Report on the CLEF Experiment: Combining Image and Multi–lingual Search for Medical Image Retrieval

Henning Müller[1], Antoine Geissbühler[1] and Patrick Ruch[12]
[1]University and University Hospitals of Geneva, Service of Medical Informatics
21 Rue Micheli-du-Crest, CH-1211 Geneva 4, Switzerland
[2]Swiss Federal Institute of Technology, LITH
IN-Ecublens, CH-1015 Lausanne, Switzerland
henning.mueller@sim.hcuge.ch

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, we will concentrate most of our effort on the medical image retrieval task.

Described are the runs that were submitted by our group including technical details for each of the single runs and a short explication of the obtained results, also compared with the results of submissions from other research groups. We will also describe the problems encountered with respect to optimising the system and especially with respect to finding a balance between weighting the textual and visual features for retrieval. A much better balance seems possible when using some training data for optimisation and with the relevance judgements being available for a control of the respective 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, then the results improve significantly. Visual and textual results combined deliver the best results.

1 Introduction

The goals of imageCLEF are mainly 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 of the systems that were used to submit 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 would need a visual component in addition to the textual multi–lingual part. The medical image retrieval task is such a realistic task where a medical doctor has produced one or several image(s) and would like to get evidence for or against a certain diagnosis. Ground truthing can, for now, not be on a diagnosis basis as the image dataset is not specialised enough for this. Still, a task was born with a visual query being a starting point as also described the following document [3]. 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 [2], the main ideas for the 2004 task are described. The data for the task were taken from a medical case database called *casImage*[1] [6]. This database contains almost 9000 images from more than 2000 medical cases. Images contain annotation in XML format but these annotations are very rudimentary and are not at all controlled with respect to quality or fields that have to be filled in. About 10% of the records to not contain any annotation. A majority of the documents are in French (70%), with around 20% being in English.

Figure 1 shows a few example images from the database. These images are among others the query topic for the performance evaluation.

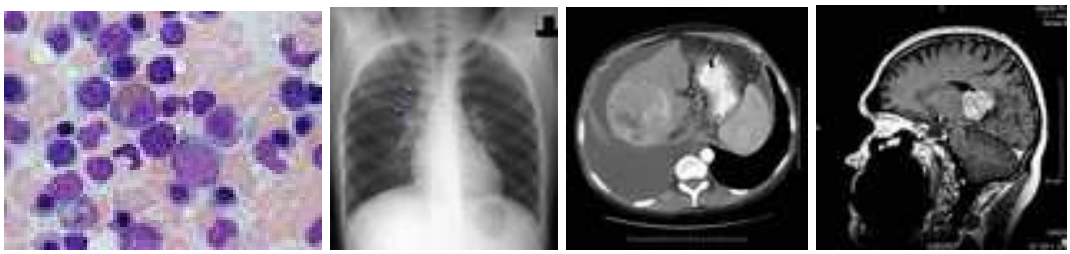

Figure 1: Examples from the casimage collection.

In total 26 images were chosen by a radiologist as query topics. These images were chosen to well represent the image set with respect to modalities used, anatomic regions represented and radiologic protocols used. A large number of images in the database was expected to not be represented by these 26 topics.

Relevance judgements were done by a total of three judges per query. Images could be judged as relevant, partially relevant or non relevant. This results in 9 sets of relevance judgements showing the overlap of all relevant judgements, intersection of relevant judgements, overlap and intersection of relevant and partially relevant. Due to a significant difference in the relevance judgements the principal evaluation was performed with a relevance set that was obtained by including all images that were judge as relevant by at least two of the judges. All data is available from the imageCLEF web sides.

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 different query tasks.

2 Basic technologies used

For our first participation at imageCLEF, we aim at combining content-based retrieval of images with cross–language retrieval applied on case reports. Considering that benchmarks are not available at the time of submission, investigating such a combination is challenging in itself, so that our study is clearly at a preliminary stage. Once training data is available, systems can be optimised for certain query tasks.

2.1 Image Retrieval

The technology used for the content–based retrieval of medical images is mainly taken from the *Viper*[2] project of the University of Geneva. Much information about this system is available [9]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[3]. This software tool is open source and can consequently also be used by other participants of imageCLEF. A ranked

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

list of visual similarity for every query task was made available for participants and will server as a baseline for the quality of submissions. Demonstration versions of a gift server were made available for participants to query visually as well 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 multi–scale descriptor;
- global colour features in the form of a colour histogram, compare by a simple histogram intersection;
- 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 the responses of the local Gabor filters in various directions and scales and with various strength.

A particularity of *GIFT* is that it uses many techniques from text retrieval. Visual features are quantised and the open a feature space that is very similar to that opened by words in texts. The distribution of feature frequency corresponds more or less 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 calculated based on a simple histogram intersection.

The medical version of the *GIFT* is called *medGIFT*[4] [4]. It is also accessible as open source and adaptations concern mainly the features used and the interface that shows the diagnosis on screen and that 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 role for medical images and their numbers are raised, especially for relevance feedback queries. The number and sort of the Gabor filter responses also has an impact on the performance and these are changed with respect to the number of directions and scales.

2.2 Textual case report search

The basic granularity of the casimage collection is the case. A case is usually gathering a textual report, which describes the case, with appropriate bibliographic references, and a set of images. Because the original queries are images, only, textual case–based retrieval is used for relevance feedback. Experiments were conducted with the easyIR engine[5]. As a single report is able to contain both French and English written parts, we decided to index the casimage collection using two different indexes: 1) using an English stemmer, 2) using a French stemmer. We used the Porters stemmer for the English and a modified version of Savoy's tool for the French language. For each index a list of stop words was used: 544 items for English, 792 for French. We also used a biomedical thesaurus, which has proven some effectiveness in the context of the TREC Genomics track [8]. For the English, 120'000 string variants were extracted from the UMLS, while the French thesaurus contains about 6000 entries. Although we tested different translations [7] and index combination strategies, our submitted runs were produced using the English index without specific translation. Finally, textual relevance feedback was also disabled.

2.3 Combining the two

As the query is an image only, we had to use some automatic mechanism to expand the query to text. In our case we used automatic query expansion of the first and the first three images. These images were analysed and the text of the case report 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.

[4] `http://www.sim.hcuge.ch/medgift/`
[5] `http://lithwww.epfl.ch/~fuch/softs/softs.html`

The free–text queries deliver a list of result cases and their similarity score. This similarity score was normalised by the highest similarity score available as it is already the case for the visual queries. Afterwards, the similarity score is transferred from the case to all the images that are part of this case. This includes a high number of visually very dissimilar images that just appear on the same case. Afterwards, visual and textual results are merged in part. This list might not contain all the images but at least images that have some similarity in 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 but the textual part does improve the final results significantly. Relevance feedback is another tool that improves the results very strongly.

3 Runs submitted for evaluation

This section gives a basic introduction into the techniques and variations used for our runs submitted to imageCLEF. It also contains the IDs for the runs that were finally submitted and that are evaluated in this paper.

3.1 Only visual retrieval with one query image

For the visual queries, the *medGIFT* system was used. This system allows a fairly easy change of a few system parameters such as the configuration of the Gabor filters and the grey level and colour quantisations. Input for these queries were only the query images. No feedback or automatic query expansion was used. 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 of imageCLEF; *(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 very 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 might be better for feedback queries with a larger number of input images [5].

The imageCLEF results finally show a slightly different picture: 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) **0.3157**). Much worse is the system when using a smaller number of colours but 16 grey levels and five scales (**0.2565**). It will have to be tested whether the five scales have an influence on these results. When using five scales, 16 grey levels and 8 directions instead of four, the results get better again (**0.2649**)

3.2 Visual retrieval with automatic query expansion

This section uses very simple query expansion, feeding back the query image and the 1 or three best images retrieved in a first query step. Some manual observations showed that the first few images seem to be very similar in most cases. Only a few queries did not turn up visually similar 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 might contain more information but has also a higher risk of error. When wrong images are used in the query expansion, the results have the risk of becoming much worse. The runs that we submitted are a mixture of these containing one quantisation with one and three images fed back

and another quantisation with only one image fed back. A forth test runs was submitted as well. The runs submitted where not analysed for their performance thus the selection of the submitted runs was more on personal intuition.

- 8 directions, 16 grey levels, one image fed back *(GE_8d_16g_qe1.txt)*;
- 4 directions, 16 grey levels, one image fed back *(GE_4d_16g_qe1.txt)*;
- 4 directions 16 grey levels three images fed back *(GE_4d_16g_qe3.txt)*;
- normal gift system with one image being fed back *(GE_4d_4g_qe1.txt)*;

The results show that with automatic query expansion the best results are again obtained with the standard gift system (MAP **0.3100**). This is actually not as good as the results without query expansion. When using 16 grey levels the results do very slightly improve over the first query step when feeding back 1 (MAP **0.2593**) and three (MAP **0.2586**) query images. The results are almost unchanged between expansion with one and three images. The system with 8 directions and 16 grey levels does improve the results stronger than with only four directions (**0.2704**). 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 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 relevance feedback for 1 step, only. We would have liked to have an evolution over several steps to show how much relevance feedback can do and when a saturation would be expected, but finally this was not attempted due to a lack of resources to perform the manual feedback. The person performing the relevance feedback does not have a medical education and some errors with respect to the feedback might be due to this as wrong images might 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 a 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. It was feared that this could make a relevance feedback query less good than the initial query.

The results show that the performance difference between a small number of grey levels and a larger number is reduced when using relevance feedback. Still, the *GIFT* base systems stays the best one in the test (**0.3791**). Worst relevance feedback system is the system with 16 grey levels and four directions (**0.3259**). Most improved system is the one with 16 grey levels but 8 directions (**0.3380**). Relevance feedback shows its enormous potential and important for visual information retrieval as the results improve significantly with the use of feedback. Taking a larger number of feedback images, an expert feedback person and also several steps of feedback has surely the potential to further improve results.

3.4 Visual and textual multi-lingual retrieval, automatic run

This combination run uses the same automatic query expansions that are based on the images retrieved with the *medGIFT* system. The first one or three images that were added for query expansion were as well used for the textual query. The text from these images was cleaned from the XML tags of the casimage case notes and unnecessary fields such as dates and the treating MD were also deleted. ACR codes are equally deleted as they are currently not translated into their correct textual description which could be an important help for the textual indexing and retrieval.

The remaining text was submitted to the *easyIR* system. We tried out both, a French and an English version but finally, only used the English version as the results did not seem to be significantly different. Making a selection between English and French case notes and thus having two indices might make a difference with respect to the results. The results list from easyIR 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 between 0 and 1. Afterwards, all images in these case notes received the value of the entire case, thus containing visually similar and very dissimilar images. A total of 200 cases was retrieved which results in a list of 800–1000 results images containing a similarity value.

The merging of the 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 really sure to have correct case in the first $N = 1..3$ images so some care might be important to not expand the query into a completely wrong direction.

Three runs were submitted using 75% visual and 25% textual retrieval:

- 4 directions, 16 grey levels, visual/textual with query expansion from one image; *(GE_4d_16g_vt1)*
- 4 directions, 16 grey levels, visual/textual with query expansion from three images; *(GE_4d_16g_vt3)*
- 4 directions, 4 grey levels, visual/textual with query expansion from one 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 being 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 simply multiplied 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, even when only using the text of a single image as in the case of 16 grays and four directions (MAP **0.2935**). This is an improvement of 0.035 and thus more than 10% over the visual query expansion with one image. When executing query expansion with 3 images, the results even improve much more (MAP **0.3370**) and are among the best automatic runs that were submitted for the competition. This is surprising as the visual query expansion with 3 images was actually worse than with 1 image and it also only improve results slightly.

Better results were again obtained when using 4 grey levels. When feeding back one image, the MAP is **0.3611** and thus better than all other submitted automatic runs. 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 **0.3749**. The selective weighting change for image that were visually similar and that appeared in the top retrievals by text also delivered very good results (MAP **0.3612**).

When analysing these results, we think that when feeding back more images with text and using a 20% weighting we could get even much better results than what we have received so far. When comparing with the change between one and three qe-images with 16 grey levels, we think that the improvement can be in the range of a MAP of **0.40**.

3.5 Visual and textual multi-lingual retrieval, manual feedback

As we do currently not have an integrated interface of our visual and textual search engines these results are based on the manual relevance feedback queries with the visual retrieval results, only. Based on the documents 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 the 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 to 1 and the expanded from cases to images. The system we used for this runs was the one with 16 grey levels and 4 directions and thus the worst system in a first visual results as well as the first in visual feedback. Still, the textual component alone improves the results significantly.

For the visual query, positive and negative feedback were taken into account. 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 at least sure that the text does contain relevant information and not 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.

The results show that the relevance feedback results are by far the best results in the entire competition. Best results are obtained when combining the results by 20% textual and 80% visual (**0.4214**). When higher weighing the textual features (25 %) the results drop significantly (**0.3824**). When lower weighting the textual features, the results drop in performance but only slightly (**0.4189**). This suggests that the optimal weighting in our case will be in between the 10 and 20% area. Having the gift base system for this run would also be an interesting option as the query results seem to be much better in a first query step.

4 Further ideas that are currently not explored

The ACR codes should be translated into text for better indexing and retrieval. They contain very valuable information and are in several case notes. We currently do not use the ACR codes that are attached to some of the images 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.

Using a gradient of the similarity scores to define hon many of the first N images might be relevant and could be sent back as automatic query expansion is another promising idea. This can allows us a more reasonable way to choose images for automatic query expansion. Currently, the values that we use are fairly conservative as a wrong query expansion can delete the quality of retrieval completely.

Work work will also need to be done with respect to quantisations of the feature space. Currently, a surprisingly small number leads to best results but it will have to be analysed 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. Thus we could not use the software tools up to their perfect performance. Especially the use of relevance feedback over several steps is expected to lead to a much better performance. The use of some ground truth data to optimise the system will also for sure lead to much better results. For further imageCLEF 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 are surely:

- a surprisingly small number of grey levels led to best results in a first query step;
- query expansion for visual retrieval does not change the performance much;
- a larger number of grey levels is better for relevance feedback;
- textual features improve performance with automatic query expansion as well as with manual relevance feedback;
- relevance feedback improves results enormously and remains a power tool for information retrieval;
- relevance feedback and visual/textual combinations led to the vest overall results in the 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.

References

[1] P. Clough and M. Sanderson. The clef 2003 cross language image retrieval task. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2004)*, 2004 (submitted).

[2] P. Clough, M. Sanderson, and H. Müller. A proposal for the clef cross language image retrieval track (imageclef) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, July 2004. Springer LNCS.

[3] H. Müller, A. Rosset, A. Geissbuhler, and F. Terrier. A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, 2004 (to appear).

[4] H. Müller, A. Rosset, J.-P. Vallée, and A. Geissbuhler. Integrating content–based visual access methods into a medical case database. In *Proceedings of the Medical Informatics Europe Conference (MIE 2003)*, St. Malo, France, May 2003.

[5] H. Müller, A. Rosset, J.-P. Vallée, and A. Geissbuhler. 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, February 2004.

[6] A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib. Casimage project – a gidital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[7] P. Ruch. Query translation by text categorization. In *Proceedings of the conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, August 2004.

[8] P. Ruch, C. Chichester, G. Cohen, G. Coray, F. Ehrler, H. Ghorbel, H. Müller, and V. Pallotta. Report on the trec 2003 experiment: Genomic track. In *Proceedings of the 2003 Text REtrieval Conference (TREC)*, Gaithersburg, MD, USA, 2004.

[9] D. M. Squire, W. Müller, H. Müller, and T. Pun. 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(13-14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.