

The MedGIFT group at ImageCLEF 2009

Xin Zhou¹, Ivan Eggel², 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

MedGIFT is a medical imaging research group of the Geneva University Hospitals and the University of Geneva, Switzerland. Since 2004, the medGIFT group has participated in the ImageCLEF benchmark each year, focusing mainly on the medical imaging tasks.

For the medical image retrieval task, two existing retrieval engines were used: the GNU Image Finding Tool (GIFT) as image retrieval engine and Apache Lucene as textual retrieval engine. To improve the retrieval performance, automatic query expansion was used. In total, 13 runs were submitted as well for the image-based topics and the case-based topics. The baseline setup used for the past five years already obtained the best result among all our visual submissions.

For the medical image annotation task, two approaches were tested. One approach is using GIFT for image retrieval and kNN (k-Nearest Neighbors) for the classification, which was already used for the past 5 years. The second approach used Scale-Invariant Feature Transform (SIFT) technology with a Support Vector Machine (SVM) classifier. Three runs were submitted in total, two with the GIFT-kNN-based approach and one using a combination of the SIFT-SVM-based approach and GIFT-kNN-based approach. For the medical image classification task, the GIFT-kNN-based approach gives stable results, although not in the quality of the best groups. The SIFT-SVM-based approach implementation did not achieve the expected better performance. We think that the SVM kernel may be a key factor that requires further optimization.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

General Terms

Image Retrieval, image classification, medical imaging

Keywords

Image Retrieval, Image Classification, Medical Imaging

1 Introduction

A medical retrieval task has been part of ImageCLEF¹ since 2004 [1, 2, 5] (Image retrieval task at the Cross Language Evaluation Forum). The MedGIFT² research group has participated in all

¹<http://www.imageclef.org/>

²<http://www.sim.hcuge.ch/medgift/>

these competitions using the same technology as a baseline and trying to improve the performance of this baseline over time using simple modifications. The GIFT³ (GNU Image Finding Tool, [6]) has been the technology used for visual information retrieval. Visual runs using GIFT have also been made available to other participants of ImageCLEF who do not possess visual retrieval technology.

For textual retrieval the Lucene⁴ system was employed in 2009. The full text of the articles was indexed and the query texts were not treated for the retrieval at all but used as they were given.

More information concerning the setup and the collections of the medical retrieval task of 2009 can be read in [4]. The following sections describe first the retrieval tools used, and then present the results in comparison with the best techniques of the ImageCLEF benchmark.

2 Retrieval Tools Reused

This section describes the basic technologies used for retrieval. The modifications to the base technologies will be detailed in the results section.

2.1 Text Retrieval Approach

The text retrieval approach used in 2009 is based on the Apache Lucene retrieval engine. No specific terms such as MeSH (Medical Subject Headings) were used. Only one textual run was submitted. Five mixed visual/textual runs were submitted combining the textual retrieval with visual retrieval in various ways. The texts were indexed entirely from the html which contains the articles used by ImageCLEFmed task, removing all links and metadata keeping only the text. The query text was not modified, either.

2.2 Visual Retrieval Techniques

GIFT has been used for the visual retrieval for the past five years. This tool is open source and can be used by other participants of ImageCLEF as well. The goal of using the standard GIFT is also to provide a baseline to facilitate the evaluation of other techniques. GIFT uses a simple partitioning of the image into fixed regions to obtain local features.

The performance obtained by GIFT remained limited over the past years. Various strategies were applied in order to improve the results. Aspect-ratio was integrated as feature and simple query expansion with the most similar images was added for the retrieval task [3]. For the annotation task a technique to treat the annotation axes of the multi-axial IRMA (Image Retrieval in Medical Applications) code of the images separately was employed [7], and also a dynamic kNN approach for classification [8]. The strategies used in ImageCLEFmed 2009 are detailed in Section 3.

3 Results

In this section the results and technical details for the two medical tasks of ImageCLEF 2009 are detailed.

3.1 Medical Image Retrieval Task

All the runs submitted for evaluation are based on visual retrieval using GIFT with 8 gray levels as this had consistently given best results in past years. Various strategies to improve the results were tried out. A simple improvement strategy is to separately query the images belonging to one topic as these images are quite often dissimilar. Then, the result can be combined in a late

³<http://www.gnu.org/software/gift/>

⁴<http://lucene.apache.org/>

fusion. Another approach is to supply images of other topics as negative feedback assuming that this can create more discriminative queries.

The following steps were thus performed:

- for each topic, the similarities between all query image pairs were calculated;
- based on the similarity between the images pairs, queries are treated either as a query of multiple images or as several separate queries;
- for each topic, a random selection of 2 images from other topics were added as negative feedback;

In total, 16 automatic runs were submitted. 10 runs were submitted for the image-based retrieval topics and 6 for the case-based topics. These runs can be divided into 2 textual, 10 visual, and 4 mixed runs. Run identifiers are composed of the strategies applied to the run. The following list explains the labels and their meaning.

- *txt*: textual;
- *vis*: visual;
- *mix*: mixed;
- *sep*: separate queries for the various images of a topic;
- *withAR*: aspect ratio added as feature;
- *withNegImg*: query expansion with negative examples of other topics;
- *sum*: simple sum to combine scores of several queries;
- *max*: maximum score of an image is taken to combine several queries;
- *0.x* for mixed runs, 0.x is the weight for the visual retrieval and (1-0.x) for the textual retrieval;
- *EN* the language used for textual retrieval is English;
- *BySim* For the case-based topics the similarity score of images appearing in the case is used;
- *ByFreq* For the case-based topics the frequency of images of a case in the first 1000 results is used.

Results for the medical retrieval task are shown in two parts. The results of 25 image-based topics are shown in Table 1 and those of the case-based topics (topics 26–30) are shown in Table 2. Mean average precision (MAP), binary preference (Bpref), and early precisions (P10, P30) were selected for our tables. The ImageCLEFmed 2009 retrieval task has an unbalanced number of visual, mixed, and textual runs. We submitted no manual or feedback runs at all.

3.1.1 Image-Based Topics

As in past years, most of the submissions are textual or mixed and only few groups worked on improving the visual runs. Only 5 groups submitted a total of 15 visual runs. All our runs and the best runs in each category can be seen in Table 1 Our best visual run is the standard GIFT with 8 gray levels that was not submitted for the competition but has actually the best MAP. The other submitted runs give worse results but in general the performance of all runs is rather low with a MAP below 2%. Also the other measures such as early precision and number of relevant images found are much lower for the purely visual runs. None of the visual runs obtains extremely good results.

Table 1: Results of the runs for the image-based retrieval topics.

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
best textual	Textual	0.4293	0.4568	0.664	0.552	1814
HES-SO-VS_txt_EN	Textual	0.3179	0.3498	0.600	0.4987	1462
best visual run	Visual	0.0136	0.0363	0.072	0.0507	295
medGIFT_vis_GIFT8	Visual	0.0153	0.0347	0.068	0.0467	284
medGIFT_vis_sep_max	Visual	0.0131	0.0276	0.076	0.056	266
medGIFT_vis_sep_sum_withAR	Visual	0.013	0.0303	0.072	0.052	262
medGIFT_vis_sep_sum	Visual	0.0114	0.0282	0.052	0.0573	259
medGIFT_vis_sep_max_withAR	Visual	0.0102	0.0303	0.076	0.0547	253
medGIFT_vis_sum_withNegImg	Visual	0.0098	0.028	0.044	0.053	210
medGIFT_vis_max_withNegImg	Visual	0.0079	0.0248	0.044	0.044	201
best mixed run	Mixed	0.3682	0.386	0.544	0.4827	1754
medGIFT_mix_0.3withNegImg_EN	Mixed	0.29	0.3216	0.604	0.516	1176
medGIFT_mix_0.5_EN	Mixed	0.2097	0.2456	0.592	0.4293	848
medGIFT_mix_0.5withNegImg_EN	Mixed	0.1354	0.1691	0.488	0.3267	547

Table 2: Results of the runs for the case-based retrieval topics.

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
best run	Textual	0.3355	0.2766	0.34	0.2267	74
HES-SO-VS_txt_case	Textual	0.1906	0.1531	0.32	0.2	71
medGIFT_mix_0.5BySim_EN	Mixed	0.0655	0.0488	0.14	0.0867	74
medGIFT_vis_maxBySim_withAR	Visual	0.021	0.029	0.04	0.0533	41
medGIFT_vis_sumBySim_withAR	Visual	0.019	0.026	0.06	0.0533	42
medGIFT_vis_maxByFreq_withAR	Visual	0.0025	0.0035	0	0.0067	26
medGIFT_vis_sumByFreq_withAR	Visual	0.0025	0.0035	0	0.0067	26

A total of 52 textual runs were submitted. We only submitted a single textual run using an extremely simple strategy of Lucene with no pre-treatment of neither queries nor documents. The results looking at MAP are not as good as other approaches but early precision is fairly good and looking at the simplicity of the approach this is rather surprising. The performance is above the average of the textual runs.

In total, 25 mixed runs were submitted. We submitted three mixed runs that have a large variety of combination strategies. It can be seen that the combination strategy has a very strong influence on the results. Our best mixed run is slightly better in early precision than our text runs but worse in all other measures. In early precision our mixed run can rival the best runs submitted and reaches the second overall position.

3.1.2 Case-based topics

In Table 2, all our runs for the case-based topics are shown and the best run of the competition for comparison. Again it can be seen that the pure text retrieval has by far the best results, despite the simplicity of our approach. Our mixed approach was worse in all categories. The purely visual runs have a fairly low MAP similar to the image-based topics and even the early precision was rather low.

In both the visual and mixed categories our system performed best but this is also partly due to the low participation.

Table 3: Results of the runs submitted to the medical image annotation task.

run ID	2005	2006	2007	2008	SUM
best system	356	263	64.3	169.5	852.8
GE_GIFT8_AR0.2_vdca5_th0.5.run	618	507	190.73	317.53	1633.26
GE_GIFT16_AR0.1_vdca5_th0.5.run	641	527	210.93	380.41	1759.34
GE_GIFT8_SIFT_commun.run	791.5	612.5	272.69	420.91	2097.6

3.2 Medical Image Annotation

For the medical image annotation task 6 groups submitted in total 18 runs. The medGIFT group submitted 3 runs. Two runs used the same strategy as in the past 2 years using GIFT as a basic retrieval technology, then adding aspect ratio to the similarity score. For classification the 5 nearest neighbors (5NN) was used. More details on the strategy can also be found in [7, 8]. One run was submitted to test a SIFT-SVM based approach. The standard Gaussian kernel is used for the SVM classifier. Another run combined the results of the SVM run and a standard run. The results obtained by the three submitted runs are shown in Table 3. The simple 5NN approach using GIFT with 8 grey levels gave consistently the best results over all tasks. The test of the SVM classifier did not reach the performance hoped for but much if this can be due to the lack of optimization performed on the level of the Kernel. The results of GIFT can not be seen as much more than a simple baseline.

4 Discussion

Textual runs significantly out-performed visual runs, which was coherent with past experiences. Even using the basic Lucene with no pretreatment obtained fairly good results compared to the much more complex approaches of other groups. Since 2004 GIFT with 8 gray levels has been used as a baseline. Generally, the result obtained were below the average MAP of purely visual runs. The strategy to perform separate queries for each query image of a topic increased early precision but decreased MAP: Using aspect ration for the retrieval has a slight influence on early precision but is otherwise not better. Combinations of text and visual result can obtain a higher early precision.

For the image annotation task a new SVM-based strategy was employed but did not increase results at all. The basic GIFT strategy still had our best results but not comparable with the results of the best groups. The combination of using GIFT features and SIFT did not improve results either. Using a small number of results for the combination does not seem appropriate.

5 Conclusion

The paper summarized the participation of the medGIFT group in ImageCLEF2009. The medical image retrieval and medical image annotation tasks were addressed. Results show that textual retrieval techniques outperform visual techniques but that visual retrieval can be an addition to improve early precision of retrieval.

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] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 cross-language image retrieval track. In *Cross Language Evaluation Forum (CLEF 2005)*, Springer Lecture Notes in Computer Science, pages 535–557, September 2006.
- [2] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross-language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael 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] Tobias Gass, Antoine Geissbuhler, and Henning Müller. Learning a frequency-based weighting for medical image classification. In *Medical Imaging and Medical Informatics (MIMI) 2007*, pages 137–147, Beijing, China, 2007.
- [4] Henning Müller, Jayashree Kalpathy-Cramer, Ivan Eggers, Steven Bedrick, Radhouani Said, Brian Bakke, Charles E. Kahn Jr., and William Hersh. Overview of the 2009 medical image retrieval task. In *Working Notes of CLEF 2009 (Cross Language Evaluation Forum)*, Corfu, Greece, September 2009.
- [5] Henning Müller, Jayashree Kalpathy-Cramer, Charles E. Kahn Jr., William Hatt, Steven Bedrick, and William Hersh. Overview of the ImageCLEFmed 2008 medical image retrieval task. In Carol Peters, Danilo Giampiccolo, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Mikko Kurimo, editors, *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.
- [6] 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.
- [7] Xin Zhou, Adrien Depeursinge, and Henning Müller. Hierarchical classification using a frequency-based weighting and simple visual features. *Pattern Recognition Letters*, 29(15):2011–2017, 2008.
- [8] Xin Zhou, Julien Gobeill, and Henning Müller. The medgift group at imageclef 2008. In *CLEF 2008 Proceedings*, Lecture Notes in Computer Science (LNCS), Aarhus, Denmark, 2009 – submitted. Springer.