The MedGIFT group at ImageCLEF 2009

Xin Zhou[1], Ivan Eggel[2], Henning Müller[12],
[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. Baseline setup used for the last three years obtained the best result among all our submissions.

For the medical image annotation task, two approaches were tested. One approach is using GIFT for similar image retrieval and kNN (k-Nearest Neighbors) for the classification, which has already been used for the past 4 years. The second approach used Scale–Invariant Feature Transform (SIFT) technology with a Support Vector Machines (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 medical image classification task, the GIFT–kNN–based approach gives stable results just as for the last 4 years. The SIFT–SVM–based approach implementation did not achieve the expected better performance. We believe that the SVM kernel seems is the 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 ImageCLEF[1] since 2004 [3, 4, 8]. The MedGIFT[2] research group has participated in all these competitions using the same technology as a baseline and trying to improve the performance of this baseline over time. The GIFT[3] (GNU Image Finding Tool, [9]) has been the technology used for visual information retrieval. Visual runs using GIFT have also been made available to other participants of ImageCLEF.

For textual retrieval the Lucene[4] system was employed in 2009. The full text of the articles was indexed.

More information concerning the setup and the collections of the medical retrieval task of 2009 can be red in [7]. 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 partitioning of the image into fixed regions to obtain local features.

During the last 3 years, the performance obtained by GIFT remained unsatisfying. Various strategies were tried out in order to get improvements: integrate aspect–ratio as addition features, add the most similar images to form an automatic query expansion [5], best threshold searching for each label axis for the annotation task [11], using dynamic kNN approach for classification [12], etc. The improvement strategies used in ImageCLEFmed 2009 are detailed in Section 3.

Best visual results obtained both in 2008 used local–feature–based feature space and machine–learning–based similarity distance metric. Popular combination is to use Scale–Invariant Feature Transform(SIFT) [6] with Support Vector Machine(SVM) [2], which obtained the best results [1, 10] in medical annotation task. A preliminary comparison of various feature spaces using the same distance metrics on ImageCLEF annotation dataset was done in [12] to understand to what percentage of difference comes from the feature space, and to what percentage comes from the distance metric. In ImageCLEFmed2009, medGIFT decided to use both SIFT–SVM–based and GIFT–kNN–based strategies for the ImageCLEF annotation task.

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

[2] http://www.sim.hcuge.ch/medgift/

[3] http://www.gnu.org/software/gift/

[4] http://lucene.apache.org/

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

All the runs submitted for evaluation are based on visual retrieval using GIFT with 8 gray levels. Several runs are submitted as various improvement strategies were tried out. By visual observation, more than half topics contain the images which are not visually similar. An evident improvement strategy is to query separately the images belonging to one topic, and then to combine the result together. Another thought is based on the fact that images assigned to different topics should be semantically different. Thus randomly selection of images from the other topics should give valuable negative examples for the query.

In one word, the improvement strategy focused on two key points based on the nature of ImageCLEFmed retrieval task :

- whether images of each topic are visually similar;
- whether images of different topics are not visually similar;

The improvement is following this route:

- for each query, the similarities between all query image pairs are measured;
- queries are treated either as a query of multiple images or as several separated queries according to the similarity analysis;
- randomly select 2 images from other queries and put them as negative images to extend the queries;

In total, 16 automatic runs were submitted : 10 for ad–hoc retrieval topics and 6 for the case–based retrieval topics. Technically these runs can be divided into 2 textual–based, 10 visual–based and 4 mixed–media based. Runs are labeled by the different improvement strategies. The employed labels and their signification are:

- *txt* textual based retrieval;
- *vis* visual based retrieval;
- *mix* combination of textual retrieval and visual retrieval;
- *sep* one query per image is performed to produce a list of similar images for each image;
- *withAR* combining the aspect ratio as an additional feature set;
- *withNegImg* query expansion by taking some images from other topic as negative examples;
- *sum* a basic result fusion strategy: if one item has several similarity scores, sum them up as its similarity score for reordering;
- *max* a basic result fusion strategy: if one item has several similarity scores, take the max value as its similarity score for reordering;
- *0.x* for a mixed run, 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 result fusion, each result is weighted by the similarity score given by the research engine;

Table 1: Results of the runs for the ad–hoc 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	0.3738	0.3883	0.56	0.5053	1803
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

- *ByFreq* for result fusion, each result is weighted by the frequency of appearance.

Results for the medical retrieval task are shown in two parts. The results of 25 ad–hoc topics are shown in Table 1 and those of 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 evaluation. ImageCLEFmed retrieval task has unbalanced number of visual runs and textual runs. To avoid biasing the evaluation, 2362 relevant images for ad–hoc topics and 95 cases for case–based retrieval topics were manually selected by clinical users to provide a standard as performance evaluation. The number of number of relevant results returned (num_rel_ret) is thus a significant criteria for the evaluations as it is independent to the users' submissions.

3.1.1 Ad–hoc topics

59 textual runs were submitted for ImageCLEFmed 2009, the best run found 1814 relevant images (77% of total relevant images). On average, each run found 1418 relevant images. Lucene research engine with standard setup (*HES-SO-VS_txt.txt*) performs slightly better than the average.

Only 5 groups submitted 16 visual runs. Our best run is the baseline that used GIFT with 8 gray levels. It is ranked as second. The other submitted runs with additional improvement strategies give worse results. The MAP of all the visual runs are below 2%, which is slightly biased by the numerous textual retrieval results. However, even using the number of found relevant results, visual runs are largely behind the textual runs. Only an average of 200 relevant images were found. The best run found around 300 relevant images.

There are 29 mixed textual/visual runs. The average of the number of relevant images returned is 1108. Our best mixed run is slightly above the average. The best mixed run found 1803 relevant images, which is close to the best text run.

3.1.2 Case–based topics

In Table 2, instead of listing the best run, three best runs are shown as they can be all evaluated as best according to the selected evaluation metric. One of the best runs found almost all the relevant cases. Lucene with standard configuration can find 71 relevant cases. 2 groups submitted 5 purely visual runs in total, among which 4 runs are from medGIFT. The combination of the visual and textual technologies gives slight improvement on number of relevant images returned compared to textual retrieval alone.

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

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
ceb-cases-essie2-automatic	Textual	0.3355	0.2766	0.34	0.2267	74
sinai_TA_cbt	Textual	0.2626	0.2264	0.34	0.2267	89
aueb_ipl	Textual	0.1912	0.1252	0.24	0.1867	93
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

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

6 groups submitted 18 runs in total for annotation task. Among them 3 runs were submitted by medGIFT. Two runs used the same strategy as the past 2 years:

- using GIFT to find a list of similar images;
- reordering the list by integrating the aspect ratio;
- 5 nearest neighbors (5NN) were used to do the classification for each axis by voting using a descending weight.

Details can be found in the paper of ImageCLEFmed 2007 [11] and 2008 [12]. One run is submitted to test the SIFT–SVM based approach. The standard Gaussian kernel is used with SVM. For the lack of SVM optimization experience, instead of submitting the run obtained by SIFT–SVM approach, a combined run of GIFT16–5NN and SIFT–SVM was selected. In both list of similar images, 15 common similar images are selected for each test image. The results are shown in Table 3. The results of using GIFT–5NN approach obtained coherent results with the results obtained in the last two years. The common results obtained by GIFT–5NN and SIFT–SVM gives worse result, which will be addressed in the discussion.

4 Discussion

4.1 Medical image retrieval

Textual runs significantly out–performed visual runs, which was coherent with the past experiences. Considering the performance gap between the textual runs and visual runs, the score for mixed runs depend mainly on the performance of textual runs. By combining the visual runs, the performance (MAP) is not highly improved, but the early precision is better then the textual retrieval alone.

Since 2006 GIFT with 8 gray levels has been used as a baseline. Generally the result obtained were below the average MAP of purely visual runs, but the MAP was around 3%. This year the baseline run with lower performance is better ranked, which is a indicator of decreasing quality for visual retrieval.

In total 8 visual runs were submitted by medGIFT group for ad–hoc topics. Except for the baseline run, all the visual runs are all based on the strategy that each image was queried separately. This strategy has been evaluated visually by observation of the first 50 results for each query and proved to provide better result. Surprisingly the improvement strategies decrease significantly the performances. It is probably because the ”improvement strategies” are only optimization for the early precision, P10 and P30 were indeed improved, but less relevant images were found and the average precision became low. Using aspect ratio improved the performance in ImageCLEF medical annotation task. As the annotation task uses only the 5 15 most similar images to do the classification [11], this is also an ”optimization” only valid for early precision.

It can be also identified that the performance gap between visual retrieval and textual retrieval of case–based topics is much smaller than the one for the ad–hoc topics. For image–based topics, the number of relevant images found is 1/10 of textual retrieval. For the case–based topics, it is 1/2 only. Textual retrieval relies on the semantic information, which already contains the information that which image belongs to the same case. It is predictable that integrating ”case information” should improve the retrieval performance for the purely visual runs.

4.2 Medical image annotation

This year the SIFT–SVM approach is used for getting better result. However, the SIFT–SVM performs worse than GIFT–kNN approach in the training stage. The GIFT–kNN approach has been used for 3 years and many optimization strategies were added. Less experience is cumulated for SIFT–SVM approach and further optimization is required.

One of the idea is to combine GIFT and SIFT based approach by selecting the common similar images. Yet this strategy gave worse result than using GIFT alone. By analyzing the several test images and assigned code, a simple reason behind this is that the fusion strategy is not adapted. We selected 15 common images which appeared in both list of similar images. However, with different feature spaces, GIFT–based runs and SIFT–based runs have very different similar images. The common images are not really relevant for both sides, which decreases the performance of classification. Moreover, the distance metric used for SIFT–SVM approach is a probability value, whereas GIFT–based approach used histogram intersection. Usually a small difference of histogram intersection value is not significant, but a small difference of probability can refer to a big difference. Even result fusion used normalization for each distance, the fusion itself is only a linear combination.

Since 2007, other groups used SIFT–SVM approach which out–performed all the other groups in this task. The best result of annotation task for this year is also based on SIFT–SVM approach. A discussion with the group which got best result in 2007 and 2008 using SIFT–SVM approach was organized to find out the optimization they used. The main different strategies are the following:

- Other group used 5 test images for each class, and all the rest is used for training. Therefore the test set is balanced for all the classes. We used a standard 50%–50% for cross–validation, fewer training images are used and the evaluation is biased to the big class;

- One key factor of getting worse result is the SVM kernel. It was told that ”chi square” kernel largely outperforms standard Gaussian kernel;

- Only the two most possible results given by SVM are taken into account as SVM gives precise result. We used the first 5 similar images, which increase the error possibilities;

- The smallest classes which has less than 10 images will be regrouped with the other large groups.

5 Conclusion & future work

The paper summarized the participation of medGIFT group in ImageCLEF2009 competition. Medical image retrieval and medical image annotation tasks were addressed.

The preliminary analysis of the results shows that visual retrieval is able to improve the early precision, but a big weakness of ”fusion” strategies exists for both tasks. SIFT–SVM based approach continues to show its potential in ImageCLEFmed competition and worths more investigations.

Future work included several possibilities for improvement:

- Using ”case–based” retrieval strategy for the ad–hoc topics of medical image retrieval task;
- Using ”chi square” kernel and 99% of training images to train the SVM for annotation task.

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] Uri Avni, Jacob Goldberger, and Hayit Greenspan. TAU MIPLAB at ImageClef 2008. In *Working Notes of the 2008 CLEF Workshop*, Aarhus, Denmark, Sep. 2008.

[2] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] 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.

[4] 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.

[5] 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.

[6] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] 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.

[8] 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.

[9] 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.

[10] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. CLEF2008 image annotation task: an SVM confidence–based approach. In *Working Notes of the 2008 CLEF Workshop*, Aarhus, Denmark, Sep. 2008.

[11] 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.

[12] 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.