]. In our case database system casimage (http://www.casimage.com/, [10]), we also have the problem of having English and French case descriptions.
For our tests, we use a database that contains a total of almost 9000 images of more than 2000 medical cases. Case descriptions are mixed in English and French with a few hundred images not containing any annotation at all. The quality of the texts is extremely varied as no control of the text input was applied. To avoid problems with the annotations we extract only a small number of around three MeSH terms with a tool called easyIR that takes as an input the entire text cleaned of the XML tags and returns an ordered list of most likely corresponding MeSH terms. English and French MeSH terms can thus be identified with one unique identifier. The first 3-5 terms have shown to deliver high accuracy in several tests on other, similar data sets.

Results

As an image retrieval framework we use the medGIFT system [11] that is based on the GNU Image Finding Tool. The system uses several techniques from text retrieval applied to images and their visual features such as frequency-based feature weights based on standard tf/idf (term frequency, inverse document frequency) [12]. This means that features (or words) that are rare in the collection are weighted higher than very frequent features or words. A first step in this direction has been achieved [13], the automatic assignment of UMLS concepts to textual reports can be performed with high precision (up to 92%); therefore the integration of the extracted MeSH terms into this framework is be easy to implement and is a natural extension of the used feature weighting.

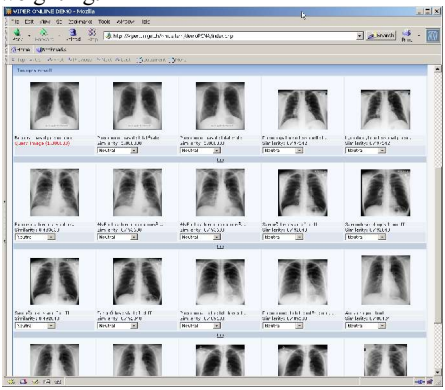

Figure 1: Screenshot of the medGIFT web interface.

Visual features that are used for retrieval include a simple colour histogram intersection as well as local colour blocks at different scales and locations. As

texture features, we use the responses of Gabor filters in various scales and directions as well in the form of a histogram (globally) as locally, in fixed image regions. Relevance feedback is possible with as many input images as needed, and a web-based user interface (see Figure 1) allows easy querying as well as a connection to the complete textual description and other images stored for the same case in the teaching file (http://www.sim.hcuge.ch/medgift/).
First results show that the quality using textual and visual features combined is superior to either one of the technologies. Images with bad annotation can still be found due to a high visual similarity and the MeSH terms add semantics and reduce the rate of false positives.

Discussion

The use of MeSH terms for image retrieval in addition to automatically extracted visual features has shown good first results and is one possibility to bridge the semantic gap. Much work still needs to be done with respect to a quantitative evaluation of this combination. We also still need to figure out which will be the best combination between visual and textual features to optimize retrieval results.
An effort to evaluate medical image retrieval algorithms using visual or visual and textual data is underway in connection with the CLEF conference (Cross Language Evaluation Forum, http://ir.shef.ac.uk/imageclef2004/index.html).

References

[1] C. Le Bozec, E. Zapletal, M.-C. Jaulent, D. Heudes, and P. Degoulet. Towards content-based image retrieval in HIS-integrated PACS. In *Proceedings of the Annual Symposium of the American Society for Medical Informatics* (AMIA), pages 477-481, Los Angeles, CA, USA, November 2000.

[2] E. El-Kwae, H. Xu, and M. R. Kabuka. Content-based retrieval in picture archiving and communication systems. *Journal of Digital Imaging*, 13(2):70-81, 2000.

[3] L. H. Tang, R. Hanka, H. H. S. Ip, and R. Lam. Extraction of semantic features of histological images for content-based retrieval of images. In *Proceedings of the IEEE Symposium on Computer-Based Medical Systems* (CBMS 2000), Houston, TX, USA, June 2000.

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

[5] H.J. Lowe, I Antipov, W Hersh, C.A. Smith, Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation and knowledge-based retrieval, *Proceedings AMIA Symposium*:882-6, 1998.

[6] D. Keysers, J. Dahmen, H. Ney, B. B. Wein, and T. M. Lehmann. A statistical framework for model-based image retrieval in medical applications. *Journal of Electronic Imaging*, 12(1):59-68, 2003.

[7] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. 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.

[8] A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, A. Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment, *Radiology* 228:265-270, 2003.

[9] P. Ruch, R. Baud, and A. Geissbühler. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. *International Journal of Medical Informatics*, 67:75-83, 2002.

[10] A. Rosset, O. Ratib, A. Geissbuhler, and J.-P. Vallée. Integration of a multimedia teaching and reference database in a PACS environment. *RadioGraphics*, 22(6):1567-1577, 2002.

[11] H. Müller, A. Rosset, J.-P. Vallée, and A. Geissbuhler. Integrating content-based visual access methods into a medical case database. In *Proceedings of MIE 2003*, St. Malo, France, May 2003.

[12] G. Salton, C. Buckley, Term weighting approaches in automatic text retrieval, *Information Processing and Management* 24(5):513-523, 1988.

[13] P. Ruch, R. Baud, A. Geissbühler. Learning-free Text Categorization, *Proceedings of Artificial Intelligence in Medicine in Europe (AIME), Springer Lecture Notes in Artificial Intelligence* 2780, 2003.