ImageCLEF 2004: Combining Image and Multi-lingual Search for Medical Image Retrieval

Henning Müller[1], Antoine Geissbühler[1], and Patrick Ruch[1,2]

[1] University and University Hospitals of Geneva, Service of Medical Informatics,
21 Rue Micheli-du-Crest, CH-1211 Geneva 4, Switzerland
{henning.mueller, antoine.geissbuhler, patrick.ruch}@sim.hcuge.ch
[2] Swiss Federal Institute of Technology, LITH
IN-Ecublens, CH-1015 Lausanne, Switzerland

Abstract. This article describes the technologies used for the various runs submitted by the University of Geneva in the context of the 2004 ImageCLEF competition. As our expertise is mainly in the field of medical image retrieval, most of our effort was concentrated on the medical image retrieval task. Described are the runs that were submitted including technical details for each of the single runs and a short explication of the obtained results compared with the results of submissions from other groups. We describe the problems encountered with respect to optimising the system and with respect to finding a balance between weighting textual and visual features for retrieval. A better balance seems possible when using training data for optimisation and with relevance judgements being available for a control of the retrieval quality.

The results show that relevance feedback is extremely important for optimal results. Query expansion with visual features only gives minimal changes in result quality. If textual features are added in the automatic query expansion, the results improve significantly. Visual and textual results combined deliver the best performance.

1 Introduction

The goals of ImageCLEF are in the field of cross–language information retrieval. From our point of view, this is of extremely high importance for a country such as Switzerland with four official languages and equally within the European Union with an even larger variety. CLEF has been held since 2000 as an independent workshop, always following the European conference on digital libraries (ECDL). 2003 saw the first ImageCLEF conference [1] and all submitted runs took into account the textual but not the visual data of the images supplied. The goal of the 2004 conference was clearly to create an image retrieval task with a realistic outline description that needs a visual component in addition to the textual multi–lingual part. The medical image retrieval task is such a (at least partly) realistic task where a medical doctor has produced one or several image(s) and likes to get evidence for or against a certain diagnosis. Ground truthing can, for now, not be on a diagnosis basis as the image dataset contains mainly one

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 718–727, 2005.
© Springer-Verlag Berlin Heidelberg 2005

example per diagnosis. Still, a task is born with a visual query being a starting point [2]. Relevant documents were in this case images that show the same anatomic region, were taken with the same modality, from the same viewing direction and the same radiologic protocol if applicable (for example, contrast agent or not, T1 vs. T2 weighting when using the MRI). In [3], the main ideas for the 2004 task are described. The data for the task were taken from a medical case database called *casImage*[3] [4]. The database contains almost 9000 images from 2000 medical cases. Images are annotated in XML format but very rudimentary and not at all controlled with respect to quality or fields that have to be filled in. About 10% of the records do not contain any annotation. A majority of the documents are in French (70%), with around 20% in English.

In this paper we will mainly discuss the un–interpolated mean average precision of every run that we submitted as this measure was used for the official ranking of systems. Other measures might change the ranking of systems and might be more appropriate for certain tasks.

2 Basic Technologies Used

For our first ImageCLEF participation, we aim at combining content-based retrieval of images with cross–language retrieval applied on textual case reports. Considering that benchmarks are not available, investigating such a combination is challenging in itself. Once training data is available, systems can be optimised.

2.1 Image Retrieval

The technology used for the content–based retrieval of medical images is mainly taken from the *Viper*[4] project of the University of Geneva. Much information about this system is available [5]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[5]. This software tool is open source and can in consequence also be used by other participants of ImageCLEF. A ranked list of visually similar images for every query task was made available for participants and will serve as a baseline to measure the quality of submissions. Demonstration versions of gift were made available for participants to query visually as not everybody can be expected to install an entire Linux tool for such a benchmark. The feature sets that are used by GIFT are:

- Local colour features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode colour of each region as a descriptor;
- global colour features in the form of a colour histogram, compared by a simple histogram intersection;

[3] `http://www.casimage.com/`
[4] `http://viper.unige.ch`
[5] `http://www.gnu.org/software/gift/`

- local texture features by partitioning the image and applying Gabor filters in various scales and directions. Gabor responses are quantised into 10 strength;
- global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

A particularity of *GIFT* is that it uses many techniques from text retrieval. Visual features are quantised and the feature space is very similar to the distribution of words in texts, corresponding to a Zipf distribution. A simple *tf/idf* weighting is used and the query weights are normalised by the results of the query itself. The histogram features are compared based on a histogram intersection.

The medical version of the *GIFT* is called *medGIFT*[6] [6]. It is also accessible as open source and adaptations concern mainly visual features and the interface that shows the diagnosis on screen and is linked with a radiologic teaching file so the MD can not only browse images but also get the textual data and other images of the same case. Grey levels play a more important role for medical images and their numbers are raised, especially for relevance feedback (RF) queries. The number of the Gabor filter responses also has an impact on the performance and these are changed with respect to directions and scales.

2.2 Textual Case Report Search

The basic granularity of the casimage collection is the case. A case gathers a textual report, and a set of images. Because the original queries are images, textual case–based retrieval is used for feedback only.

Indexes. Textual experiments were conducted with the easyIR engine[7]. As a single report is able to contain both French and English written parts, it would have been necessary to detect the boundaries of each language segment. Ideally, French and English textual segments would be stored in different indexes. Each index could have been translated into the other language using a general translation method, or more appropriately using a domain-adapted method [7]. However, such a complex architecture would require to store different segments of the same document in separate indexes. Considering the lack of data to tune the system, we decided to index the casimage collection using a unique index: 1) using an English stemmer, 2) using a French stemmer. We use 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 the TREC Genomics track [8]. For English, 120'000 string variants were extracted from UMLS, while the French thesaurus contains about 6'000 entries. Both resources were merged for the experiments. Our submitted runs were produced using the English index without specific translation.

Weighting Schema. Because queries were not provided, a generally good weighting schema of the term frequency - inverse document frequency family.

[6] `http://www.sim.hcuge.ch/medgift/`
[7] `http://lithwww.epfl.ch/~ruch/softs/softs.html`

Following weighting convention of the SMART engine, cf. Table 1, we used atc-ltn parameters, with $\alpha = \beta = 0.5$ in the augmented term frequency.

Table 1. Usual *tf-idf* weight; for the cosine normalisation factor, the formula is given for Euclidean space: $w_{i,j}$ is the document term weight, $w_{j,q}$ is the query term weight

Term Frequency	
First Letter	$f(tf)$
n (natural)	tf
l (logarithmic)	$1 + log(tf)$
a (augmented)	$\alpha + \beta \times (\frac{tf}{max(tf)})$, where $\alpha = 1 - \beta$ and $0 < \alpha < 1$
Inverse Document Frequency	
Second Letter	$f(\frac{1}{df})$
n(no)	1
t(full)	$log(\frac{N}{df})$
Normalisation	
Third Letter	$f(length)$
n(no)	1
c(cosine)	$\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{j,q}^2}$

2.3 Combining the Two

Combinations of visual and textual features are rather scarce in the literature [9], so many of the mechanism and fine tuning of the combinations will need more work. As the query is an image only, we had to use some automatic mechanism to expand the query to text. We use automatic query expansion (QE) to the first and the first three images retrieved visually. The text of the case report of these images was taken as free text for the query. XML tags of the casimage files were removed and unnecessary fields such as MD name or date of the entry were removed. All terms in the case reports were used. These free–text queries deliver a ranked list of cases and their similarity score. This score was normalised by the highest score available to have a result within $[0; 1]$. Afterwards, the similarity score is extended from the case to all the images that are part of the case. This includes a high number of visually very dissimilar images that just appear on the same case but are from a differing modality. Afterwards, visual and textual result list are merged. Such a list might not contain all images but at least images that have similarity in the visual and in the textual part will be ranked highly. Problem is to find a balance between the visual and the textual component. In our experience, the visual part needs to be ranked higher than the textual part as the task description is based on visually similar images but the textual part does improve the final results significantly. Manual RF is another tool that improves the results very strongly.

3 Runs Submitted for Evaluation

This section gives an introduction to the techniques and variations used for our submitted runs and their identifications.

3.1 Only Visual Retrieval with One Query Image

For the visual queries, the *medGIFT* system was used that allows an easy change of system parameters such as the configuration of Gabor filters and grey level/colour quantisations. Input for these queries were the query images. The following system parameters were submitted:

- 18 hues, 3 saturations, 3 values, 4 grey levels, 4 directions and 3 scales of the Gabor filters, the GIFT base configuration made available to all participants; *(GE_4g_4d_vis)*
- 9 hues, 2 saturations, 2 values, 16 grey levels, 8 directions and 5 scales of the Gabor filters; *(GE_16g_8d_vis)*
- 9 hues, 2 saturations, 2 values, 16 grey levels, 4 directions and 5 scales of the Gabor filters. *(GE_16g_4d_vis)*

It is hard to actually analyse visually and without ground truth, which of the runs performed best. The three runs were submitted as a trial and because previous results suggest that a small number of grey levels performs better, especially within the first few images retrieved. Studies show that a larger number of grey levels performs better for feedback with a larger number of input images [10]. The ImageCLEF results show that the best of the visual runs is the *GIFT* base system that uses only 4 grey levels, 3 scales and four directions of the Gabor filters (mean average precision (MAP) **0.3757**). Much worse is the system when using 16 grey levels and five scales (**0.2998**). We have to test whether the five scales have a strong influence on these results. When using five scales, 16 grey levels and 8 directions instead of four, the results get better (**0.3084**)

3.2 Visual Retrieval with Automatic Query Expansion

This section describes QE, automatically feeding back the query image and the 1 or 3 best images retrieved in a first step. Manual observations showed that the first few images seem to be very similar in most configurations. Only a few queries did not turn up visually relevant images as the first response. Thus, we attempted to feed back the first retrieved image as feedback with the initial query image. In a second try we submitted the first three retrieved images automatically which can contain more information but has also a higher risk of error as non-relevant images can be used. When wrong images are used in the expansion, the results risk becoming much worse. The submitted runs are a mixture of these containing one quantisation with 1 and 3 images fed back and two other quantisation with only 1 image fed back. The runs submitted were not analysed for performance beforehand thus the selection of the submitted runs was arbitrary.

- 8 directions, 16 grey levels, 1 image fed back *(GE_8d_16g_qe1.txt)*;
- 4 directions, 16 grey levels, 1 image fed back *(GE_4d_16g_qe1.txt)*;

– 4 directions 16 grey levels 3 images fed back *(GE_4d_16g_qe3.txt)*;
– normal gift system with 1 image being fed back *(GE_4d_4g_qe1.txt)*;

Results show that with automatic QE the best results are again obtained with the standard gift system (MAP **0.3727**). This is actually not as good as the results without QE. When using 16 grey levels the results do slightly improve over the first query step when feeding back 1 (MAP **0.3020**) but not with 3 (MAP **0.2977**) query images. Results are almost unchanged between expansion with 1 and 3 images. The system with 8 directions and 16 grey levels improves stronger than with only four directions (**0.3126**). This seems to underline the idea that a small number of grey levels is much better in the first query step but with expansion it is better to have more information on the images in form of grey levels and Gabor filter responses.

3.3 Visual Retrieval with Manual Feedback

This RF part was performed in a manual way with the same three quantisations as were used in the one–shot queries. Only difference is that a user was retrieving the first 20 images for every query and performed manual RF for 1 step. We would have liked to have an evolution over several steps to show how much RF can do and when a saturation is expected, but finally this was not attempted due to a lack of expert resources to perform the manual feedback. The person performing the RF does not have a medical education and some errors with respect to the feedback might be due to wrong images being be fed back.

– *(GE_4d_4g_rf)*;
– *(GE_4d_16g_rf)*;
– *(GE_8d_16g_rf)*.

The result images from the first query step were taken to query the system and observe the first 20 results for the run. Positive and negative images were marked for feedback to optimise the system response. A few images were marked as neutral when they were regarded as irrelevant but visually similar to the correct images or when the feedback person was not sure about the relevance of the image.

The results show that the performance difference between a small number of grey levels and a larger number is reduced when using RF. Still, the *GIFT* base systems stays the best in the test (**0.4469**). Worst RF system is the system with 16 grey levels and four directions (**0.3791**). Most improved system is the one with 16 grey levels but 8 directions (**0.3921**). RF shows its enormous potential and importance for visual information retrieval as the results improve significantly with its use. Taking a larger number of feedback images, an expert feedback person and several steps of feedback can further improve results.

3.4 Visual and Textual Multi-lingual Retrieval, Automatic Run

This combination uses the same automatic QE based on the images retrieved with the *medGIFT* system. The first 1 or 3 images added for visual QE are also

used for the textual query. The text from these images was cleaned from the XML tags of the casimage case notes and unnecessary fields. ACR codes are equally deleted as they are currently not translated into their correct textual description which could be an important help for textual indexing and retrieval.

The remaining text was submitted to the *easyIR* system. The result list contains the most similar case notes with respect to the text and a weighting. This weighting was normalised based on the highest weighting in the list to get values in $[0; 1]$. Afterwards, all images of the case receives the value of the case, thus containing visually similar and very dissimilar images. A total of 200 cases was retrieved, which results in a list of 800–1000 images containing a similarity value.

Merging of visual and textual results was done in various ways. As the unit for retrieval and similarity assessment is the image, the visual similarity plays an important role. Textual similarity might be better with respect to the semantics of the case but a case contains relevant and also many irrelevant images that are in the same case but of a different modality. Thus, visual similarity had to be weighted higher than textual similarity, so visually non-similar images were not weighted higher than visually similar but textually dissimilar ones. We were not sure to have correct cases in the first $N = 1..3$ images so care is be important to not expand the query into a wrong direction. Three runs were submitted using 75% visual and 25% textual retrieval:

- 4 directions, 16 grey levels, visual/textual with QE 1 image; *(GE_4d_16g_vt1)*
- 4 directions, 16 grey levels, visual/textual with QE 3 images; *(GE_4d_16g_vt3)*
- 4 directions, 4 grey levels, visual/textual with QE 1 image; *(GE_4d_4g_vt1)*

Another run was submitted with a ratio of 80% for the visual and 20% for the textual features:

- *(GE_4d_4g_vt2).*

Another idea was based on the fact that most visually important images should be within the upper part of the visually similar images retrieved. This means that the goal should be to augment the value of those in the list of the visually similar that also appear in the list of the textually similar. For this run we multiplied the score of all those images that were within the first 200 cases retrieved textually and within the first 1000 images visually by a factor of 1.5. The resulting series has the tag:

- *(GE_4d_16g_vtx).*

Evaluation results show that the use of textual information significantly improves the retrieval, also when using the text of a single image, for example with 16 grays and four directions (MAP **0.3361**). This is an improvement of 0.036 and thus more than 10% better than the visual QE with one image. When executing QE with 3 images, the results improve slightly less strong (MAP **0.3322**).

Better results were again obtained when using only 4 grey levels. When feeding back one image, the MAP is **0.3901** and thus better than all other submitted

automatic runs in the competition. Best results in our test were obtained when changing the weighting between visual and textual features from 25% to 20% which delivered a MAP of **0.4020**. The selective weighting change for images that were visually similar and that appeared in the top retrieved by text also delivered very good results (MAP **0.3940**).

When analysing these results, we think that feeding back (maybe more) images with text using a 20% weighting can produce better results. We think that the optimal result with this technique can be in the range of a MAP of **0.45**.

3.5 Visual and Textual Multi-lingual Retrieval, Manual Feedback

As we do not yet have an integrated interface of visual and textual search engines, these results are based on manual RF queries based on the visual retrieval results, only. Based on images marked relevant after a first visual query step a query was constructed. For the textual query, only positive documents were taken whereas for the visual part positive and negative images were taken into account. The text was generated in the same way as before by adding case notes without names, dates, XML tags and ACR codes into one large file. If there were several images of the same case, the text was copied several times. These texts were submitted to the easyIR system. Again, the resulting list of case results and scores was normalised. The system we use employs 16 grey levels and 4 directions and thus the worst system in a first visual result as well as the worst in visual feedback, which we did not know at that point. Still, the textual component alone improves the results significantly. For the visual query, the results were equally normalised to a range between 0 and 1. For merging the results we used three different ratios between visual and textual characteristics:

- 25% textual, 75% visual; *GE_rfvistex1*
- 20% textual, 80% visual; *GE_rfvistex20*
- 10% textual, 90% visual; *GE_rfvistex10*

At this point we were sure that the text contains relevant information and not only automatically expanded case texts. Still, it is important to not have a too strong influence of the textual features as they are on a case and not an image basis whereas the gold standard is generated based on an image basis. The gradient of similarity within the textual results list is much higher than within the visual result list which explains part of the risk of too strongly weighing the textual features.

Results show that our RF results are by far the best results in the entire competition. Best results are obtained when combining the results by 10% textual and 90% visual (**0.4847**). When weighing the textual features high (25 %) the results drop significantly (**0.4520**). When weighting the textual features at 20%, the results drop in performance but only very slightly (**0.4827**). This suggests that the optimal weighting in our case should be around the 10%. Tests will have to confirm whether it is lower or higher than the 10Having the gift base system with 4 grey levels for this run would also improve retrieval quality as the query results seem to be much better in a first query step. Optimal results might be

obtained with 4 grey in a first query steps and then several grey levels for the feedback.

4 Further Ideas That Are Currently not Explored

The ACR codes should be translated into text for better indexing and retrieval. They contain valuable information and are part of several case notes. We currently do not use ACR codes attached to cases at all.

Image normalisation should be applied to avoid that images which lye in a different grey spectrum are not properly retrieved. Currently, this can be the case quite often as there is no control on the level/window settings for a medical doctor when inserting images. Images are in JPEG and so information from the original DICOM images might have got lost. Other pretreatment steps such as background removal can also improve results.

Using a gradient of the similarity scores to define how many of the first N images might be relevant and could be sent back as automatic QE is another promising idea. This can allow a more reasonable way to choose images for automatic QE. Currently, the values used are fairly conservative as a wrong QE can delete the quality of retrieval completely.

Work is also needed with respect to quantisations of the feature space. Currently, a surprisingly small number of grey levels leads to best results but we need to analyse which queries were responsible for this and which other factors such as directions, scales and quantisations of Gabor filters might play a vital role.

5 Conclusions

We had a lack of manpower to do a proper adaptation and evaluation of the parameters that we could use within our system. Thus we could not use the software tools up to their full potential. Especially the use of RF over several steps is expected to lead to a much better performance. The use of ground truth data to optimise the system will also lead to much better results. For further ImageCLEF competitions it is expected to have training data accessible before the conference, and a different database during the conference. There was also a lack of experience with combining textual and visual features for retrieval. Many ideas can be performed for this combination to optimise retrieval results.

The most important conclusions for the evaluation of our system are surely:

- a surprisingly small number of grey levels led to best results in a first query step;
- QE for visual retrieval does not change the performance;
- a larger number of grey levels is better for RF;
- textual features improve performance with automatic QE as well as with manual RF;

- RF improves results enormously and remains a power tool for information retrieval;
- RF and visual/textual combinations led to the best overall results in the ImageCLEF competition;
- there is still a lot to be tried out!

This leaves us with several important outcomes and many ideas to prove now that the ground truth is available. Next year's ImageCLEF will for sure deliver much better systems and more optimised solutions based on training data.

References

1. Clough, P., Sanderson, M.: The clef 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003). (2004)
2. Müller, H., Rosset, A., Geissbuhler, A., Terrier, F.: A reference data set for the evaluation of medical image retrieval systems. Computerized Medical Imaging and Graphics (2004 (to appear))
3. Clough, P., Sanderson, M., Müller, H.: A proposal for the clef cross language image retrieval track (imageclef) 2004. In: The Challenge of Image and Video Retrieval (CIVR 2004), Dublin, Ireland, Springer LNCS (2004)
4. Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallée, J.P., Ratib, O.: Casimage project – a gidital teaching files authoring environment. Journal of Thoracic Imaging **19** (2004) 1–6
5. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content–based query of image databases: inspirations from text retrieval. Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99) **21** (2000) 1193–1198 B.K. Ersboll, P. Johansen, Eds.
6. Müller, H., Rosset, A., Vallée, J.P., Geissbuhler, A.: Integrating content–based visual access methods into a medical case database. In: Proceedings of the Medical Informatics Europe Conference (MIE 2003), St. Malo, France (2003)
7. Ruch, P.: Query translation by text categorization. In: Proceedings of the conference on Computational Linguistics (COLING 2004), Geneva, Switzerland (2004)
8. Ruch, P., Chichester, C., Cohen, G., Coray, G., Ehrler, F., Ghorbel, H., Müller, H., Pallotta, V.: Report on the trec 2003 experiment: Genomic track. In: Proceedings of the 2003 Text REtrieval Conference (TREC), Gaithersburg, MD, USA (2004)
9. La Cascia, M., Sethi, S., Sclaroff, S.: Combining textual and visual cues for content–based image retrieval on the world wide web. In: IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98), Santa Barbara, CA, USA (1998)
10. Müller, H., Rosset, A., Vallée, J.P., Geissbuhler, A.: Comparing feature sets for content–based medical information retrieval. In: Proceedings of the SPIE International Conference on Medical Imaging, SPIE Vol. 5371, San Diego, CA, USA (2004)