Report on the imageCLEF Experiment: How to visually retrieve images from the St. Andrews collection using GIFT

Henning Müller[1], Antoine Geissbühler[1] and Patrick Ruch[12]
[1]University and University Hospitals of Geneva, Service of Medical Informatics
24 Rue Micheli-du-Crest, CH-1211 Geneva 14, Switzerland
[2]Swiss Federal Institute of Technology, LITH
IN-Ecublens, CH-1015 Lausanne, Switzerland
henning.mueller@sim.hcuge.ch

Abstract

The imageCLEF task of the Cross Language Evaluation forum has as its main goal the retrieval of images from multi–lingual test collections, or retrieval of images where the query is in a different language than the collection itself. The 2003 imageCLEF task saw no group using the visual information of the images that is inherently language independent. In 2004, this changed and a few groups among them the university hospitals of Geneva are submitting visual runs for the queries.

The query topics are definitely defined in a way that makes visual retrieval extremely hard as pure visual similarity plays a marginal role whereas semantics and background knowledge are extremely important, that can only be obtained from textual captions. This article describes the submission of an entirely visual result set to the task. This article will also define possible improvements for visual retrieval systems with the current data. Most important is Section 4 that explains possible ways to make this query task more appealing to visual retrieval research groups, explaining problems of content–based retrieval and what such a task could do to help overcome the present problems. A benchmarking event is needed for visual information retrieval to lower current barriers in retrieval performance. ImageCLEF can help to define such an event and identify areas where visual retrieval might be better than textual and vice–versa. The combination of visual and textual features together is another important field where research is needed.

1 Introduction

Visual retrieval of images has been an extremely active research area for more then ten years now [5, 16]. Still, there has not been neither a benchmarking event nor the use of standard datasets to compare the performance of several systems or techniques. Despite efforts such as the Benchathlon[1] [6] and several articles on evaluation [8, 11, 12, 17], no common framework has been created, yet. This is different in textual information retrieval where several initiatives such as TREC[2] [7] (Text REtrieval conference) and CLEF[3] [15] (Cross Language Evaluation Forum) exist. In 2003, CLEF added a cross language image retrieval task [1] using a collection of historic photographs. The task in 2004 uses the same collection but adds an interactive and a medical task [2]. Figure 1 shows a few examples from the St Andrews collection.

Images are annotated in English and query topics are formulated in another language containing a textual description of the query and an example image. English retrieval performance is taken

[1] `http://www.benchathlon.net/`
[2] `http://trec.nist.gov/`
[3] `http://www.clef-campaign.org/`

(a) (b) (c) (d)

Figure 1: Some example images of the St. Andrews database.

as a baseline. The topics for which results can be submitted look as follows (a French example for image 1(a)):

```
<title>
Portraits photographiques de pasteurs d'église par Thomas Rodger
</title>
<narr>
Les images pertinentes sont des portraits photographiques de pasteurs ou
de leaders d'église pris par Thomas Ridger.  Les images de nimporte quelle
époque sont pertinentes, mais ne doivent montrer qu'une personne dans un
studio, c'est-à-dire posant pour la photo.  Des photos de groupes ne sont
pas pertinentes.
</narr>
```

From this topic description we only took the image to start queries with our system, the textual information was discarded. No manual relevance feedback or automatic query expansion was used. This means that important information on the query task has not been obtained. With the visual information only, we do not know that we are searching for church ministers and we do not know who actually took the picture. Only a very good domain expert might be able to get this information from the image alone. Actually, all this information is only findable if the annotation is of a very high quality and is known to be complete. It has to be assured that all images with church ministers have these words in the text, otherwise we can not be sure whether the person is a church minister or might have a similar function. The producer (photographer) of the images also needs to be marked, otherwise a relevance judge would not be able to mark a result as relevant, although two images might be extremely similar in style. What about images where we do not have any name of the photographer but that look very similar to images from "Thomas Ridger"? What about collections with a mediocre text quality such as those that we often find in the real world, for example the Internet?

Some retrieval tasks led to subjectively good results with a visual retrieval system whereas others did not manage to show any relevant images in the top 20 results. Figure 2 shows one example result of a visual retrieval system. The first image is the query image and we can see that the same image was found as well as a few other images with the queen that apparently show the same scene.

Although this might look like a reasonable retrieval results, we can definitely tell that the system had no idea that we were looking for the queen or a military parade. The images were basically retrieved because they have very similar properties with respect to the grey levels contained, and especially with respect to the frame around the image. These images were most likely taken with the same camera and digitised with the same scanner. These properties can be found with a visual retrieval system.

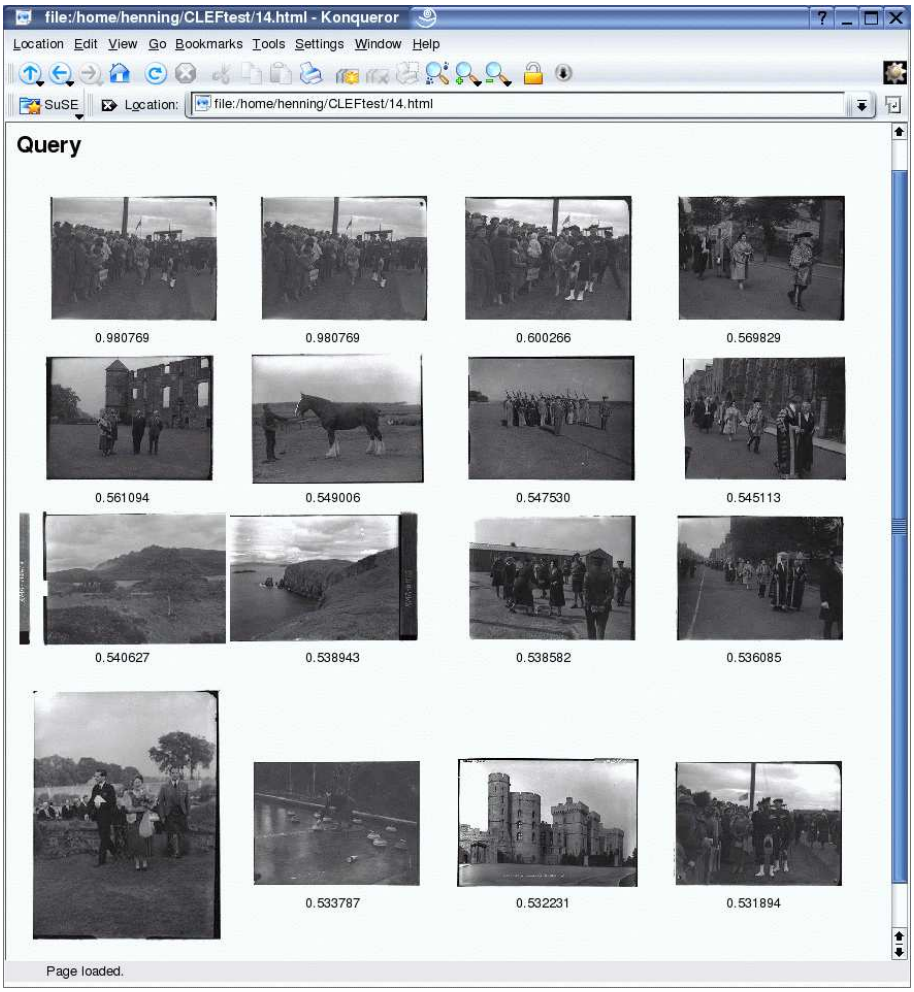

Figure 2: Example for a "good" query result based on visual properties.

2 Basic technologies used for the task

The technology used for the content–based image retrieval is mainly taken from the *Viper*[4] project of the University of Geneva. Much information is available on the system [18]. Outcome of the Viper project is the GNU Image Finding Tool, *GIFT*[5]. We used a version that slightly modifies the feature space and is called *medGIFT*[6] as it was mainly developed for the medical domain. These software tools are open source and can consequently also be used by other participants of imageCLEF. Demonstration versions for participants were made available as well as not everybody can be expected to install an entire Linux tool for such a benchmarking event, only. The feature sets that are used by *medGIFT* are:

- Local colour features at different scales by partitioning the images successively four times into four subregions and taking the mode colour of each region as a feature;
- global colour features in the form of a colour histogram;

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

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

A particularity of GIFT is that it uses many techniques from text retrieval. Visual features are quantised/binarysed, and open a feature space that is very similar to the distribution of words in texts (similar 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. This allows us to apply a variety of techniques that are common in text retrieval to the retrieval of images. Experiments show that especially relevance feedback queries on images are much better using this feature space whereas one-shot queries might be done more performant with other techniques.

3 Runs submitted for evaluation

Unfortunately, there was not enough time this year to submit a mixed visual and textual run for imageCLEF but we are working on this for next year.

3.1 Only visual retrieval with one query image

For the visual queries, the *medGIFT* system was used. This system allows to fairly easy change 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, 4 directions and 5 scales of the Gabor filters. *(GE_16g_4d_vis)*

Some queries delivered surprisingly good results but this was not due to a recognition of image features with respect to the topic but rather due to the fact that images from a relevance set were taken at a similar time and have a very similar appearance. Content–based image retrieval can help to retrieve images that were taken with the same camera or scanned with the same scanner if they are similar with respect to their colour properties. Mixing text and visual features for retrieval will need a fair amount of work to optimise parameters and really receive good results.

The evaluation results show the very low performance of all visual only runs that were submitted. Mean average precision (MAP) is 0.0919 for the GIFT base system and 0.0625 for the modified version. It is actually surprising that the system with only four grey levels performed better than a system having a larger number. Most of the images are in grey and brown tones so we expected to obtain better results when giving more flexibility to this aspect. It will have to be show whether other techniques might obtain better results such as a normalisation of the images or even a change of the brown tones into grey tones to make images better comparable. Still, these results will be far away from the best systems that reach a MAP of 0.5865 such as the Daedalus system suing text retrieval only. Several systems include some visual information into the retrieval and some of these systems are indeed ranked high. All systems that relied on visual features, only, receive fairly bad results, in general the worst results in the competition.

3.2 Techniques to improve visual retrieval results

Some techniques might be of help to further increase the performance of the retrieval results. One such techniques is a pre–processing of images to bring all images to a standard grey level

distribution and maybe removing colour completely. At least the brown levels should be changed to grey levels so images can be retrieved based on real content and not based on general appearance.

Another possibility is the change of the colour space of the image. Several spaces have been analysed with respect to invariance regarding lighting conditions with good results [4]. For the tasks of imageCLEF it might be useful to reduce the number of colours and slightly augment the number of grey levels for best retrieval. Some form of normalisation could also be used as some images used the entire grey spectrum whereas others only use an extremely limited number of grey levels. A proper evaluation will have to show what actually works best.

Mixed visual/textual strategies can lead to a better result. If in a first step only the textual information is taken as a query and then the first N images are visually fed back to the system the results can be much better and can manage to find images that are without text or with a bad annotation and that would not have been found otherwise. More research is definitely needed on mixed textual/visual strategies for retrieval to find out which influence each one can have. It might also be possible to have a small influence of the visually most similar images in a first query step as well but the text will need to be the dominating factor for best results as the query topics are semantics–based.

4 How to make the queries more appealing to visual retrieval research groups?

Although CLEF is on cross–language retrieval and thus mainly on text, image information should exploited in this context for the retrieval of visual data. Images are inherently language–independent and they can provide important additional information for cross–language retrieval tasks. To foster these developments it might even be the best to have an entirely visual task to attract the content–based retrieval community and later come back to a combination of visual/textual techniques. This can also help to develop partnerships between visual and textual retrieval groups to submit common runs for such a benchmark.

Techniques for visual information retrieval are currently not good enough to respond properly to semantic tasks [3]. Sometimes the results look indeed good but this is most often linked to secondary parameters and not really to the semantic concepts being searched for or the low–level features being used.

4.1 More visual information for the current topics

The easiest way to make the St. Andrews cross–language retrieval task more attractive to visual retrieval groups is simply to supply more visual information as task description. Having three to five example images instead of one might help visual retrieval significantly as systems can search for the really important information that these images have in common. A single image for retrieval is a little bit "a shot in the dark" but several images do supply a fair amount of information.

Besides positive examples, an important improvement would be to supply several negative examples to have an idea of what not to look for. Negative relevance feedback has shown to be extremely important in visual information retrieval [10] and feedback with negative examples substantially changes the result sets whereas positive examples only do a slight reordering of the highest–ranked results. Finding three to five negative examples per query task in addition to the positive examples should not be a big problem.

4.2 Topics based on the visual "appearance" of an image

It has been discussed a lot what visual image retrieval cannot do but there are quite a few things that visual image retrieval can indeed do. Although search on semantics seems currently infeasible, similarity based on the appearance of the images can be obtained in a fairly good quality. Visual appearance is often described as a first impression of an image or preattentive

similarity of images [14]. Tasks can also contain fairly easy semantics that are basically modelled by the visual appearance. Possible topics could be:

- Sun sets – modelled by a yellow round object somewhere in the middle and mainly variations of red.
- Mountain views – upper part blue and in the middle sharp changes, in grey/white tones, bottom sometimes/often green.
- Beach – Lower part yellow and the upper part in blue with a clear line between the two.
- City scenes – very symmetric structures with a large number of horizontal lines and right angles.

It will need to be analysed whether these queries do actually respond to what real users are looking for in retrieval systems, but they have the potential to attract a much larger number of visual information retrieval groups to participate and compare their techniques in such a benchmarking event.

4.3 Easy semantic topics

TRECVID[7] introduced in 2003 several topics for video retrieval that can also be used for visual image retrieval, maybe with slight variations. These are fairly easy semantic topics such as finding out whether there are people in images. Some examples for topics are:

- People: segment contains at least three humans.
- Building: segment contains a building. Buildings are walled structures with a roof.
- Road: segment contains part of a road - any size, paved or not.
- Vegetation: segment contains living vegetation in its natural environment.
- Animal: segment contains an animal other than a human .

ImageCLEF could define topics similar in style for the image collections being available (topics that actually do correspond to the images in the collection). Retrieval systems can then try to find as many of the images with respect to the topic as possible based on visual features only or based on visual and textual features. This could also help to find out the influence of text and visual information on fairly low–level semantic concepts.

This can especially stimulate the creation of simple detectors for simple semantic concepts. These detectors can later be combined for the retrieval of higher–level semantic retrieval, so they do deliver important intermediary results.

4.4 An easier image collection

The St. Andrews collection is definitely a very hard collection for purely visual analysis. The images do not contain many clearly separated objects and the small amount of colour pictures and variances in sharpness/quality make automatic analysis extremely hard. Other collections such as the Corel Photo CDs are much easier for automatic analysis and query/retrieval [9]. This collection contains 100 images each for a large number of topics (tigers, planes, eagles, ...). Often the collections have a distinct object in each of the sets, sometimes the sets also correspond to regions (Paris, California, Egypt, ...). Only problem might be to get a collection without to strong copyright constraints. As the Corel Photo CDs are not sold anymore, this might be a possibility if Corel agrees to make the images in a lower resolution available to participants. The Corbis[8] image archive also offers a limited selection of around 15.000 images for research purposes that are annotated in a hierarchical code. Such a collection might be an easier topic for visual and combined visual/textual retrieval.

[7] http://www-nlpir.nist.gov/projects/trecvid/
[8] http://www.corbis.com/

4.5 Interactive tasks evaluated by users

A different idea is the evaluation of interactive systems based on real users performing queries. Normally, image retrieval is not extremely good in a first query step but with feedback, very good results can be obtained [10, 13]. Similar to the interactive task using text introduced in 2004 we can imagine a task with only a visual query description with an example image. Users can subsequently perform queries until they are satisfied with the results. Evaluation could be done directly by the users, for example by counting how many relevant images they found with which system, and how many refinement steps were necessary to find a satisfactory result. It has to be stated that the user satisfaction can vary considerable with respect to his knowledge of the content of the database. When not knowing anything about the total number of relevant images, users tend to be satisfied fairly easily.

5 Conclusions

This article explained a simple submission to the imageCLEF task using the St. Andrews historical image collection. The two submitted runs were based on visual features of the images only, without using the text supplied for the queries. No other techniques were used such as manual relevance feedback or automatic query expansion. The results show the problems of purely visual image retrieval: no semantics are currently included in the visual low–level features and as a consequence the performance is low.

Still, visual information retrieval based on low–level non–semantic features can be an important part in the general information retrieval picture. Visual information retrieval can be used to find images with a similar visual appearance or with simple semantic concepts if learning data for these concepts are available. Thus, it is important for evaluation events such as imageCLEF to create topics that are more suitable to visual retrieval groups and that correspond to desires of real users as well. Visual and textual retrieval need to be brought together with overlapping retrieval tasks to find out where each one works best and where the two can be combined for optimal results. Currently, there is no experience in this domain, hence the importance of benchmarking events such as imageCLEF but also the creation of retrieval tasks suitable for visual retrieval. This article gives a few ideas on how to make the imageCLEF task more appealing for visual retrieval groups. Hopefully, these changes will be able to attract more attention in the visual retrieval community so people start working on the same data sets and start comparing systems and techniques. To advance retrieval systems, a critical evaluation and comparison of existing systems is currently more needed than new techniques. ImageCLEF might be an important factor in advancing information retrieval and especially visual information retrieval.

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] D. A. Forsyth. Benchmarks for storage and retrieval in multimedia databases. In *Storage and Retrieval for Media Databases*, volume 4676 of *SPIE Proceedings*, pages 240–247, San Jose, California, USA, January 21–22 2002. (SPIE Photonics West Conference).

[4] T. Gevers and A. W. M. Smeulders. A comparative study of several color models for color image invariants retrieval. In *Proceedings of the First International Workshop ID-MMS'96*, pages 17–26, Amsterdam, The Netherlands, August 1996.

[5] A. Goodrum. Image information retrieval: An overview of current research. *Journal of Information Science Research*, 3(2):–, 2000.

[6] N. J. Gunther and G. Beretta. A benchmark for image retrieval using distributed systems over the internet: BIRDS–I. Technical report, HP Labs, Palo Alto, Technical Report HPL–2000–162, San Jose, 2001.

[7] D. Harman. Overview of the first Text REtrieval Conference (TREC–1). In *Proceedings of the first Text REtrieval Conference (TREC–1)*, pages 1–20, Washington DC, USA, 1992.

[8] C. Leung and H. Ip. Benchmarking for content–based visual information search. In R. Laurini, editor, *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, number 1929 in Lecture Notes in Computer Science, pages 442–456, Lyon, France, November 2000. Springer–Verlag.

[9] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel – evaluation in image retrieval. In *Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002)*, London, England, July 2002.

[10] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. In A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alcézar, J.-O. Eklundh, and Y. Aloimonos, editors, *Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000)*, pages 1043–1046, Barcelona, Spain, September 2000. IEEE.

[11] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content–based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, April 2001.

[12] A. D. Narasimhalu, M. S. Kankanhalli, and J. Wu. Benchmarking multimedia databases. *Multimedia Tools and Applications*, 4:333–356, 1997.

[13] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content–based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).

[14] S. Santini and R. Jain. Gabor space and the development of preattentive similarity. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96)*, pages 40–44, Vienna, Austria, August 1996. IEEE.

[15] J. Savoy. Report on clef–2001 experiments. In *Report on the CLEF Conference 2001 (Cross Language Evaluation Forum)*, pages 27–43, Darmstadt, Germany, 2002. Springer LNCS 2406.

[16] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content–based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 No 12:1349–1380, 2000.

[17] J. R. Smith. Image retrieval evaluation. In *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98)*, pages 112–113, Santa Barbara, CA, USA, June 21 1998.

[18] 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.