The MedGIFT group at ImageCLEF 2008

Xin Zhou[1], Julien Gobeill[1], Henning Müller[1,2]

[1]Geneva University Hospitals and University of Geneva, Switzerland
[2]University of Applied Sciences Western Switzerland, Sierre, Switzerland
henning.mueller@sim.hcuge.ch

Abstract. This article describes the participation of the MedGIFT research group at the 2008 ImageCLEFmed image retrieval benchmark. We concentrated on the two tasks concerning medical imaging. The visual information analysis is mainly based on the GNU Image Finding Tool (GIFT). Other information such as textual information and aspect ratio were integrated to improve our results. The main techniques are similar to past years, with tuning a few parameters to improve results. For the visual tasks it becomes clear that the baseline GIFT runs do not have the same performance as some more sophisticated and more modern techniques. GIFT can be seen as a baseline for the visual retrieval as it has been used for the past five years in ImageCLEF. Due to time constraints not all optimizations could be performed and no relevance feedback was used, one of the strong points of GIFT. Still, a clear difference in performance can be observed depending on the various optimizations applied, and the difference with the best groups is smaller than in past years.

1 Introduction

The MedGIFT group of the Geneva University Hospitals and the University of Geneva contribute regularly to ImageCLEF[1]. The principle domains of interest are medical retrieval and medical image annotation [1]. More details on the ImageCLEF databases, topics, and a comparison of all medical retrieval results can be found in [2]. In [10] the medical classification is detailed.

2 Basic Retrieval Strategies

This section describes the basic technologies that were used for the retrieval by the mdGIFT group. More details on optimizations per task are given in the results section.

2.1 Text Retrieval Approach

The text retrieval approach used in 2008 is detailed in a paper of the text retrieval group of the Geneva University Hospitals [3]. It is similar to approaches in past years, where queries and documents were translated into MeSH (Medical Subject Heading) terms.

[1] http://www.imageclef.org/

2.2 Visual Retrieval Techniques

The technology used for the visual retrieval is mainly taken from the *Viper*[2] (Visual Information Processing for Enhanced Retrieval) project [4]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[3]. This tool is open source and can be used by other participants of ImageCLEF as well. A ranked list of visually similar images for all query topics was made available for participants and serves as 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 in various directions and scales.

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Most visual features are quantized and the feature space is similar to the distribution of words in texts. A standard *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 [5].

3 Results

In this section, the results and technical details for the two medical tasks of ImageCLEF 2008 are detailed.

3.1 Medical Image Retrieval

Results of our runs for the medical retrieval task are shown in Table 1 highlighting the most important performance measures such as MAP (Mean Average Precision), Bpref, and early precision. 3 purely visual retrieval runs using GIFT with 4 gray levels (*GIFT4*), 8 gray levels (*GIFT8*), and 16 gray levels (*GIFT16*) were submitted for evaluation. Using GIFT with 8 gray levels gives the best result for purely visual retrieval. Increasing the number of gray levels further decreases basically all performance measures.

Purely visual retrieval results in past years were often not robust [6]. Thus, more effort was invested into mixing visual retrieval and textual retrieval. The textual retrieval run (*HUG–BL–EN*) was provided by the text retrieval group of

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

Table 1. Results of the runs submitted to the medical retrieval task.

Run	run_type	MAP	bpref	P10	P30	num_ret
best system	Mixed	0.2908	0.327	0.4267	0.3956	30000
HUG–BL–EN	Textual	0.1365	0.2053	0.26	0.24	28095
GE–GE_GIFT8_EN0.5	Mixed	0.0848	0.1927	0.2433	0.2378	29999
GE–GE_EN_reGIFT8	Mixed	0.0815	0.1896	0.2267	0.2267	29452
GE–GE_EN_GIFT8_mix	Mixed	0.0812	0.1867	0.24	0.2467	29999
GE–GE_GIFT8_EN0.9	Mixed	0.0731	0.1248	0.2733	0.25	30000
GE–GE_GIFT8_reEN	Mixed	0.0724	0.1244	0.2433	0.2544	30000
GE–GE_GIFT4	Visual	0.0315	0.0901	0.1433	0.12	30000
GE–GE_GIFT8	Visual	0.0349	0.0898	0.17	0.1511	30000
GE–GE_GIFT16	Visual	0.0255	0.0715	0.1333	0.1111	30000

the Geneva University Hospitals [3]. This text retrieval run was used for several combinations with our best–performing visual run (*GIFT8*). In total, 5 mixed–media automatic runs were generated based on these runs with the following combination strategies:

- combination of textual and visual runs with equal weight (*GIFT8_EN0.5*);
- reordering of the ranked lists of the textual run based on the visual run (*EN_reGIFT8*);
- mixing visual and textual retrieval by giving varying weights based on the kind of topic: for visual topics the visual run is at 90%, for textual topics the visual run is at 10%, for mixed topics the visual run is at 50% (*EN_GIFT8_mix*);
- combining textual and visual runs but favoring the text (90%) over the visual information (10%) (*GIFT8_EN0.9*);
- reordering the visual run based on the textual run (*GIFT8_reEN*).

Mixing two runs with varying weights based on the topic type (*EN_GIFT8_mix*) gives second best early precision (P30), and third best MAP among the 5 runs. The best MAP is reached by simply combining textual and visual runs with equal weight (*GIFT8_EN0.5*). Favoring the textual run (*GIFT8_EN0.9*) gives best early precision, but surprisingly poor MAP. Compared to the original text runs, the combination with our visual run improves early precision slightly, but reduces MAP significantly.

3.2 Medical Image Annotation

For the medical image annotation task, the basic GIFT system was used for the feature extraction as in previous years but with significant changes [7]. Aspect ratio as feature and annotation by axis were again used for our participation in 2008. Main new approaches for 2008 were a modified classification strategy and changed parameter settings.

The annotation is based on the known labels of similar images of the training set retrieved by GIFT. In [7], the classification strategies were regrouped around

a kNN (k Nearest Neighbor) approach and a voting–based approach. The voting–based approach takes into account the n most similar images. In 2008, we took into account two other factors: the frequency of images of each class in the training data and the hierarchy information inside each axis of the IRMA (Image Retrieval in Medical Applications) code.

One problem of classifying images with training data is that the classification strategy most often favors large classes in the training data and punishes small ones, as images of large classes have a higher chance to be selected. The frequency of each class in the training data is analyzed to avoid this bias. Such a dynamic kNN approach is then used instead of a standard kNN approach to give a different k value for each class. The disadvantages for the smaller classes are thus reduced. In previous years, the distribution of classes in the test data was the same as in the training data, which is not the case in 2008. Thus, using a dynamic kNN approach to avoid the bias is even more necessary.

Another useful information is the hierarchy information inside each code axis (the IRMA code in total contains four). The output of the classification per axis is usually an entire axis or a wild card for the entire axis. Another possibility is to chop only the lowest level (the last letter) of each axis. The remainder can then be used for a second round of classification. This additional step allows to use less wild cards in the classification process and thus can potentially improve the score.

Table 2. Results of the main runs submitted by MedGIFT to the medical image annotation task.

run ID	score
best system	74.92
GE–GIFT0.9_0.5_vad_5.run	209.70
GE–GIFT0.9_0.5_vcad_5.run	210.93
GE–GIFT0.9_0.5_vca_5.run	217.34
GE–GIFT0.9_adkNN_2.run	233.02
GE–GIFT0.9_akNN_2.run	241.11
GE–GIFT0.9_kNN_2.run	251.97

The results of our basic runs and the best overall system are presented in Table 2. Three submitted runs use the kNN approach with classification for the entire code (*kNN*), classification per axis (*akNN*), and dynamic kNN classification per axis (*adkNN*). Dynamic kNN obtains the best result of these three approaches. Three other runs use a voting–based approach described in [7]: per axis with descending vote (*vad*), per axis with chopping letter by letter with a descending vote (*vcad*), and per axis with chopping letter by letter using equal weights (*vca*). The confidence thresholds were all set to 0.5 (as this obtained good results in past years) and we submitted the runs that take into account the first 5 similar images, only. In tests this lead to good results and no optimization

for this parameter was tried. The best results among these runs is obtained using the voting strategy per axis with descending vote(vad). Surprisingly, chopping the lowest level and redoing the classification for the rest gives slightly worse results. To detail the two best–performing techniques and optimize results a further comparison is performed with varying parameters and presented in Table 3. Chopping at the lowest level and re–classification performs better but only when

Table 3. Classification per axis with and without a chopping strategy.

run ID	score
GE–GIFT0.9_0.5_vad_5.run	209.70
GE–GIFT0.9_0.6_vad_5.run	198.79
GE–GIFT0.9_0.7_vad_5.run	198.79
GE–GIFT0.9_0.8_vad_5.run	198.79
GE–GIFT0.9_0.9_vad_5.run	208.23
GE–GIFT0.9_0.5_vcad_5.run	210.93
GE–GIFT0.9_0.6_vcad_5.run	191.53
GE–GIFT0.9_0.7_vcad_5.run	191.53
GE–GIFT0.9_0.8_vcad_5.run	191.53
GE–GIFT0.9_0.9_vcad_5.run	181.17

using a high threshold.

The two best groups (IDIAP and MIPLAB) in the classification competition in 2008 both use a similar approach for their visual characteristics. This *bag of features* approach is based on neighborhoods of interest points randomly selected from the image, followed by a Support Vector Machine (SVM)–based classification approach [8, 9]. Both use a large number of features (1'000–5'000 patches per image) and Principle Component Analysis (PCA) to reduce the dimensionality. The IDIAP group duplicated the instances of small classes in the training data in order to reduce the possibilities that the large classes mask the small ones [8].

An important aspect of the evaluation is to understand how much of the performance is based on the visual features used, and how much based on the machine learning techniques. Table 4 shows a comparison to have an idea about the influence of the visual features only. To minimize the impact of machine learning techniques, classifiers for patch–based approach (such as SVM) will be replaced by a simple Euclidean distance, which translates an annotation approach into a retrieval one. GIFT is used as it is to give a baseline. The presumption is: appropriate features should rank images of the same class as "close" without the help from machine learning algorithm. The evaluation is based on the 1000 images in test dataset. For each of them, with selected feature and distance function, 100 nearest images were extracted from the training dataset. The goal is to know among the 1000 images in test dataset, how many of them have found at least one image of the same class. Results were obtained with 100, 30 and 10 nearest

images. The comparison shows that particularly the axis anatomy and thus also for the full code, the patch based features work significantly better.

Table 4. Comparison of GIFT features with a patch–based approach with respect to the number of test images that have at least one exact correspondence in the top N results of the system (on the axis level and for the full code; T=type, modality, D=direction, A=Anatomy, B=Bio system).

Feature	entire code	axis T	axis D	axis A	axis B
in the 100 most similar images					
GIFT	736	996	949	798	987
random patches	821	994	972	882	984
in the 30 most similar images					
GIFT	691	993	919	754	976
random patches	752	990	923	821	980
in the 10 most similar images					
GIFT	621	985	847	682	966
random patches	682	982	870	743	967

On the other hand we can also see, that a large number of images has no correspondence in the top N=10 results for neither of the two feature sets. This means that a large part of the higher performance of these approaches is not due to the features alone but to a combination of features, distance measures and learning approach.

4 Conclusions

For the medical retrieval task only very few purely visual runs (8 runs among 111) were submitted by the participants. The pools of the relevance judgments can thus be slightly biased and even further worsen results such as MAP for these runs. All visual approaches obtain poor scores underlying the high–quality annotations, and tasks that are much more oriented towards text–based approaches. The use of text alone is in our test even better than the combinations with visual retrieval. Few groups actually manage to increase performance with a visual approach over purely textual retrieval. Only early precision can be improved through the combination of textual runs with visual runs. The visual baseline seems to be of insufficient quality for really improving the combined runs significantly and better visual approaches seem necessary. A small number of gray levels still gives best results in our tests.

Differently from previous years, the training dataset and the test dataset do not have the same distribution of classes. Goal of this was to force participants to use the supplied hierarchy for classification including wild cards [10]. An analysis on the wild card frequency of participants is also given in the overview article, indicating a relationship between the wild card frequency and the number of

training images available. The difference between our runs and the best techniques was reduced compared to previous years. The voting–based approaches perform generally better than the simple kNN approaches. Classifying each axis separately with a suitable threshold gives best results in our tests. When the threshold cannot be reached in the first step, chopping the lowest level and repeating the classification for the remaining levels can improve the result slightly. The advantage of the chopping strategy is that the classification is repeated iteratively. High threshold values increase the confidence without totally blocking the classification. The idea of the IDIAP group of oversampling the small classes in the training data is easy to implement and considerably increases performance.

Acknowledgments

This study was partially supported by the Swiss National Science Foundation (Grant 200020–118638/1), the HES SO with the BeMeVIS project, and the European Union in the 6th Framework Program through the KnowARC project (Grant IST 032691).

References

1. 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: CLEF 2007 Proceedings. Volume 5152 of Lecture Notes in Computer Science (LNCS)., Budapest, Hungary, Springer (2008)
2. Müller, H., Kalpathy-Cramer, J., Kahn Jr., C.E., Hatt, W., Bedrick, S., Hersh, W.: Overview of the ImageCLEFmed 2008 medical image retrieval task. In Peters, C., Giampiccolo, D., Ferro, N., Petras, V., Gonzalo, J., Peñas, A., Deselaers, T., Mandl, T., Jones, G., Kurimo, M., eds.: Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum. Lecture Notes in Computer Science, Aarhus, Denmark (September 2009 – to appear)
3. Gobeill, J., Ruch, P., Zhou, X.: Text-only cross–language image search at medical ImageCLEF 2008. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (September 2008)
4. 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**(13–14) (2000) 1193–1198 B.K. Ersboll, P. Johansen, Eds.
5. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1) (1991) 11–32
6. Zhou, X., Gobeill, J., Ruch, P., Müller, H.: University and hospitals of geneva at imageclef 2007. In: CLEF 2007 Proceedings. Volume 5152 of Lecture Notes in Computer Science (LNCS)., Budapest, Hungary, Springer (2008) 649–656
7. Zhou, X., Depeursinge, A., Müller, H.: Hierarchical classification using a frequency–based weighting and simple visual features. Pattern Recognition Letters **29**(15) (2008) 2011–2017

8. Tommasi, T., Orabona, F., Caputo, B.: CLEF2008 image annotation task: an SVM confidence–based approach. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (Sep. 2008)
9. Avni, U., Goldberger, J., Greenspan, H.: TAU MIPLAB at ImageClef 2008. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (Sep. 2008)
10. Deselaers, T., Deserno, T.M.: Medical image annotation in ImageCLEF 2008. In Peters, C., Giampiccolo, D., Ferro, N., Petras, V., Gonzalo, J., Peñas, A., Deselaers, T., Mandl, T., Jones, G., Kurimo, M., eds.: Evaluating Systems for Multilingual and Multimodal Information Access — 9th Workshop of the Cross–Language Evaluation Forum. Lecture Notes in Computer Science, Aarhus, Denmark (September 2009 – to appear)