# University and Hospitals of Geneva Participating at ImageCLEF 2007

Xin Zhou[1], Julien Gobeill[1], Patrick Ruch[1], and Henning Müller[1,2]

[1] Medical Informatics Service, University and Hospitals of Geneva, Switzerland
[2] Business Information Systems, University of Applied Sciences, Sierre, 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. 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 techniques such as visual patch histograms do have. GIFT can be seen as a baseline for the visual retrieval as it has been used for the past four years in ImageCLEF in the same configuration. Whereas in 2004 the performance of GIFT was among the best systems it now is towards the lower end, showing the clear improvement in retrieval quality. Due to time constraints no further optimizations could be performed and no relevance feedback was used, one of the strong points of GIFT. The text retrieval runs have a 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. English is by far the language with the best results; even a mixed run of the three languages was lower in performance.

## 1 Introduction

ImageCLEF[1] [1] has started within CLEF[2] (Cross Language Evaluation Forum, [2]) 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 [3].

More about the ImageCLEF tasks, topics, and results in 2007 can also be read in [4,5,6].

[1] http://www.imageclef.org/
[2] http://www.clef-campaign.org/
[3] http://ir.ohsu.edu/image/

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 649–656, 2008.

## 2 Retrieval Strategies

This section describes the basic technologies that are used for the retrieval. More details on small optimizations per task 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 [7]. The full text of the documents in the collection and of the queries were mapped to a fixed 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 a MeSH categorizer [8,9] 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 [10] 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 an additional 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 [11]. An outcome of the *Viper* project is the *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, quantized into 10 strengths;
- global texture features represented as a simple histogram of responses of the local Gabor filters at the smallest size in various directions and scales.

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

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Visual features are quantized 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 normalized by the results of the query itself. The histogram features are compared based on a histogram intersection [12].

## 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 optimizations and are a simple baseline using the GIFT system to compare the improvement of participants over the years. Only visual retrieval was attempted and no text was used. The two runs are fully automatic.

**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 |

Table 1 shows the results of the two submitted runs with GIFT compared to the 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 (visual, textual, mixed). All runs were automatic and so the results are classified by the media used.

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

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

| Run | num_ret | num_rel_ret | MAP | R–prec | bpref | P10 | P30 |
|---|---|---|---|---|---|---|---|
| best visual run | 30000 | 1376 | 0.2427 | 0.264 | 0.283 | 0.48 | 0.3756 |
| GE_4_8 | 30000 | 245 | 0.0035 | 0.0144 | 0.0241 | 0.04 | 0.0233 |
| GE_GIFT8 | 30000 | 245 | 0.0035 | 0.0143 | 0.024 | 0.04 | 0.0233 |
| GE_GIFT4 | 30000 | 244 | 0.0035 | 0.0144 | 0.024 | 0.04 | 0.0233 |

Table 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 but very close to other visual runs. Early precision becomes slightly better in the combination runs using a combination of two gray level quantizations.

**Textual Retrieval.** Textual retrieval was performed using each of the query languages separately and in addition in a combined run.

**Table 3.** Results for purely textual retrieval

| Run | num_ret | num_rel_ret | MAP | R–prec | bpref | P10 | P30 |
|---|---|---|---|---|---|---|---|
| best textual run | 28537 | 1904 | 0.3538 | 0.3643 | 0.3954 | 0.43 | 0.3844 |
| GE_EN | 27765 | 1839 | 0.2369 | 0.2537 | 0.2867 | 0.3333 | 0.2678 |
| GE_MIX | 30000 | 1806 | 0.2186 | 0.2296 | 0.2566 | 0.2967 | 0.2622 |
| GE_DE | 26200 | 1166 | 0.1433 | 0.1579 | 0.209 | 0.2 | 0.15 |
| GE_FR | 29965 | 1139 | 0.115 | 0.1276 | 0.1503 | 0.1267 | 0.1289 |

Results of our four runs 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 creating both a potential bias towards relevant documents in English. 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. In comparison to the best overall runs our system is close in number of relevant items found and still among the better systems in all other categories.

**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). In these runs we discovered a problem we had with the evaluation of the treceval package that does not take into account the order of the items in the submitted runs. Some runs assumed the order to be the main criterion and had same weightings for many items. This can result in very different scores and for this reason we add in this table a recalculated map where the score is simply set to 1/rank.

**Table 4.** Results for the combined media runs

| Run | num_ret | num_rel_ret | MAP | new MAP | R–prec | bpref | P10 | P30 |
|---|---|---|---|---|---|---|---|---|
| best mixed run | 21868 | 1778 | 0.3415 | 0.4084 | 0.3808 | 0.4099 | 0.4333 | 0.37 |
| GE_VT1_4 | 30000 | 1806 | 0.2195 | 0.2199 | 0.2307 | 0.2567 | 0.3033 | 0.2622 |
| GE_VT1_8 | 30000 | 1806 | 0.2195 | 0.2204 | 0.2307 | 0.2566 | 0.3033 | 0.2622 |
| GE_VT5_4 | 30000 | 1562 | 0.2082 | 0.2090 | 0.2328 | 0.2423 | 0.2967 | 0.2611 |
| GE_VT5_8 | 30000 | 1565 | 0.2082 | 0.2082 | 0.2327 | 0.2424 | 0.2967 | 0.2611 |
| GE_VT10_4 | 30000 | 1192 | 0.1828 | 0.1829 | 0.2125 | 0.2141 | 0.31 | 0.2633 |
| GE_VT10_8 | 30000 | 1196 | 0.1828 | 0.1839 | 0.2122 | 0.214 | 0.31 | 0.2633 |
| 3gift-3fire-4ohsu | 29651 | 1748 | 0.0288 | 0.1564 | 0.0185 | 0.1247 | 0.0067 | 0.0111 |
| 4gift-4fire-2ohsu | 29651 | 1766 | 0.0284 | 0.2194 | 0.0135 | 0.1176 | 0.0233 | 0.0156 |
| 1gift-1fire-8ohsu | 29709 | 1317 | 0.0197 | 0.0698 | 0.0184 | 0.1111 | 0.0067 | 0.0133 |
| 3gift-7ohsu | 29945 | 1311 | 0.0169 | 0.1081 | 0.0108 | 0.1309 | 0.0033 | 0.0044 |
| 5gift-5ohsu | 29945 | 1317 | 0.0153 | 0.1867 | 0.0057 | 0.1151 | 0.0033 | 0.0022 |
| 7gift-3ohsu | 29945 | 1319 | 0.0148 | 0.2652 | 0.0042 | 0.1033 | 0.0033 | 0.0022 |

The combinations of our visual with our own English retrieval run were all better in quality than the combinations with the FIRE and OHSU runs in the initial results but when re–scoring the images taking into account the rank information this changes completely! Combinations are all simple, linear combinations 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 concerning MAP, although not as high as the purely textual run alone. Concerning early precision the runs with a higher visual proportion are on the other hand better than with a lower percentage. Differences between the two gray level quantizations (8 and 4) are extremely small.

### 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 [13] that the features are not too well suited for image classification as they do not include any invariance and are on a very low semantic level. Performance as shown in Table 5 is low compared to the best systems for our runs submitted for the competition.

The strategy was to perform the classification in an image retrieval way. No training phase was carried out. Visually similar images with known classes were used to classify images from the test set. In practice, the first 10 retrieved images

**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 |

of every image of the 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. This is similar to a classical kNN (k Nearest Neighbors) approach.

Thresholds and voting strategies varied slightly. Three voting 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 their rank. A code at a certain level will be chosen if more than 66% of the maximum was reached for a 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 a fixed value.

Results in Table 5 show that the easiest method gives the best results. It can be concluded that a high similarity score is not a significant parameter to classify images.

*New Runs.* Based on our first experiences with the classification, several parameters were tried out to optimize performance without learning for the existing system. One clear idea was that taking only the first ten images was not enough, so up to the first 100 images were taken into account. The threshold was also regarded as too high favoring non–classification over taking chances. Another approach was to classify images not only on the entire hierarchy but also fixed on a full axis level or fixed for the entire code. In the competition the best systems did not take into account the hierarchy at all. Adding a simple aspect ration as feature further improved our results significantly (reduction in error score of around 100). All this brought down the overall classification to 234 instead of an initial 391, which is an enormous gain. Table 6 details the best results obtained with these changes. The best run actually performs classification on an axis–bases, thus takes into account part of the hierarchy.

Despite the enormous improvements in the error score it can clearly be seen that new feature sets and a learning strategy still have an strong potential for our approach.

**Table 6.** Results of some new runs to search for the optimization

| run ID | score |
|---|---|
| GE_GIFT13_0.4vad_withAR | 234.469972 |
| GE_GIFT11_0.4vae_withAR | 238.0446107 |
| GE_GIFT100_vakNN_withAR | 262.249183 |

## 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 than a very simple and general retrieval system that did perform well in benchmarks three years ago. 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 English speakers. In a truly multilingual setting with unbiased relevance judges, such an approach to map terms onto an ontology should even perform much better than the other approaches mixing languages.

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

For the classification of images our extremely easy approach was mainly hindered by the simple base features that were used and the absence of using the training data for optimization. Simple improvements such as the use of the aspect ration and slightly modified voting schemes improved the results already enormously.

## Acknowledgements

This study was partially supported by the Swiss National Science Foundation (Grants 3200–065228 and 205321–109304/1) and the European Union (SemanticMining Network of Excellence, INFS–CT–2004–507505) via OFES Grant (No 03.0399).

## References

1. Clough, P., Müller, H., Sanderson, M.: The CLEF cross–language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)

2. Savoy, J.: Report on CLEF–2001 experiments. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 27–43. Springer, Heidelberg (2002)
3. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: CLEF working notes, Alicante, Spain (September 2006)
4. Deselaers, T., Hanbury, A., et al.: Overview of the ImageCLEF 2007 object retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
5. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
6. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
7. Gobeill, J., Müller, H., Ruch, P.: Translation by text categorization: Medical image retrieval in ImageCLEFmed 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
8. Ruch, P.: Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22(6), 658–664 (2006)
9. Ruch, P., Baud, R.H., Geissbühler, A.: Learning-free text categorization. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS (LNAI), vol. 2780, pp. 199–208. Springer, Heidelberg (2003)
10. Ruch, P., Jimeno Yepes, A., Ehrler, F., Gobeill, J., Tbahriti, I.: Report on the trec 2006 experiment: Genomics track. In: TREC (2006)
11. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content–based query of image databases: inspirations from text retrieval. In: Ersboll, B.K., Johansen, P. (ed.) Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA 1999), B.K, vol. 21(13-14), pp. 1193–1198 (2000)
12. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
13. Gass, T., Geissbuhler, A., Müller, H.: Learning a frequency–based weighting for medical image classification. In: Medical Imaging and Medical Informatics (MIMI) 2007, Beijing, China (2007)