The MedGIFT Group at ImageCLEF 2009

Xin Zhou[1], Ivan Eggel[2], 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. MedGIFT is a medical imaging group of the Geneva University Hospitals and the University of Geneva, Switzerland. Since 2004, the group has participated ImageCLEF each year, focusing on the medical imaging tasks. For the medical image retrieval task, two existing retrieval engines were used: the GNU Image Finding Tool (GIFT) for visual retrieval and Apache Lucene for text. Various strategies were applied to improve the retrieval performance. In total, 16 runs were submitted, 10 for the image–based topics and 6 for the case–based topics. The baseline GIFT setup used for the past three years obtained the best results among all our submissions.

For medical image annotation two approaches were tested. One approach is using GIFT for retrieval and kNN (k–Nearest Neighbors) for classification. The second approach used the Scale–Invariant Feature Transform (SIFT) with a Support Vector Machine (SVM) classifier. Three runs were submitted, two with the GIFT–kNN approach and one using the common results of the two approaches. The GIFT–kNN approach gave stable results. The SIFT–SVM approach did not achieve the expected performance, most likely due to the SVM Kernel used that was not optimized.

1 Introduction

A medical retrieval task has been part of ImageCLEF[1] since 2004 [1, 2]. The MedGIFT[2] research group has participated in all these competitions using the same technology as a baseline and tried to improve the performance of this baseline over time. The GIFT[3] (GNU Image Finding Tool, [3]) has been the technology used for visual retrieval. Visual runs using GIFT have also been made available to other participants of ImageCLEF. For text retrieval, Lucene[4] was employed in 2009. The full text of the articles was indexed with no optimization. More information concerning the setup and collections of the medical retrieval task can be found in [4].

2 Retrieval Tools Reused

This section describes the basic technologies used for retrieval.

[1] http://www.imageclef.org/
[2] http://www.sim.hcuge.ch/medgift/
[3] http://www.gnu.org/software/gift/
[4] http://lucene.apache.org/

2.1 Text Retrieval Approach

The text retrieval used in 2009 is based on Lucene. No specific terminologies such as MeSH (Medical Subject Headings) were used. Only one textual run was submitted. The texts were indexed entirely from the HTML (Hyper Text Markup Language), removing links and metadata. The query text was not modified.

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 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, such as integration of aspect–ratio as feature, automatic query expansion and threshold optimization for axes for the annotation task. In ImageCLEF 2009, query expansion with negative examples was carried out for the image retrieval task, and SIFT features were integrated into the image annotation task.

3 Results

This section describes our main results for the two medical tasks.

3.1 Medical Image Retrieval

All the runs were obtained by using GIFT with 8 gray levels. Various strategies were tried to increse performance. One strategy is to query the images belonging to one topic separately, and then to combine the obtained results. Another strategy is to apply negative feedback using the query images of other topics as we assume that the topics are sufficiently different. Adding aspect–ratio is another feature that has worked well in the past. In total, 16 automatic runs were submitted : 2 textual, 10 visual and 4 mixed. 10 run were for the image–based retrieval topics and 6 for the case–based topics. Runs were labeled by the strategies applied. The labels and their signification are:

- *txt* textual retrieval;
- *vis* visual retrieval;
- *mix* combination of textual and visual retrieval;
- *sep* one query per image is performed to produce a list of similar images for each query image;
- *AR* adding aspect ratio;
- *NgRan* query expansion by randomly taking images from other topics as negative examples;

- *sum* basic results fusion: if one item has several similarity scores, the sum of all scores is used;
- *max* basic results fusion: if one item has several similarity scores, the maximum value is used;
- *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 results fusion, each result is weighted by the similarity score given by Lucene/GIFT;
- *ByFreq* for results fusion, each result is weighted by the number of appearances.

The results of the 25 ad–hoc topics are shown in Table 1 and those of the case–based topics (26–30) are shown in Table 2. Mean average precision (MAP), binary preference (Bpref), and early precisions (P10, P30) are used as measures.

Table 1. Results of the runs for the image–based topics.

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
best textual run (LIRIS)	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
MedGIFT_vis_GIFT8 (best visual run)	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_AR	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_AR	Visual	0.0102	0.0303	0.076	0.0547	253
MedGIFT_vis_sum_negRan	Visual	0.0098	0.028	0.044	0.053	210
MedGIFT_vis_max_negRan	Visual	0.0079	0.0248	0.044	0.044	201
best automatic mixed run (DEU)	Mixed	0.3682	0.386	0.544	0.4827	1753
MedGIFT_mix_0.3NegRan_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_0.5NegRan_EN	Mixed	0.1354	0.1691	0.488	0.3267	547

Image–Based Topics In total, 59 textual runs were submitted for ImageCLEFmed 2009. The average score (MAP) for the textual runs is around 0.3. The Lucene search engine with a standard setup(*HES–SO–VS_txt.txt*) performed slightly better than the average. The best textual runs used mapping of text to Medical Subject Headings (MeSH) or the Unified Medical Language System (UMLS) to reach an improvement [5–7].

5 groups submitted 16 visual runs. Our best run is the baseline that used GIFT with 8 gray levels(*MedGIFT_vis_GIFT8.txt*). The baseline obtained the highest MAP among all visual runs. The run using the one query per image

strategy was officially ranked as second but it outperformed the other visual runs on early precision. As the performance was fairly limited, additional tests were performed and are described in Section 3.1.

The second best visual run was submitted by the Image and Text Integration(ITI) group from the National Library of Medicine. Various low level global features were used and a linear combination of these features was applied [8]. SVMs were used to map visual features to semantic terms based on a predefined visual concept tree built from the consolidated ImageCLEFmed collection. Despite the integration of a visual concept tree with machine learning, the results were not extremely high.

There were 29 mixed textual/visual runs. The MedGIFT runs are among the five best runs. However, as textual runs outperform the visual runs, many mixed runs are not even as good as the corresponding textual runs. Compared with our textual baseline run all mixed runs obtained worse performance. Several other groups had similar conclusions [8, 9]. York University declared that the Color and Edge Directivity Descriptor(CEDD) slightly boosted the performance of a textual run [10]. Both the group from York University and our group used a similar linear combination strategy for fusing the results. Considering the fact that visual runs submitted by York University obtained the worst results among all submitted visual runs, the improvement detected by York University might require further investigation. The best mixed run is from the DEU group, that combined visual and textual features into a single feature matrix [11]. The results show that fusion in the feature space can obtain good results.

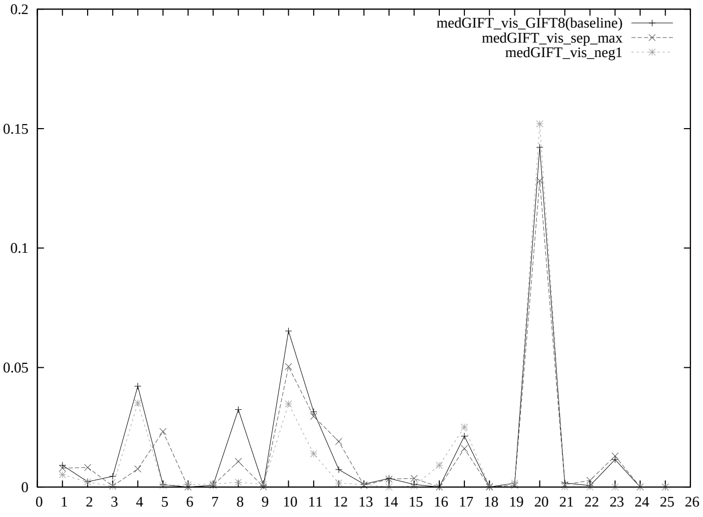

Fig. 1. The performance obtained by GIFT configurations for visual retrieval per topic.

Follow–Up Analyses Follow–up analyzes were performed once the ground–truth was available. Using the one query per image strategy and negative query expansion did not improve visual retrieval. The performance for each topic with the three main techniques is shown in Figure 1. The similarities among topic images for each topic are shown in Figure 2 to show homogeneous and heterogeneous topics. To obtain the similarity among topic images, all topic images were indexed and queries with each topic image were performed. In a pairwise comparison the images of one topic were analyzed. The result shown is the average score among all pairwise per topic using the GIFT baseline run.

For the submitted visual runs with negative query expansion, negative examples were randomly selected. In an additional approach, negative examples were selected based on the similarity score obtained through visual queries. These runs slightly outperformed the submitted runs using negative examples. In Figure 1, the new run using one negative example is also presented.

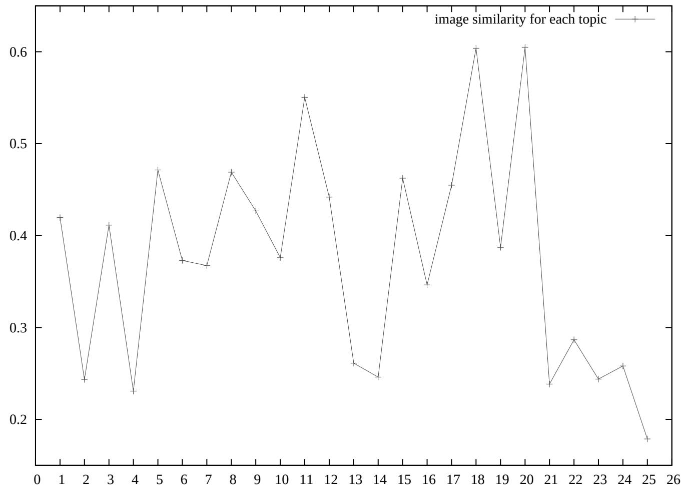

Fig. 2. The similarity among the images of one topic showing whether the images depict a similar topic or different aspects of the search topic.

Comparing the baseline(MedGIFT_vis_GIFT8) with one query per image (MedGIFT_vis_sep_max) shows that the performance for a topic is not correlated with similarity among the topic images. On the one hand, topics 2, 4, 14, 24, 25 contain images with little similarity, with only topic 2 being improved using one query per image. On the other hand, topics 5, 15, 23 contain very similar images and still the one query per image strategy gave significantly better results. For all other topics the baseline obtained better scores.

Using negative examples outperformed the baseline run and the one query per image strategy only rarely (for example topics 6, 16, 18).

Case–Based Topics In 2009, the MedGIFT group submitted 1 mixed run, 4 visual runs and 1 textual run for case–based topics. In total, 11 textual runs, 2 mixed runs and 5 visual runs were submitted for this task. In Table 2, the three best runs of other groups are shown, all of them were textual. The MedGIFT runs were best for visual and mixed retrieval. Our best textual run used Lucene with its standard configuration(*HES-SO-VS_txt_case*). By combining the visual run with a textual run (*MedGIFT_mix_0.5BySim_EN*) the MAP decrease significantly but slightly more relevant cases could be found.

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_AR	Visual	0.021	0.029	0.04	0.0533	41
MedGIFT_vis_sumBySim_AR	Visual	0.019	0.026	0.06	0.0533	42
MedGIFT_vis_maxByFreq_AR	Visual	0.0025	0.0035	0	0.0067	26
MedGIFT_vis_sumByFreq_AR	Visual	0.0025	0.0035	0	0.0067	26

3.2 Medical Image Annotation

In the medical image annotation task 6 groups submitted a total of 18 runs. Three of these runs were submitted by the MedGIFT group. Two runs used the same strategy as in the past 2 years:

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

Details can be found in the papers of ImageCLEF 2007 [12] and 2008 [13]. One run was submitted to test a SIFT–SVM approach. The standard Gaussian kernel was used for the SVMs. No optimizations of the SVMs were tried. As the results of the SIFT–SVM approach were not optimal we used this run in combination with one of our standard runs for the submission. In both cases, the N most similar images were retrieved for each test image and then used for the classification. The results are shown in Table 3. Best results were obtained using GIFT–5NN as in the past years. Using a combination with SIFT–SVM gave worse results.

Two groups (Biomed and IDIAP) submitted runs significantly outperforming all other techniques. Very similar techniques were used as Biomed was inspired from by IDIAP [14]. Their system uses the following approach:

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

run ID	2005	2006	2007	2008	SUM
best system (TAU Biomed)	356	263	64.3	169.5	852.8
second best system (IDIAP)	393	260	67.23	178.93	899.16
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

- extract local features from a sub–set of images using random points;
- use k–means clustering to create a dictionary of visual words;
- sample each image with a denser grid and represent each image as a histogram of the visual words;
- train a classifier using SVMs with a $\mathcal{X}^2$ kernel.

This approach has proven to obtain best results for the past three years.

4 Conclusions

This paper summarizes the participation of the MedGIFT group in ImageCLEF2009. The medical image retrieval and medical image annotation tasks were addressed. A preliminary analysis of our results for the medical retrieval task shows that visual retrieval is able to improve early precision. Overall performance (measure by MAP) of mixed–media runs relied highly on the performance of the textual run. Textual/visual run fusion strategies require further study as currently the MAP of mixed runs is often lower than that of the corresponding textual run.

An additional analysis were carried out to better understand the obtained results. Query performance of a topic is not directly related to the similarity among the images of the topic.

There is still a big gap of performance between textual and visual retrieval. Keywords are naturally linked to semantic topics and this for semantic topics text–based approaches perform much better, although even for the visual topics the text retrieval results obtain better results.

Using SVMs together with local features based on salient points shows to obtain reasonable results but requires further optimization as our obtained results were by far not as good as those groups obtaining the best results.

Acknowledgments

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

References

1. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Cross Language Evaluation Forum (CLEF 2005). Springer Lecture Notes in Computer Science (September 2006) 535–557
2. 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.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613
3. 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.
4. Müller, H., Kalpathy-Cramer, J., Eggers, I., Bedrick, S., Said, R., Bakke, B., Kahn Jr., C.E., Hersh, W.: Overview of the 2009 medical image retrieval task. In: Working Notes of CLEF 2009 (Cross Language Evaluation Forum), Corfu, Greece (September 2009)
5. Maisonnasse, L., Harrathi, F.: Analysis combination and pseudo relevance feedback in conceptual language model. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
6. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.C.: MIRACLE at ImageCLEFmed 2009: Reevaluating strategies for automatic topic expansion. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
7. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A., García-Cumbreras, M.A.: SINAI at ImageCLEF 2009 medical task. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
8. Simpson, M., Rahman, M.M., Demner-Fushman, D., Antani, S., Thoma, G.R.: Text– and content–based approaches to image retrieval for the ImageCLEF 2009 medical retrieval track. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
9. Boutsis, I., Kalamboukis, T.: Combined content–based and semantic image retrieval. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
10. Ye, Z., Huang, X., Lin, H.: Towards a better performance for medical image retrieval using an integrated approach. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
11. Berber, T., Alpkoçak, A.: DEU at ImageCLEFmed 2009: Evaluating re-ranking and integrated retrieval model. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)
12. 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
13. Zhou, X., Gobeill, J., Müller, H.: The medgift group at imageclef 2008. In: CLEF 2008 Proceedings. Volume 5706 of Lecture Notes in Computer Science (LNCS)., Aarhus, Denmark, Springer (2009) 712–718
14. Avni, U., Goldberger, J., Greenspan, H.: TAU MIPLAB at ImageClef 2008. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (Sep. 2008)