MedGIFT at ImageCLEF 2008

Xin Zhou[1], Julien Gobeill[1], Henning Müller[1 2],
[1] Medical Informatics, Geneva University Hospitals and University of Geneva, Switzerland
[2] University of Applied Sciences Western Switzerland (HES SO), Sierre, Switzerland
xin.zhou@sim.hcuge.ch

Abstract

This article describes the participation of the Geneva University Hospitals and the University of Geneva at the 2008 ImageCLEF image retrieval benchmark. We concentrated on the two tasks concerning medical imaging. The visual information analysis is based on the GNU Image Finding Tool (GIFT). Other information such as textual information and aspect ratio are integrated to improve the results. The main techniques are the same as in past years, with a little tuning to slightly improve results.

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 five years in ImageCLEF. Due to time constraints no optimizations could be performed and no relevance feedback was used, usually one of the strong points of GIFT.

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 Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image Retrieval, Text categorization, Image Classification, Medical Imaging

1 Introduction

The Geneva University Hospitals and the University of Geneva contribute regularly to the ImageCLEF[1] campaign. The domains of interest are medical image retrieval and medical image annotation [2].

2 Retrieval strategies

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

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

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. It is very similar to approaches in pas years, where queries and documents are tranlated into MeSH (Medical Subject Headings) terms.

2.2 Visual retrieval techniques

The technology used for the visual retrieval is mainly taken from the *Viper*[2] project [3]. 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 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 in various directions and scales.

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 [4].

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 for the medical retrieval task are shown in Table 1 highlighting the most important performance measures such as MAP, Bpref, and early precision. 3 purely visual retrieval runs using GIFT with 4 gray levels (*GIFT4*), 8 gray levels (*GIFT8*), 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 decreases basically all performance measures.

Purely visual retrieval results proved to be little robust [6]. Thus, more effort was invested into mixing visual retrieval and textual retrieval. The textual retrieval run (*HUG–BL–EN*) was provided our collaborator [1], we use it to combine with our best visual run (*GIFT8*). In total 5 mixed–media automatic runs were generated with the following combination strategies:

- combining textual and visual runs with equal weight (*GIFT8_EN0.5*);
- reordering the textual runs based on a visual run (*EN_reGIFT8*);
- mixing two runs by giving varying weights based on the topic, for visual topics, the visual run is weighted 90%, for textual topics, the visual run is weighted 10%, and for mixed topics, the visual run is weighted 50% (*EN_GIFT8_mix*);

[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

- combining textual and visual runs but favoring the text (90%) over visual information (10%) (*GIFT8_EN0.9*);
- reordering the visual runs based on a textual run (*GIFT8_reEN*).

Mixing two runs with varying weights based on the topics (*EN_GIFT8_mix*) gives second best early precision (P30), and third best MAP among the 5 runs. The best MAP is given 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. The work of this year followed work performed in 2007 [5]. The techniques showed to be stable. Adding aspect ratio as feature and performing annotation by axis were reused for our participation in 2008 as well. Main new approaches were a modified classification strategy and changed parameter settings.

The annotation is based on the known labels of similar images retrieved by the GIFT system. In [5], the classification strategies were regrouped around a kNN 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 code.

One problem of classifying images with similar images with known labels is that the classification strategy 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. As a result, the disadvantages of small classes are reduced.

Another useful information is the hierarchy information inside each code axis (there are four in the IRMA code). The output of classification per axis is usually an entire axis or a wildcard 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 gives the possibility to use less wildcards in the classification process and thus can potentially improve the score.

The results of our submitted 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 gives the

Table 2: Results of the runs submitted 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

Table 3: Classification per axis with and without a ”chopping” strategy with descending vote.

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

best results in our tests. Three submitted runs use a voting–based approach as described in [5], respectively per axis with descending vote (*vad*), per axis with chopping letter by letter with descending vote (*vcad*), and per axis with chopping letter by letter using equal weights (*vca*). The thresholds were all set to 0.5 and we submit the runs which take into account the first 5 similar images. The best result among the submitted 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. As the difference between the strategies with and without ”chopping” is not significant, a further comparison is given and the results are presented in Table 3. Chopping at the lowest level and redoing the classification performs better but only with a high threshold.

4 Discussion

For the medical retrieval task the use of text alone is still better than our combinations with visual retrieval, which means that combination techniques still need significant work to preform reasonably well and stable. 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. A small number of colors still gives best results. For a significant improvement in visual retrieval quality new visual features seem necessary.

For the classification of images the difference between our runs and the best techniques is 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 always good results. When the threshold cannot be reached in the first step, chopping the lowest level and redoing the classification for the remaining levels can further improve the result significantly. The advantage of the ”chopping” strategy is that the classification is redone iteratively, thus high threshold values increase the confidence without totally blocking the classification.

Acknowledgments

This study was partially supported by the Swiss National Science Foundation (Grants 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] Julien Gobeill, Patrick Ruch, and Xin Zhou. Text-only cross-language image search at medical imageclef 2008. In *Working Notes of the 2008 CLEF Workshop*, Aarhus, Denmark, September 2008.

[2] 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 *CLEF 2007 Proceedings*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, Budapest, Hungary, 2008. Springer.

[3] 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.

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

[5] Xin Zhou, Adrien Depeursinge, and Henning Müller. Hierarchical classification using a frequency–based weighting and simple visual features. *Pattern Recognition Letters*, Special Issue on Medical Image Annotation in ImageCLEF 2007, 2008–to appear.

[6] Xin Zhou, Julien Gobeill, Patrick Ruch, and Henning Müller. University and hospitals of geneva at imageclef 2007. In *CLEF 2007 Proceedings*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, Budapest, Hungary, 2008–to appear. Springer.