Enriching content-based image retrieval with multi-lingual search terms

Müller H, Ruch P, Geissbuhler A

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

Summary:

Content-based visual image retrieval is a research domain with the goal to allow efficient access to the large amount of visual information that is being produced in medical institutions. Currently, visual retrieval is most often strictly separated from textual information extraction and retrieval in medical records. The complementary nature of the two methods by contrast invites to use the two together in an integrated fashion. We use a visual retrieval system (*medGIFT*) and a textual search engine powered with biomedical terminological resources (*easyIR*) together on a data set presented at the *imageCLEF* image retrieval competition. Both systems are available free of charge as open source. The dataset is also publicly available to make results reproducible and comparable. Results show that a simple combination of visual and textual features for retrieval improves performance significantly for fully automatic retrieval as well as for runs with manual relevance feedback. The currently applied techniques are fairly simple combinations and better results can be expected when optimizing the combined weighting based on learning data. Visual and textual features should be used together for information retrieval whenever they are both available to allow optimal access to varied data sources.

Keywords:
Medical image retrieval, Multilingual search, Multimodal information retrieval, Cross-media indexing, Content-based data access

Introduction

The rising amount of digitally produced images and other visual/audiovisual documents such as signal curves and videos in medical departments creates a need to develop new tools to manage these data. The radiology department of the university hospitals of Geneva alone produces currently (2004) more than 20,000 images per day. Connections of other departments such as cardiology, hematology and pathology to the PACS are planned and will continue to augment the amount of data produced. Standard access methods to these visual data are most often limited to access by numerical patient identification. Sometimes, search by textual key words from the radiology report [1] or electronic patient record is possible. Content-based image retrieval (CBIR), on the other hand, allows browsing and searching in large image collections based on visual features that are automatically extracted from images and consequently cheap to produce [2]. Content-based image retrieval has been one of the most active research areas in computer vision over the last ten years with hundreds of systems being developed. The use in the medical domain has been proposed for almost ten years and important clinical benefits are expected [3,4,5]. Still, content-based access is rarely used in clinical practice as the paradigm of using positive and negative example images to formulate queries is not straightforward and users will need to get used to it. First applications are expected for access to medical teaching files [6] where retrieval quality is not critical but where a user is interested in browsing large collections. The final goal of content-based image retrieval is the use as a diagnostic aid. An example for such a possible use as a diagnostic aid is shown in Figure 1, where based on a query image, other visually similar cases are found to provide evidence for or against a certain diagnosis.
Such functionalities can be beneficial in modern information systems to use the images up to their full potential. For fields such as case-based reasoning [7] or evidence-based medicine [8], there is a need for finding similar medical cases. When one or several image(s) are available for diagnostics, an image retrieval system can deliver pointers to similar cases 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

color/grey level properties of the images. Thus, medically similar cases are cases with visually similar images that can be found by CBIR.

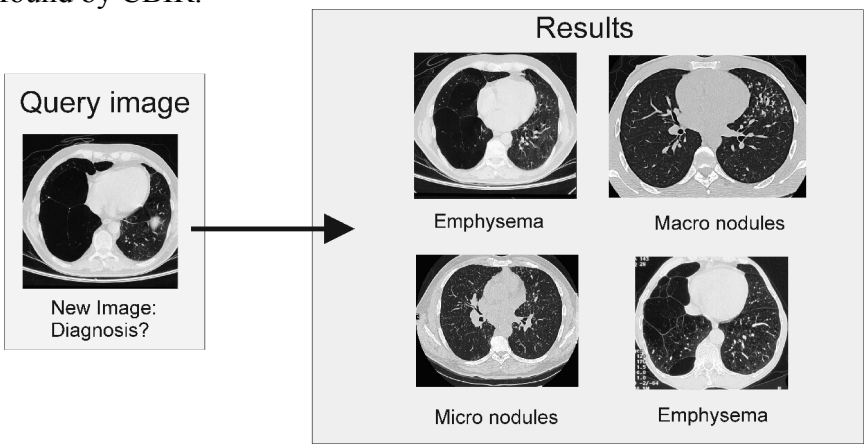

Figure 1: Based on an example query image, visually similar other images are found in a database based on automatically extracted visual features.

Few active research projects in medical image retrieval exist such as *IRMA* (Image Retrieval in Medical Applications[1], [9]), *ASSERT* [10] and *medGIFT*[2] [6]. In a first test, as a tool for diagnostic aid, *ASSERT* has shown to improve the diagnostic quality significantly [11], especially among radiologists not being specialized in chest CTs. Although retrieval quality is sufficient for many tasks and automatic extraction of visual features is convenient, there is still a *semantic gap* between the low-level visual features (textures, colors, shapes) automatically extracted and the high-level concepts that users normally search for (tumor, abnormal tissue). Only in narrow domains, higher-level features and automatic segmentations can be used to extract higher-level visual image features.

Most of the stored cases including images are not unconnected in a medical context but do contain a radiology report or textual data from the patient record attached to them. These data can further on be used to improve the retrieval quality and allow semantic retrieval through automatic query expansion (QE) even when the initial query only contains images and no or little text. Natural language processing and information extraction from medical texts are two other very active domains of research. Often, UMLS concepts or MeSH terms are tried to be extracted from texts and show good results [12]. Still, there is currently no connection of visual retrieval projects and textual medical information retrieval. Sometimes the use is proposed [13]. The complementary nature of text and visual image features for retrieval promises good combined results. Most current systems only try to classify the images based on their visual content into known classes [9]. In the non-medical domain, several articles describe connections between visual and textual characteristics for retrieval. Usually, well-constructed annotations with few words [14] or captions of newspapers for images [15] are used.

Radiology reports and teaching files that the images are attached to have other problems. Often, the quality of the text is mediocre. The texts are often not produced for being searched by keywords afterwards. Spelling errors, various and differing abbreviations and non-standardized coding hinder efficient retrieval [12]. In our case database system casimage[3] [16], we also have the problem of having English and French descriptions mixed, as well as several empty case notes (~7%). For our tests, we index this *casimage* database that contains a total of 9000 images of 2000 medical cases used in the *imageCLEF* image retrieval competition. *ImageCLEF*[4] [17] started in 2003, and in 2004 a medical image retrieval task was added. Although current tasks do not reflect a real medical target, the goal is to augment the difficulty of tasks successively and create medically relevant tasks within the next one or two years. The methodology is based on evaluation standards well known and accepted of text retrieval conferences such as TREC[5] (Text Retrieval Conference). Goal of *imageCLEF* 2004 was to motivate research groups to combine visual and textual features. Based on this goal, medical query topics were 26 images without annotation. A total of 18 groups participated in the image retrieval competition, 11 in the medical task. Our *medgift/easyIR* system presented the best performance in the medical task of all participants in 2004.

[1] http://www.irma-project.org/
[2] http://www.sim.hcuge.ch/medgift/
[3] http://www.casimage.com/
[4] http://ir.shef.ac.uk/imageclef2004/
[5] http://trec.nist.gov/

Methods

Both retrieval systems used in our approach are available as open source. *medGIFT* is based on the GNU Image Finding Tool (*GIFT*[6]). *EasyIR* can also be downloaded[7].

easyIR

Before actually indexing the texts, several steps are needed to preprocess the data. XML tags were removed from the documents as well as the unimportant fields of the annotation such as the name of the medical doctor who included the case, etc. This information is able to improve retrieval quality as the same medical doctor will add cases of the same anatomic region but is not our retrieval goal.

As reports can contain both French and English written parts, boundary detection of language segments and storage in separate indexes would have been best. Considering the lack of time, we decided to index the *casimage* collection using a unique index. We used the Porters stemmer for English and a modified version of Savoy's conflation tool for French. Depending on the index, a list of stop words was used, 544 items for English, 792 for French. We also use a biomedical thesaurus, which has proven its effectiveness in the context of other evaluation campaigns [18]. For English, 120'000 string variants were extracted from UMLS[8], while the French thesaurus contains about 6'000 entries [19]. Both resources were merged for the experiments. Our submitted runs were produced using the English index without specific translation. All documents are indexed as full text using a "bag of words" approach. Queries are single case reports pretreated in the same way or several reports simply pasted together as a single text. A well-known weighting schema, atc.ltn [20] was selected a priori for our experiments showing good feedback performance (Table 1).

Term Frequency	
First Letter	$f(tf)$
n (natural)	tf
l (logarithmic)	$1 + log(tf)$
a (augmented)	$0.5 + 0.5 \times (tf/max(tf))$
Inverse Document Frequency	
Second Letter	$f(1/df)$
n(no)	1
t(full)	$log(N/df)$
Normalization	
Third Letter	$f(length)$
n(no)	1
c(cosine)	$\frac{1}{\sqrt{\rho_1^2 + \rho_2^2 + ... + \rho_n^2}}$

Table 1: Weighting schemes used for the indexing

medGIFT

As image retrieval framework we use *medGIFT* [6]. The system uses several techniques from text retrieval applied to images and their visual features such as inverted files and frequency-based feature weights based on standard tf.idf (term frequency, inverse document frequency) [20]. Features (or words) that are rare in the collection are weighted higher than frequent features. Relevance feedback (positive and negative) is possible with as many input images as needed. A web-based user interface (Figure 2) allows easy querying and a connection to the radiology teaching file.

[6] http://www.gnu.org/software/gift/
[7] http://lithwww.epfl.ch/~ruch/softs/softs.html
[8] http://www.nlm.nih.gov/research/umls/

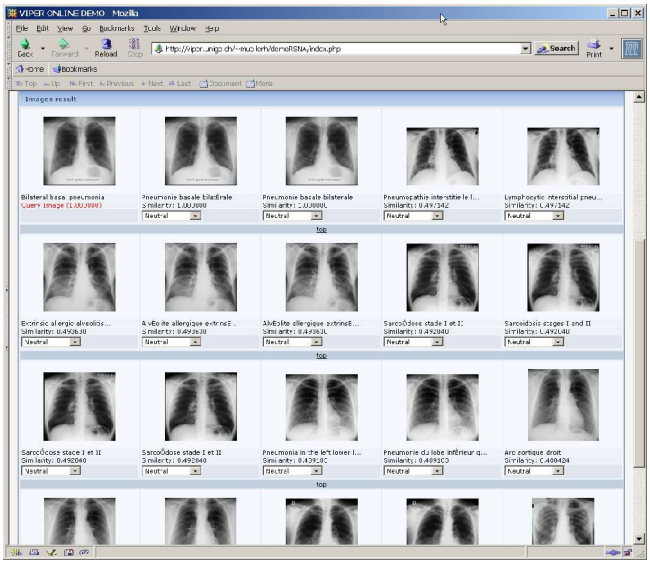

Figure 2: Screenshot of the medGIFT web interface.

Visual features used for retrieval include a color histogram intersection as well as local color blocks at different scales and locations (by dividing the image symmetrically into four sub regions, repeating this four times for each sub block, Figure 3). As texture features, we use the responses of Gabor filters in various scales and directions in the form of a histogram (globally) and locally, in fixed image regions.

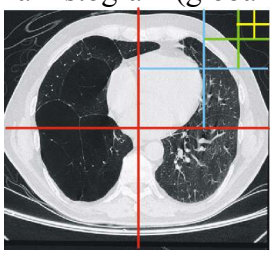

Figure 3: An image is subdivided into blocks of varying size and scale for feature extraction.

The histogram features are compared via a histogram intersection while the block features use a tf.idf weighting. The four feature groups are evaluated separately and then normalized. This normalization before combination is necessary to avoid that frequent features dominate the result. To avoid common problems appear with too much negative feedback, we weight positive and negative parts of a query separately and combine them afterwards, so-called Rocchio feedback [21].

combination of results from the two systems

Both retrieval systems are separate entities. For connecting the two, perl scripts are combining the results. After an initial query (an image only) with the visual system, the first N=1..3 images are used to expand the query to text. The text of the cases of these images was simply combined and submitted to *easyIR*. The results from *easyIR* are normalized by the case with the highest score to have a result within [0;1]. The text scores are based on the case and not the image. We achieve a list of images by simply expanding the textual score of a case to all images that are part of this case, whether visually similar or not. Another query with the N images plus the initial query image was performed visually with the results equally being normalized. Afterwards we can simply combine the two normalized values as follows:

$$score_{total} = \alpha \cdot score_{visual} + \beta \cdot score_{textual}$$

The main question is now how to weight the visual and the textual parts in this combination. Optimal weighting depends on the task and the ground truth. In our case, relevance was defined as an image being created with the same modality, same anatomic region, same viewing direction, and depending on the case same radiological protocol. This puts more stress on the visual similarity, which means that the visual part of the score has to be weighted higher. Due to a lack of training data and expert opinion to optimize the weighting, we use three different settings: 75%, 80% and 90% for the visual component.

Results

Retrieval results show that the quality using textual and visual features combined is superior to either one of the technologies alone. Images with bad or missing annotation can still be found due to a high visual similarity and text terms add semantics and reduce the rate of false positives. The lead measure for *imageCLEF* as for other benchmarks is *mean average precision* (MAP). Precision is defined as follows:

Precision=Number of relevant images retrieved/Number of all images retrieved

Mean average precision is the average precision at the points where relevant images are retrieved. MAP gives a good overview of system performance. Other measures such as precision vs. recall graphs are as well available for all participating systems.

The best visual system is a system from Aachen University (**MAP=0.3858**) using manually set weightings. The best automatic visual *medGIFT* run received a score of **MAP=0.3757**, thus slightly lower. The best automatic run using text and image information is our system (**MAP=0.4020**) leading in front of the next best system in the competition from the State University New York (**MAP=0.3858**). When using relevance feedback medGIFT is the best visual and the best visual/textual system in the competition. The visual feedback result is at **MAP=0.4469**. The best visual/textual feedback run is at **MAP 0.4847**. Weighting the visual part differently strong shows that a relatively high weighting of the visual part (90%) leads to best results whereas only 75% leads to worst results. The optimum is expected to be in between 80% and 90% for the visual part. These results show that the *medGIFT*/*easyIR* combination delivers the best results in the competition and it also shows that combinations of visual and textual combinations lead to best results whether for automatic or manual feedback runs. Although the text quality is not optimal, the retrieval results improved strongly.

Discussion and conclusions

Although the *imageCLEF* benchmark is still several steps away from medically relevant retrieval tasks, it managed to activate 18 research groups and thus a large community of visual retrieval researchers. This shows that there is a strong need for standardized datasets, tasks and evaluations. Clearly, the tasks will need to be oriented more towards real user needs in the future.

Our evaluation of *medGIFT* shows that the system is among the best in the competition. Especially when using relevance feedback the system delivers the best results. For automatic queries as well as for manual feedback the optimal results were received when combining visual and textual cues for retrieval. This underlines that both, visual and textual features should be used together whenever the two are available. Currently they are most often used in completely separate ways. Collections, not only in the medical domain, are getting increasingly multimodal containing free text, numerical values as well as signals, images and also videos. All these data need to be used together in an integrated fashion for information retrieval and for allowing optimal access to large data collections. Our current teaching file already contains more than 60’000 images of more than 12’000 cases and is growing steadily.

Such new search forms and multimodal interaction paradigms also need to be evaluated based on real data. Currently used tasks are not really medically relevant but it is important to develop these benchmarks towards medically relevant tasks. Currently, these techniques are not used in clinical practice but it is important to develop models for their use. We are currently organizing a user survey among people using images in the hospital and the way that these users currently search for and organize images. Standardized evaluation is extremely important in this context and comparisons of techniques even more to identify promising research directions.

References

[1] Le Bozec C, Zapletal E, Jaulent MC, Heudes D, Degoulet P. Towards content-based image retrieval in HIS-integrated PACS. In *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, pp 477-81, Los Angeles, CA, USA, November 2000.

[2] 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* 22(12):1349-80, 2000.

[3] Müller H, Michoux N, Bandon D, Geissbuhler A, A review of content-based image retrieval applications – clinical benefits and future directions, *International Journal of Medical Informatics* 73:1-23, 2004.
[4] Tagare HD, Jaffe C, Duncan J. Medical image databases: A content-based retrieval approach. *Journal of the American Medical Informatics Association,* 4(3):184-98, 1997.
[5] Lowe HJ, Antipov I, Hersh W, Smith CA, 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] Müller H, Rosset A, Vallée JP, Geissbuhler A. Integrating content-based visual access methods into a medical case database. In *Proceedings of MIE 2003*, St. Malo, France, May 2003.
[7] Abidi SSA, Manickam S, Leveraging XML-based medical records to extract experimental clinical knowledge – an automated approach to generate cases for medical case-based reasoning systems, *International Journal of Medical Informatics* 68:187-203, 2002.
[8] Bui AAT, Taira RK, Dioniso JD, Aberle DR, El-Saden S, Kangarloo H, Evidence-based radiology – requirements for electronic access, *Academic Radiology* 9:662-669, 2002.
[9] Keysers D, Dahmen J, Ney H, Wein BB, Lehmann TM. A statistical framework for model-based image retrieval in medical applications. *Journal of Electronic Imaging*, 12(1):59-68, 2003.
[10] Shyu CR, Brodley CE, Kak AC, Kosaka A, Aisen AM, Broderick LS. 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.
[11] Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Pavlopoulou C, Dy J, Shyu CR, Marchiori A, Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment, *Radiology* 228:265-270, 2003.
[12] Ruch P, Baud R, Geissbühler A. 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.
[13] Orphanoudakis S, Chronaki C, Vamvaka D. I2C net: Content-based Similarity Search in Geographically Distributed Repositories of Medical Images. *Computerized Medical Imaging and Graphics*, 20 (4): 193-207, 1996.
[14] Sclaroff S, La Cascia M, Sethi S. Unifying textual and visual cues for content-based image retrieval on the world wide web. *Computer Vision and Image Understanding* 75(1/2):86-98, 1999.
[15] Westerveld T, Image Retrieval: Content versus Context, In *Content-Based Multimedia Information Access (RIAO 2000)*, pp 276-284, Paris, France, 2000.
[16] Rosset A, Müller H, Martins M, Dfouni N, Vallée JP, Ratib O, Casimage Project–a digital teaching files authoring environment, *Journal of Thoracic Imaging* 19(2):1-6, 2004.
[17] P. Clough, M. Sanderson, H. Müller, A proposal for the CLEF cross language image retrieval track (imageCLEF), In *Proceedings of the Challenge of Image and Video Retrieval (CIVR 2004)*, *Springer Lecture Notes in Computer Science*, Dublin, 2004.
[18] Ruch P, Chichester C, Cohen G, Coray G, Ehrler F, Ghorbel H, Müller H, Pallotta V. Report on the TREC 2003 Experiment: Genomic Track, *TREC 2003*, Gaitherburg, USA, 2004.
[19] Zweigenbaum P, Baud R, Burgun A, Namer F, Jarrousse E, Grabar N, Ruch P, Le Duff F, Thirion B, Darmoni S. UMLF: A Unified Medical Lexicon for French. *Proc AMIA Symp*, Washington DC, USA, 2003.
[20] Salton G, Buckley C, Term weighting approaches in automatic text retrieval, *Information Processing and Management* 24(5):513-523, 1988.
[21] Rocchio JJ, Relevance feedback in information retrieval. In *The Smart Retrieval System, Experiments in Automatic Document Processing*, Prentice Hall, 313-323, 1971.