University and Hospitals of Geneva at ImageCLEF 2007

Xin Zhou, Julien Gobeill, Patrick Ruch, Henning Müller
University and Hospitals of Geneva, Switzerland
xin.zhou@sim.hcuge.ch

Abstract

This article describes the participation of the University and Hospitals of Geneva at three tasks of the 2007 ImageCLEF image retrieval benchmark. Two of these tasks were medical tasks and one was a photographic retrieval task. The visual retrieval techniques relied mainly on the GNU Image Finding Tool (GIFT) whereas multilingual text retrieval was performed by mapping the full text documents and the queries in a variety of languages onto MeSH (Medical Subject Headings) terms, using the EasyIR text retrieval engine for retrieval.

For the visual tasks it becomes clear that the baseline GIFT runs do not have the same performance as more sophisticated modern techniques do. GIFT can be seen as a baseline for the visual retrieval as it has been used for the past four years in ImageCLEF. Whereas in 2004 the performance of GIFT was among the best systems it now is towards the end of the spectrum, showing the clear improvement in retrieval quality of participants over the years. Due to time constraints no optimisations could be performed and no relevance feedback was used, usually one of the strong points of GIFT. The text retrieval runs have a fairly good performance showing the effectiveness of the approach to map terms onto an ontology. Mixed runs are in performance slightly lower than the best text results alone, meaning that more care needs to be taken in combining runs other than a simple linear combination. English is by far the language with the best results; even a mixed run of the three languages was lower in performance. This can partly be explained with the judges as they are all native English speakers. Thus, a bias towards relevance for English documents is unfortunately possible.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, text categorization, multimodal retrieval, automatic annotation

1   Introduction

ImageCLEF[1] [1, 2] has started within CLEF[2] (Cross Language Evaluation Forum [12]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. A medical image retrieval task[3] was added in 2004 to explore domain–specific multilingual information retrieval and also multi modal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task are both presented as part of ImageCLEF [8].

More about the ImageCLEF tasks, topics, and results in 2007 can also be read in [3, 6, 7].

2   Retrieval Strategies

This section describes the basic technologies that are used for the data retrieval. More details on optimisations per tasks are given in the results section.

2.1   Text retrieval approach

The text retrieval approach used in 2007 is similar to the techniques already applied in 2006 [5]. The full text of the documents in the collection and of the queries were mapped to a fix number of MeSH terms, and retrieval was then performed in the MeSH–term space. Based on the results of 2006, when 3, 5, and 8 terms were extracted we increased the number of terms further. It was shown in 2006 that a larger number of terms lead to better results, although several of the terms might be incorrect, these incorrect terms create less damage than the few additionally correct terms add in quality. Thus 15 terms were generated for each document in 2007 and 3 terms from every query, separated by language. Term generation is based on the MeSH categorizer [9, 10] developed in Geneva. As MeSH exists in English, German, and French, multilingual treatment of the entire collection is thus possible. For ease of computation an English stemmer was used on the collection and all XML tags in the documents were removed, basically removing all structure of the documents. The entire text collection was indexed with the easyIR toolkit [11] using a pivoted–normalization weighting schema. Schema tuning was discarded due to the lack of time.

Queries were executed in each of the three languages separately and one run combined the results of the three languages.

2.2   Visual retrieval techniques

The technology used for the visual retrieval of images is mainly taken from the *Viper*[4] project [13]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[5]. This 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 serves as a baseline to measure the quality of submissions. Feature sets 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;

---

[1] http://ir.shef.ac.uk/imageclef/
[2] http://www.clef-campaign.org/
[3] http://ir.ohsu.edu/image/
[4] http://viper.unige.ch/
[5] http://www.gnu.org/software/gift/

Table 1: Our two runs for the photographic retrieval task.

run ID	MAP	P10	P30	Relevant retrieved
best visual run	0.1890	0.4700	0.2922	1708
GE_GIFT18_3	0.0222	0.0983	0.0622	719
GE_GIFT9_2	0.0212	0.0800	0.0594	785

- 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 well–known from text retrieval. Visual features are quantised and the feature space is similar to the distribution of words in texts. 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 [14].

3 Results

This section details the results obtained for the various tasks. It always compares our results to the best results in the competition to underline the fact that our results are a baseline for comparison of techniques.

3.1 Photographic Image Retrieval

The two runs submitted for the photographic retrieval task do not contain any optimisations and are a simple baseline using the GIFT system to compare performance of other techniques and their improvement over the years. Only visual retrieval was attempted and no text was used. The two runs are fully automatic.

Table 1 shows the results of the two submitted runs with gift compared to best overall visual run submitted. MAP is much lower than the best run, almost by a factor of ten, whereas early precision is about a factor of five lower. The best run uses the standard GIFT system whereas the second run uses a smaller number of colors (9 hues instead of 18) and a smaller number of saturations as well. The results with these changes are slightly lower but the number of relevant images found is significantly higher, meaning that more fuzziness in the feature space is better for finding relevant images but less good concerning early precision.

3.2 Medical Image Retrieval

This section describes the three categories of runs that were submitted for the medical retrieval task. All runs were automatic and so the results are classified by media.

3.2.1 Visual Retrieval

The purely visual retrieval was performed with the standard GIFT system using 4 grey levels and with a modified gift using 8 grey levels. A third run was created by a linear combination the two previous runs.

Figure 2 shows the results of the best overall visual run and all of our runs. It is actually interesting to see that all but three visual runs have very low performance in 2007. These three runs used training data on almost the same collection of the years 2005 and 2006 to select and weight features, leading to an extreme increase in retrieval performance. Our runs are on the lower end of the spectrum concerning MAP. Early precision becomes much better in the combination runs using a combination of two grey level quantisations.

Table 2: Results for purely visual retrieval at the medical retrieval task.

run ID	MAP	P10	P30
best visual run	0.2328	0.4867	0.4333
GE_4_8	0.0041	0.0400	0.0322
GE_GIFT8	0.0041	0.0194	0.0322
GE_GIFT4	0.0040	0.0192	0.0322

Table 3: Results for purely textual retrieval.

run ID	MAP	P10	P30
best textual run	0.3962	0.5067	0.4600
GE_EN	0.2714	0.3900	0.3356
GE_MIX	0.2416	0.3500	0.3133
GE_DE	0.1631	0.2200	0.1789
GE_FR	0.1557	0.1933	0.2067

3.2.2 Textual retrieval

Textual retrieval was performed using each of the query languages separately and in one combined run.

Results can be seen in Table 3. The results show clearly that English obtains the best performance among the three languages. This can be explained as the majority of the documents are in English and the majority of relevance judges are also native English speakers For most of the best performing runs it is not clear whether they use a single language or a mix of languages, which is not really a realistic scenario for multilingual retrieval. Both, German and French retrieval have a lower performance than English and the run linearly combining the three languages is also lower in performance than English alone.

3.2.3 Mixed–media retrieval

There were two different sorts of mixed media runs in 2007 from the University and Hospitals of Geneva. One was a combination of our own visual and textual runs and the other was a combination of the GIFT results with results from the FIRE (Flexible Image Retrieval Engine) system and a system from OHSU (Oregon Health and Science University).

The combinations of our visual with our own English retrieval run are all better in quality than the combinations with the FIRE and OHSU runs. Combinations are all simple, linear com-

Table 4: Combined media runs.

run ID	MAP	P10	P30
best mixed run	0.3719	0.5667	0.5122
GE_VT1_4	0.2425	0.3533	0.3133
GE_VT1_8	0.2425	0.3533	0.3133
GE_VT5_4	0.2281	0.3500	0.3122
GE_VT5_8	0.2281	0.3500	0.3122
GE_VT10_4	0.1938	0.3600	0.3133
GE_VT10_8	0.1937	0.3600	0.3100
3gift-3fire-4ohsu	0.0334	0.0067	0.0111
5gift-5ohsu	0.0188	0.0033	0.0044
7gift-3ohsu	0.0181	0.0033	0.0044

Table 5: Results of the runs submitted to the medical image annotation task.

run ID	score
best system	26.847
GE_GIFT10_0.5ve	375.720
GE_GIFT10_0.15vs	390.291
GE_GIFT10_0.66vd	391.024

binations with a percentage of 10%, 50% and 90% of the visual runs. It shows that the smallest proportion of visual influence delivers the best results, although not as high as the purely textual run alone. Differences between the two grey level quantisations (8 and 4) are extremely small. All combinations runs between systems at OHSU and the FIRE system did not work very well, having a very low performance.

3.3 Medical Image Classification

For medical image classification the basic GIFT system was used as a baseline for classification. It shows as already in [4] that the features are not too well suited for image classification as they do not include any invariance and are on a very low level. Performance as shown in Table 5 is low compared to the best systems.

The strategy was to perform the classification in an image retrieval way. No training phase was carried out. Visually similar images with known classes are used to classify images from the test set. In practice, the first 10 retrieved images of every image of th test set were taken into account, and the scores of these images were used to choose the IRMA code on all hierarchy levels. When the sum of the scores for a certain code reaches a fixed threshold, an agreement can be assumed for this level. This allows the classification to be performed up to this level. Otherwise, this level and all further levels were not classified and left empty.

Thresholds and score distribution strategies varied slightly. Three score distribution strategies were used:

- Every retrieved image votes equally. A code at a certain level will be chosen only if more than half of the results are in agreement.

- Retrieved images vote with decreasing importance values (from 10 to 1) according to the rank. A code at a certain level will be chosen if more than 66% of hte maximum were reached for one code.

- The retrieved images vote with their absolute similarity value. A code at a certain level will be chosen if the average of the similarity score for this code is higher than 0.15.

The performance varies slightly depending on the chosen strategies. Results in Table 5 show that the easiest method gives the best result. It can be concluded that a high similarity score is not a significant parameter to classify images.

4 Discussion

The results show clearly that visual retrieval with the GIFT is not state of the art anymore and that more specific techniques can receive much better retrieval results. Still, the GIFT runs serve as a baseline as they can be reproduced easily as the software is open source and they have been used in ImageCLEF since 2004, which clearly shows the improvement of techniques participating in ImageCLEF since this time.

The text retrieval approach shows that the extraction of MeSH terms from documents and queries and then performing retrieval based on these terms is working well. Bias is towards the

English terms with a majority of documents being in English and also the relevance judges being all native speakers.

Combining visual and textual retrieval remains difficult and in our case no result is as good as the English text results alone. Much potential still seems to be in this combination of media.

For the classification of images our extremely easy was mainly hindered by the simple base features that were used.

Acknowledgements

References

[1] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF 2004 cross language image retrieval track. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, pages 597–613. Lecture Notes in Computer Science (LNCS), Springer, Volume 3491, 2005.

[2] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Michael Jones, Gareth J. F.and Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, 2005. Springer.

[3] Thomas Deselaers, Allan Hanbury, and et al. Overview of the ImageCLEF 2007 object retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[4] Tobias Gass, Antoine Geissbuhler, and Henning Müller. Learning a frequency–based weighting for medical image classification. In *Medical Imaging and Medical Informatics (MIMI) 2007*, Beijing, China, 2007.

[5] Julien Gobeill, Henning Müller, and Patrick Ruch. Translation by text categorization: Medical image retrieval in ImageCLEFmed 2006. In *CLEF 2006 Proceedings*, volume 4730 of *Springer Lecture Notes in Computer Science*, Alicante, Spain, 2007.

[6] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[7] Henning Müller, Thomas Deselaers, Eugene Kim, Jayashree Kalpathy-Cramer, Thomas M. Deserno, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[8] Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[9] Patrick Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664, 2006.

[10] Patrick Ruch, Robert H. Baud, and Antoine Geissbühler. Learning–free text categorization. In Michel Dojat, Elpida T. Keravnou, and Pedro Barahona, editors, *AIME*, volume 2780 of *Lecture Notes in Computer Science*, pages 199–208. Springer, 2003.

[11] Patrick Ruch, A. Jimeno Yepes, Frdric Ehrler, Julien Gobeill, and Imad Tbahriti. Report on the trec 2006 experiment: Genomics track. In *TREC*, 2006.

[12] Jacques 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.

[13] David McG. Squire, Wolfgang Müller, Henning Müller, and Thierry 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.

[14] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.