The participation of the MedGIFT Group in ImageCLEFmed 2010

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@hevs.ch

Abstract. This article presents the participation of the MedGIFT group in ImageCLEFmed 2010. Since 2004, the group has participated in the medical image retrieval tasks of ImageCLEF (ImageCLEFmed) each year. The main goal is to provide a baseline by using the same technology each year, and to search for further improvement if possible.
There are three types of tasks for ImageCLEFmed 2010: modality classification, image–based retrieval and case–based retrieval. The MedGIFT group participated in all three tasks. For ad–hoc retrieval and case–based retrieval tasks, two existing retrieval engines were used: the GNU Image Finding Tool (GIFT) for visual retrieval and Apache Lucene for text. Fusion strategies were also tried out to combine results from two engines. For the modality classification, a purely visual approach was used with GIFT for the visual retrieval and a kNN (k–Nearest Neighbors) classifier for the classification.
Results show that the best textual run outperforms the best visual run by a factor of 30 in terms of mean average precision. Baselines provided by Apache Lucene and GIFT are ranked above the average among textual runs and visual runs respectively in ad–hoc retrieval. In the case–based retrieval task the Lucene baseline is the third best automatic run. For modality classification, GIFT and the kNN–based approach perform slightly better than the average of visual approaches.

1 Introduction

ImageCLEF is the cross–language image retrieval track[1] of the Cross Language Evaluation Forum (CLEF). ImageCLEFmed is part of ImageCLEF focusing on medical images [1, 2]. The MedGIFT[2] research group has participated in ImageCLEFmed using the same technology as baselines since 2004, with additional modifications of the basic techniques. Visual and textual baseline runs have been made available to other participants of ImageCLEFmed. The visual baseline is based on GIFT[3] (GNU Image Finding Tool, [3]) whereas Lucene[4] was used for textual 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 Methods

This section describes the basic techniques used for retrieval in ImageCLEFmed2010.

2.1 Retrieval Tools Reused

Text Retrieval The text retrieval approach in 2010 is based on Lucene using standard settings. 4 textual runs were submitted, 2 for case–based retrieval and 2 for image–based retrieval. For case– and image–based retrieval, image and full text were used.

The full text approach used all texts as downloaded as HTML. Links, metadata, scripts and style information were removed and only the remaining text was indexed. For image captions, an XML file containing captions of all the images was indexed. No specific terminologies such as MeSH (Medical Subject Headings) were used.

Visual Retrieval GIFT is a visual retrieval engine based on color and texture information[3]. Colors are compared in a color histogram. Texture information is described by applying Gabor filters and quantizing the responses into 10 strengths. The image is rescaled to 256x256 and partitioned into fixed regions to extract features for both global and local levels. GIFT uses a standard *tf/idf* strategy for feature weighting. It also allows image–based queries with multiple input images.

GIFT has been used for the ImageCLEFmed tasks since 2004. Each year the default setting has been used to provide a baseline. For classification, GIFT has been used to produce the distance (similarity) value followed by a nearest neighbor (kNN) classification.

Fusion Techniques In 2009, the ImageCLEF@ICPR fusion task was organized to compare fusion techniques using the best ImageCLEFmed visual and textual results [4]. Studies such as [5] show that combSUM (1) and combMNZ(2) proposed by [6] in 1994 are robust fusion strategies. With the data from the ImageCLEF@ICPR fusion task, combMNZ performed slightly better than combSUM, the difference was small and not statistically significant.

$$S_{\texttt{combSUM}}(i) = \sum_{k=1}^{N_k} \overline{S_k(i)} \tag{1}$$

$$S_{\texttt{combMNZ}}(i) = F(i) * S_{\texttt{combSUM}}(i) \tag{2}$$

where $F(i)$ is the freqence of image i being returned by one input system with a non–zero score, and $S(i)$ is the score assigned to image i.

In ImageCLEFmed2010, the fusion approach was used in two cases:

- fusing textual and visual runs to produce mixed runs (combMNZ was used);

– fusing various images which belong to the same case (only for the case–based retrieval task, combSUM was used).

For case–based fusion, the frequency for one case is highly related to the number of images in this case. It is not optimal to include the frequency information, thus combSUM was used.

Score Normalization Techniques Studies performed by MedGIFT show that using a logarithmic function based on a rank number for score normalization was a stable solution for fusion [5]. The following formula was used:

$$\overline{S_{ln}(R)} = \ln N_{images} - \ln R, \tag{3}$$

where R is the rank given by the input system, and $\overline{S}$ is the normalized similarity score. As a large performance difference exists between textual runs and visual runs, textual runs are weighted by a factor of 0.8 whereas visual runs are weighted by 0.2. In ImageCLEFmed 2010, this score normalization strategy was applied for all fusions as well as for k–NN classification.

2.2 Image Collection

77'506 medical images were available for ImageCLEFmed 2010. Among them, 2'390 images with modality labels were used as training data, another 2'620 images are selected as test data for the modality classification task. Details about the setup and collections of the ImageCLEFmed tasks can be found in overview paper [7].

3 Results

This section describes our results for the three medical tasks.

3.1 Modality Classification

Table 1 shows the number of images for each modality in the training data. Even it was mentioned in the README file describing the data that there can be considerable intra–class heterogeneity (the PX class can contain microscopic images as well as photographs, PET can contain PET and PET/CT and XR can contain DXR and X–ray angiography), without precise labels. Manually dividing one class into sub–classes can generate errors rather than resulting in a gain.

The training data was separated by us into two parts: balanced classes and a set of remaining images. 200 images were selected randomly for each class, which created a set of 1'600 images. The remaining 790 images were used for kNN parameter tuning. Figure 1 shows the performance related to the parameter k. Performance evaluation was based on the percentage of correctly classified images. The best performance with the training data was achieved by 5NN, which was then applied on the test data.

Table 1. Number of images for each modality in the training data.

Label	Modality	Number
GX	Graphics, typically drawing and graphs	355
PX	optical imaging including photographs, micrographs, gross pathology etc	330
CT	Computerized tomography	314
US	Ultrasound including (color) Doppler	307
MR	Magnetic resonance imaging	299
XR	X-ray including X-ray angiography	296
PET	Positron emission tomography including PET/CT	285
NM	Nuclear Medicine	204
total		2390

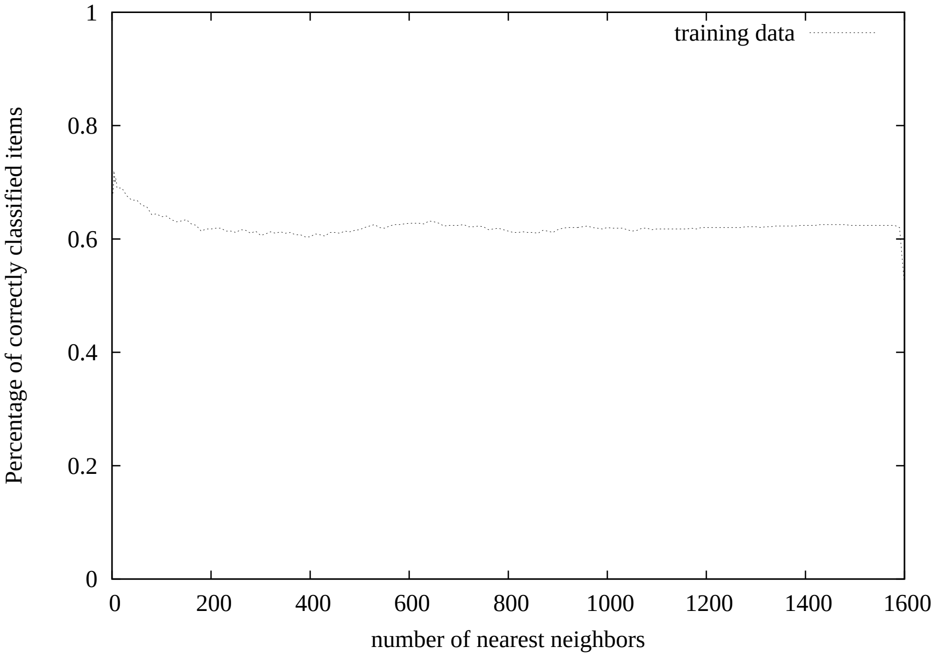

Fig. 1. The performance obtained by the kNN classification.

One run was submitted to the modality classification task. A binary classifier was used for the classification. For runs of various natures (textual, visual, mixed), the best accuracy and average accuracy are shown in Table 2. Even with default settings of GIFT and a very simple kNN classification approach, the baseline run is above the average accuracy (59%) of all visual runs. Results also show that visual runs can achieve similar performance to textual runs in accuracy, which explains the high performance of mixed runs.

Table 2. Results of the runs for the modality classification task.

run ID	best accuracy	average accuracy
mixed run	0.94	0.9
textual run	0.9	0.59
visual run	0.87	0.59
GIFT8_5NN	0.68	

3.2 Image–based Retrieval

For the image–based retrieval task, 9 groups submitted 36 textual retrieval runs, 4 groups submitted 9 visual runs and 6 groups submitted 16 mixed runs combining textual and visual information. In total 5 runs were submitted by the MedGIFT group. In addition to the three baselines (1 visual baseline and 2 textual baselines), 2 mixed runs were produced using the combMNZ approach. Results are shown in Table 3. Mean average precision (MAP), binary preference (Bpref), and early precision (P10, P30) are shown as measures. In terms of mean average precision(MAP), the best textual run (0.338) outperforms the best visual run (0.0091) by a factor of 30, which shows a big performance gap between the

Table 3. Results of the runs for the image–based topics.

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
best textual run (XRCE)	Textual	0.338	0.3828	0.5062	0.3062	667
HES–SO–VS_CAPTIONS	Textual	0.2568	0.278	0.35	0.2917	657
HES–SO–VS_FULLTEXT	Textual	0.1312	0.1684	0.1813	0.1792	658
best visual run (ITI)	Visual	0.0091	0.0179	0.0125	0.0125	66
MedGIFT_GIFT8	Visual	0.0023	0.006	0.0125	0.0042	52
best mixed run (XRCE)	Mixed	0.3572	0.3841	0.4375	0.325	762
MedGIFT_FUSION_VIS_CAPTIONS	Mixed	0.0208	0.0753	0.0375	0.0540	340
MedGIFT_FUSION_VIS_FULLTEXT	Mixed	0.0245	0.0718	0.0375	0.0479	346

two approaches. The average score of all textual runs is 0.253, whereas the average score of all visual retrieval runs is 0.0020. The performance of the baselines produced by Apache Lucene based on image caption information (HES–SO–VS_CAPTIONS) and GIFT (MedGIFT_GIFT8) are slightly above the averages. Performance of mixed runs depends largely on the fusion strategies: reordering a textual run obtains close or better performance compared with the original run, whereas merging textual runs with visual runs reduces the performance of a textual run. Two mixed runs submitted from MedGIFT are based on a merging approach and are punished by the large performance gap between textual and visual runs.

Case–Based Retrieval For the case–based retrieval task, one visual run, 43 textual runs and 4 mixed runs from 9 groups were submitted. The MedGIFT group submitted one visual run, two textual runs and two mixed runs. The visual and textual runs were obtained by processing a case–based fusion of all images of a case using the combSUM strategy. Based on visual and textual runs, mixed runs were produced by using the combMNZ strategy. Results are shown in Table 4. Best performance in terms of MAP (0.3551) was obtained by purely textual retrieval. The Lucene baseline (HES–SO–VS_FULLTEXT) is the third best run among all automatic runs. The GIFT baseline (MedGIFT_GIFT8) is the only visual run and is ranked below all textual runs. All mixed runs are ranked below the GIFT baseline, showing that the fusion strategy was not optimal.

Table 4. Results of the runs for the case–based retrieval topics.

Run	run_type	MAP	Bpref	P10	P30	num_rel_ret
best manual run (UIUCIBM)	Manual	0.3551	0.3714	0.4714	0.3857	449
best textual run (UIUCIBM)	Textual	0.2902	0.3049	0.4429	0.3524	441
HES–SO–VS_CAPTIONS	Textual	0.1273	0.1375	0.25	0.2024	342
HES–SO–VS_FULLTEXT	Textual	0.2796	0.2699	0.4214	0.3452	470
MedGIFT_GIFT8	Visual	0.0358	0.0612	0.0929	0.0786	215
best mixed run (ITI)	Mixed	0.0353	0.0509	0.0429	0.0714	316
MedGIFT_VIS_CAPTIONS	Mixed	0.0143	0.0657	0.0357	0.019	301
MedGIFT_VIS_FULLTEXT	Mixed	0.0115	0.0786	0.0357	0.0167	274

3.3 Conclusions

Comparing the ad–hoc retrieval task in ImageCLEFmed2009 and ImageCLEFmed2010, the number of relevant documents per topic decreased by 50% (94.48 in 2009 and 62.44 in 2010), which can partly explain the general decrease of performance observed. However, in 2009, the Lucune baseline using fulltext (HES–SO–VS_FULLTEXT) was ranked above the average, and the GIFT baseline (MedGIFT_GIFT8) was the best run among all purely visual runs. In 2010, both baselines are ranked slightly lower, which shows that the average performance of systems improved.

Comparison of the case–based retrieval tasks in 2009 and 2010 show a different picture. The Lucene baseline performed slightly better than the average in 2009, but is the third best run in 2010. As only 4 case–based topics were available in 2009, the results might not be representative, and the task in 2010 was definitely at a larger scale.

Both in 2009 and 2010 the performance gap between textual and visual runs is larger in image–based retrieval than in case–based retrieval. This can be explained by the fact that textual runs for both image–based and case–based topics use case information, whereas the visual approach only uses case information in case–based topics. In other words, including case information in image–based retrieval could be able to improve the performance of visual runs.

Relevance is judged based on domain knowledge, which is often case–based rather than image–based. Case–based retrieval thus seems to be more coherent. So far the visual runs for case–based topics were produced by image–based approaches plus fusion, which is not optimal. Key words extracted from fulltext articles about one case directly are a good case descriptor, whereas robust case descriptors are needed for the visual approach.

Another important aspects that went wrong in our approach is the fusion of textual and visual results that actually decreased instead of increased the results.

4 Acknowledgments

This work was partially supported by SWITCH in the context of the medLTPC project and the European Union in the context of the Khresmoi project.

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.: The ImageCLEF medical retrieval task at icpr 2010 — information fusion to combine viusal and textual information. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2010). Lecture

Notes in Computer Science (LNCS), Istanbul, Turkey, Springer (August 2010) in press.

5. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval. In: Pattern Recognition, International Conference on, Los Alamitos, CA, USA, IEEE Computer Society (2010)
6. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Text REtrieval Conference. (1993) 243–252
7. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Kahn Jr., C.E., Hersh, W.: Overview of the CLEF 2010 medical image retrieval track. In: Working Notes of CLEF 2010, Padova, Italy (September 2010)