The Use of MedGIFT and EasyIR for ImageCLEF 2005

Henning Müller, Antoine Geissbühler, Johan Marty,
Christian Lovis, and Patrick Ruch

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

Abstract. This article describes the use of *medGIFT* and *easyIR* for three of four *ImageCLEF* 2005 tasks. All results rely on two systems: the GNU Image Finding Tool (*GIFT*) for visual retrieval, and *easyIR* for text. For ad–hoc retrieval, two visual runs were submitted. No textual retrieval was attempted, resulting in lower scores than those using text retrieval. For medical retrieval, visual retrieval was performed with several configurations of Gabor filters and grey level/color quantisations as well as combinations of text and visual features. Due to a lack of resources no feedback runs were created, an area where *medGIFT* performed best in 2004. For classification, a retrieval with the target image was performed and the first $N = 1; 5; 10$ results used to calculate scores for classes by simply adding up the scores for each class. No machine learning was performed, so results were surprisingly good and only topped by systems with optimised learning strategies.

1 Introduction

Image retrieval is increasingly important in information retrieval research. Compared to text retrieval little is known about how to search for images, although it has been an extremely active domain in the fields of computer vision and information retrieval [1,2,3,4]. Benchmarks such as ImageCLEF [5,6] allow to evaluate algorithms compared to other systems and deliver insights into the techniques that perform well. Thus, new developments can be directed towards these goals and techniques of well–performing systems can be adapted. More on the tasks can be found in [7].

In 2005, the ad–hoc retrieval task created topics were better adapted for visual systems using the same database as in 2004. The tasks made available contain three images. We submitted two configurations of our system to this task using visual information only.

The medical retrieval task was performed on a much larger database than in 2004 containing a total of more than 50.000 images [8]. The annotation was also more varied, ranging from a few words in a very structured form to completely unstructured paragraphs. This made it hard to preprocess any of the information. Finally, only free–text retrieval was used for our results submission including

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 724–732, 2006.

all XML tags. The tasks were much harder and mainly semantic tasks, which made the retrieval by visual means more difficult. Due to a lack of resources we could only submit partial results that did not include any relevance feedback or automatic query expansion.

The automatic annotation task was interesting and challenging at the same time [9]. We did not take into account any of the training data and simply used *GIFT* and a nearest neighbour technique to classify results. Still, the outcome is surprisingly good (6th best submission, 3rd best group) and when taking into account the learning data using an approach as described in [10], these results are expected to get better.

ImageCLEF gave us the opportunity to compare our system with other techniques which are invaluable and will provide us with directions for future research.

2 Basic Technologies Used

For our ImageCLEF participation, we aim at combining content–based retrieval of images with cross–language retrieval applied to textual annotation of the images. Based on the results from last year (2004), we used parameters that were expected to lead to good results.

2.1 Image Retrieval

The technology used for the content–based retrieval of images is mainly taken from the *Viper*[1] project of the University of Geneva. Much information about this system is available [11]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[2]. This software tool is open source and can be used by other participants of ImageCLEF. A ranked list of visually similar images for every query topic was made available for participants and will serve as a baseline to measure the quality of submissions. Demonstration versions with a web–accessible interface of *GIFT* were also made available for participants. Not everybody can be expected to install a Linux tool only for such a benchmark. The feature sets that are used by *GIFT* are:

– Local color features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode color of each region as a descriptor;
– global color features in the form of a color histogram, compared by a simple histogram intersection;
– local texture features by partitioning the image and applying Gabor filters in various scales and directions, quantised into 10 strengths;
– global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

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

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Visual features are quantised and the feature space is very similar to the distribution of words in texts, corresponding roughly 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 [12].

The medical adaptation of the *GIFT* is called *medGIFT*[3] [13]. It is also accessible as open source and adaptations concern mainly visual features and the user 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. We used in total 4, 8 and 16 grey levels and for the Gabor filters we used 4 and 8 directions. Other techniques in *medGIFT* such as a pre–treatment of images [14] were not used for this competition due to a lack of resources.

2.2 Text Search

The basic granularity of the Casimage and MIR collections is the case. A case gathers a textual report, and a set of images. For the PathoPic and PEIR databases annotation exists for every image. The queries contain one to three images and text in three languages. We used all languages as a single query and indexed all documents in one index. Case–based annotation is expanded to all images of the case after retrieval. The final unit of retrieval is the image.

Indexes. Textual experiments were conducted with the easyIR engine[4]. As a single report is able to contain written parts in several languages mixed, it would have been necessary to detect the boundaries of each language segment. Ideally, French, German 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 [15]. 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 all collections using a unique index using an English stemmer, For simplicity reasons, the XML tags were also indexed and not separately treated.

Weighting Schema. We chose a generally good weighting schema of the term frequency - inverse document frequency family. 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.

---

[3] `http://www.sim.hcuge.ch/medgift/`

[4] `http://lithwww.epfl.ch/~ruch/softs/softs.html`

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 for retrieval are still scarce in the literature [16], so many of the mechanism and fine tuning of the combinations will need more work, especially when the optimisation is based on the actual query. For the visual query we used all images that are present, including one query containing negative feedback. For the text part, the text of all three languages was used as a combined query together with the combined index that includes the documents in all languages. Results lists of the first 1000 documents were taken into account for both the visual and the textual search. Both result lists were normalised to deliver results within the range $[0; 1]$. The visual result is normalised by the result of the query itself whereas the text was normalised by the document with the highest score. This leads to visual scores that are usually slightly lower than the textual scores in high positions.

To combine the two lists, two different methods were chosen. The first one simply combines the list with different percentages for visual and textual results (textual= 50, 33, 25, 10%). In a second form of combination the list of the first 1000 visual results was taken, and then, all those that were in the first 200 textual documents were multiplied with N–times the value of the textual results.

3 The Ad Hoc Retrieval Task

For the ad–hoc retrieval task we submitted results using fairly similar techniques as those in 2004. The 2005 topics were actually more adapted to the possibilities of visual retrieval systems as more visual attributes were taken into account for the topic creation [7]. Still, textual retrieval is necessary for good results. It is not so much a problem of the queries but rather a problem of the database containing mostly grey or brown scale images of varying quality where automatic treatment such as color indexing is difficult. This should change in 2006 with a new database using mostly consumer pictures of vacation destinations.

We used *GIFT* with two configurations. First, we used the normal *GIFT* engine with 4 grey levels and the full HSV space using the Gabor filter responses in four directions and at three scales. The second configuration took into account 8 grey levels as the 2004 results for 16 grey levels were actually much worse than expected. We also raised the number of directions of the Gabor filters to 8 instead of four. The results of the basic *GIFT* system were made available to all participants and used by several. Surprisingly the results of the basic *GIFT* system remain the best in the test with a MAP of 0.0829, being at the same time the best purely automatic visual system participating. The system with eight grey levels and eight directions for the Gabor filters performed slightly worse and a MAP of 0.0819 was reached. Other visual systems performed slightly lower. The best mono–lingual text systems performed at a MAP of 0.41. Several text retrieval systems performed worse than the visual system for a variety of languages.

4 The Automatic Annotation Task

We were new to automatic annotation as almost everyone and mainly used our system for retrieval, so far. Due to a lack of resources no optimisation using the available training data was performed. Still, the tf/idf weighting is automatically weighting rare features higher which leads to a discriminative analysis.

As techniques we performed a query with each of the 1000 images to classify and took into account the first $N = 1, 5, 10$ retrieval results. For each of these results (i.e. images from the training set) the correct class was determined and this class was augmented with the similarity score of the image. The class with the highest final score became automatically the class selected for the query. For retrieval we used three different settings of the features using 4, 8, and 16 grey levels. The runs with 8 and 16 grey levels also had eight directions of the Gabor filters for indexation. Best results obtained in the competition were from the Aachen groups (best run at 12.6% error rate) that have been working on very similar data for several years, now.

The best results for our system were retrieved when using 5NN and eight grey levels (error rate 20.6%), and the next best results using 5NN and 16 grey levels (20.9). Interestingly, the worst results were obtained with 5NN and 4 grey levels (22.1). Using 10NN led to slightly worse results (21.3) and 1NN was rather in the middle (4 grey levels 21.8; 8 grey levels: 21.1; 16 grey levels 21.7).

In conclusion we can say that all results are extremely close together 20.6-22.1%, so the differences do not seem statistically significant. 5NN seems to be the best but this might also be linked to the fact that some classes have a very small population and 10NN would simply retrieve too many images of other classes to be competitive. 8 levels of grey and 8 directions of the Gabor filters seem to perform best, but the differences are still very small.

In the future, we planned to train the system with the available training data using the algorithm described in [10]. This technique is similar to the market basket analysis [17]. A proper strategy for the training needs to be developed to especially help smaller classes to be well classified. Typically, these classes cause most of the classification problems.

5 The Medical Retrieval Task

Unfortunately, our textual retrieval results contained an indexation error and the results were almost random. The only textual run that we submitted had a MAP of 0.0226. The best textual retrieval systems were at 0.2084 (IPAL/I2R). Due to a limitation of resources, we were not able to submit relevance feedback runs, where *GIFT* usually is strongest. The best feedback system was OHSU with a MAP of 0.2116 for only textual retrieval.

The best visual system is I2R with a MAP of 0.1455. Our *GIFT* retrieval system was made available to participants and was widely used. Again, the basic *GIFT* system obtained the best results among the various combinations in feature space (MAP 0.0941), with only I2R having actually better results but using manual optimisation based on the dataset. The second indexation using 8 grey levels and eight directions of the Gabor filters performs slightly worse at 0.0872.

For mixed textual/visual retrieval, the best results were obtained by IPAL/I2R with MAP 0.2821. Our best result in this category is using 10% textual part and 90% visual part and obtains 0.0981. These results should be much better when using a properly indexed text–based system. The following results were obtained for other combinations: 20% visual: 0.0934, 25%: 0.0929, 33%: 0.0834, 50%: 0.044. When using eight grey levels and 8 Gabor directions: 10% visual: 0.0891, 20%: 0.084, 33%: 0.075, 50%: 0.0407. The results could lead to the assumption that visual retrieval is better than textual retrieval in our case, but this holds only true because of our indexation error.

A second combination technique that we applied used as a basis the results from textual retrieval and then added the visual retrieval results multiplied with a factor $N = 2, 3, 4$ to the first 1000 results of textual retrieval. This strategy proved fruitful in 2004 the other way round by taking first the visual results and then augmenting only the first N=1000 results. The results for the main *GIFT* system were: 3 times visual: 0.0471, 4 times visual 0.0458, 2 times visual 0.0358. For the system with 8 grey levels, the respective results are: 3 times visual 0.0436, 4 times visual 0.0431, 2 times visual 0.0237. A reverse order of taking the visual results first and then augment the textually similar would have led to better results in this case but when having correct results for text as well as for visual retrieval, this needs to be proven.

The MAP scores per topic shown in Figure 1 show that the textual retrieval is extremely low for all but very few queries (GE_M_TXT.txt) compare with the visual results (GE_M_88.TXT and GE_M_4g.txt). For the queries with good text results the mixed retrieval (GE_M_10.txt) is actually much better than

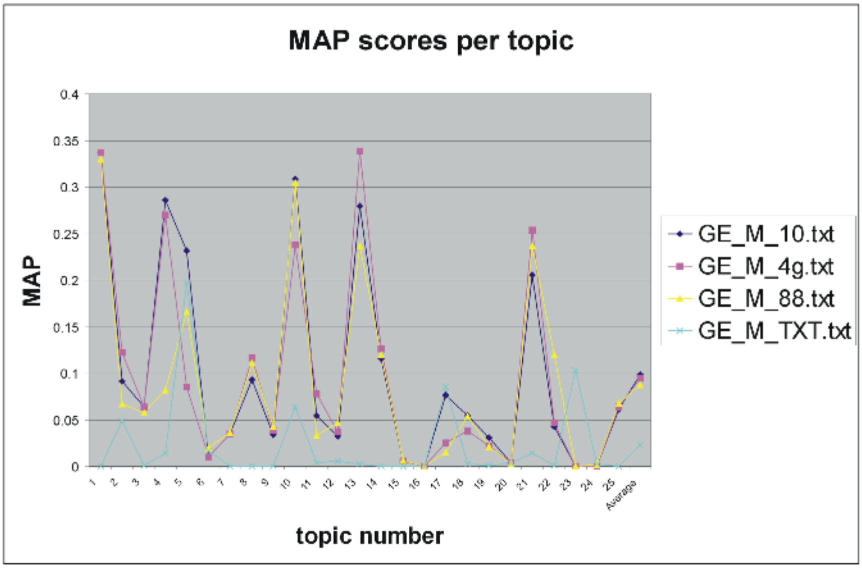

Fig. 1. MAP per topic for four different system configurations

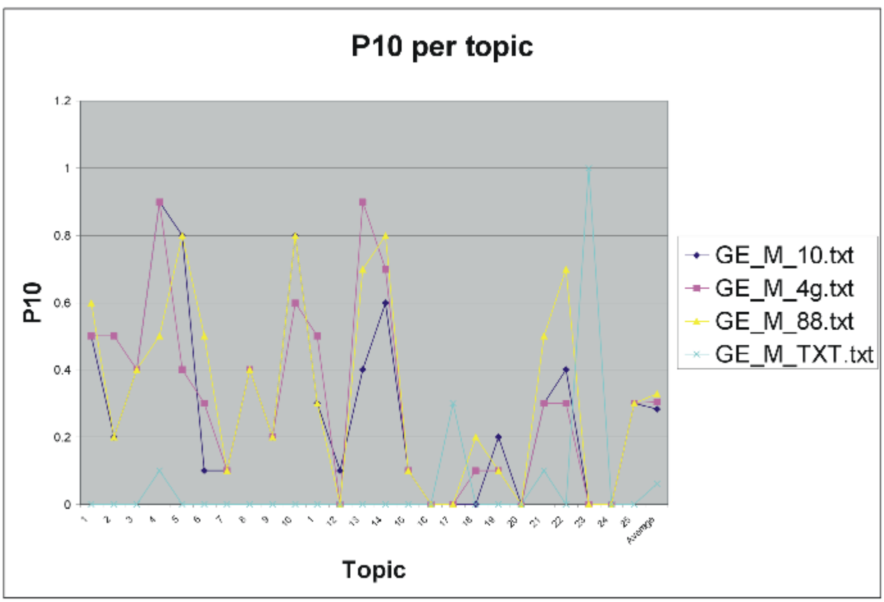

Fig. 2. Precision after ten images retrieved per topic for four different configurations

only textual retrieval. This shows the potential of the mixed runs with correct text retrieval results. Best test retrieval results were 10 times better than ours. The precision after ten retrieval image per topic can be seen in Figure 2. This underlines our previous assumption as most results for the text retrieval receive no relevant images early on, whereas visual retrieval does have very good results. Much more can not be concluded from our submission as several errors prevented better results.

6 Conclusions

Although we did not have any resources for an optimised submission we still learned from the 2005 tasks that the *GIFT* system delivers a good baseline for visual image retrieval and that it is widely usable for a large number of tasks and different images. More detailed results show that the ad–hoc task is hard for visual retrieval even with a more visually–friendly set of queries as the image set does not contain enough color information or clear objects, which is crucial for fully visual information retrieval.

The automatic annotation or classification task proved that our system delivers good results even without learning and shows that information retrieval can also be used well for document classification. When taking into account the available training data these results will surely improve significantly.

From the medical retrieval task not much can be deduced for now as we need to work on our textual indexation and retrieval to find the error responsible for the mediocre results. Still, we can say that *GIFT* is well suited and among the best systems for general visual retrieval. It will need to be analysed which features were used by other systems, especially runs performing better.

For next year we will definitely have to take into account the available training data and we hope as well to use more complex algorithms for example to extract objects form the medical images and limit retrieval to theses objects. Another strong point of *GIFT* is the good relevance feedback and this can surely improve results significantly as well. Already the fact to have a similar databases for two years in a row would help as such large databases need a large time to be indexed and require human resources for optimisation as well.

Acknowledgements

Part of this research was supported by the Swiss National Science Foundation with grants 632-066041 and 205321-109304/1. We also acknowledge the support of the EU FP6 project SemanticMining (IST NoE 507505).

References

1. Smeulders, A.W.M., 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 No 12** (2000) 1349–1380

2. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics **73** (2004) 1–23
3. Tagare, H.D., Jaffe, C., Duncan, J.: Medical image databases: A content–based retrieval approach. Journal of the American Medical Informatics Association **4** (1997) 184–198
4. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content–based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology **8** (1998) 644–655 (Special Issue on Segmentation, Description, and Retrieval of Video Content).
5. 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 3115 (2004)
6. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Lecture Notes in Computer Science, Bath, England, Springer–Verlag (2005)
7. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Springer Lecture Notes in Computer Science (LNCS), Vienna, Austria (2006 – to appear)
8. Hersh, W., Müller, H., Gorman, P., Jensen, J.: Task analysis for evaluating image retrieval systems in the ImageCLEF biomedical image retrieval task. In: Slice of Life conference on Multimedia in Medical Education (SOL 2005), Portland, OR, USA (2005)
9. Lehmann, T.M., Güld, M.O., Deselaers, T., Schubert, H., Spitzer, K., Ney, H., Wein, B.B.: Automatic categorization of medical images for content–based retrieval and data mining. Computerized Medical Imaging and Graphics **29** (2005) 143–155
10. Müller, H., Squire, D.M., Pun, T.: Learning from user behavior in image retrieval: Application of the market basket analysis. International Journal of Computer Vision **56(1–2)** (2004) 65–77 (Special Issue on Content–Based Image Retrieval).
11. 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.
12. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7** (1991) 11–32
13. 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)
14. Müller, H., Heuberger, J., Geissbuhler, A.: Logo and text removal for medical image retrieval. In: Springer Informatik aktuell: Proceedings of the Workshop Bildverarbeitung für die Medizin, Heidelberg, Germany (2005)
15. Ruch, P.: Query translation by text categorization. In: Proceedings of the conference on Computational Linguistics (COLING 2004), Geneva, Switzerland (2004)
16. 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)
17. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference, Santiago, Chile (1994) 487–499