The medGIFT Group in ImageCLEFmed 2011

Dimitrios Markonis, Ivan Eggel, Alba G. Seco de Herrera, Henning Müller

University of Applied Sciences Western Switzerland (HES–SO)
Sierre, Switzerland
dimitrios.markonis@hevs.ch

Abstract. This article presents the participation of the medGIFT group in ImageCLEFmed 2011. Since 2004, the group has participated in the medical image retrieval tasks of ImageCLEF each year. The main goal is to provide a baseline by using the same technology each year, and to search for further improvements in retrieval quality.
There are three types of tasks for ImageCLEFmed 2011: modality classification, image–based retrieval and case–based retrieval. The medGIFT group participated in all three tasks. For the image–based 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. 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 text runs outperform the best visual runs by a factor of 10 in terms of mean average precision. Baselines provided by Apache Lucene and GIFT are ranked above the average among text runs and visual runs respectively in image–based retrieval. In the case–based retrieval task the Lucene baseline is the second best automatic run for text retrieval, and our mixed and visual runs are the best overall. For modality classification, GIFT and the kNN–based approach perform slightly better than the average of the 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 [4, 5]. The medGIFT[2] research group has participated in ImageCLEFmed using the same technology as baselines since 2004. Additional modifications of the basic techniques were attempted to improve results. 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, [6]) whereas Lucene[4] was used for text retrieval.

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

This year, the bag–of–visual–words approach [1] was also used using local descriptors also called visual words. This widely used method is applied as follows: a training set of images is chosen and a number of local descriptors (in the case of SIFT, Scale Invariant Feature Transform, 128–dimensional vectors) are extracted from each image of this set. The descriptors are then clustered using a clustering method (such as k–means) and the centroids of the clusters are used as visual words. Based on this the visual vocabulary — the set of all the visual words — is created. Local features are then also extracted from each image in a database. The images are finally indexed as histograms of the visual word occurrences (bags–of–visual–words) by assigning the nearest visual word to each feature vector. When an image is queried, a similarity measure is used to compare the query image histogram and the database images histograms, providing a similarity score. In order to include spatial information to this representation, several approaches have been proposed [2, 3], improving the performance.

2 Methods

This section describes the basic techniques that we used for retrieval in ImageCLEFmed 2011.

2.1 Retrieval Tools Reused

This section details the existing retrieval tools that were reused for text and visual retrieval.

Text Retrieval The text retrieval approach in 2011 is based on Lucene using standard settings. 4 text runs were submitted, 2 for case–based retrieval and 2 for image–based retrieval. For case– and image–based retrieval, captions and full text were used.

The full text approach used all texts as obtained in the data set. 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 [6]. Colors are compared in a color histogram using a simple histogram intersection. Texture information is described by applying Gabor filters and quantizing the responses into 5 strengths. This different from the previous years' use of 10 strengths because of the size of this year's data set that can cause problems for GIFT. The image is rescaled to 256x256 and partitioned into fixed regions to extract features both global and local features. GIFT uses a standard *tf/idf* (term frequency/inverse document frequency) 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 (1NN) classification.

For the description of the images when using visual words, we used the SIFT implementation in the fiji[5] image processing package. In order to create the visual vocabulary, our implementation of the density–based clustering algorithm DENCLUE [7] was used. The reason for this choice are the features and the nature of the dataset that needs to be clustered. The data set to be clustered is large–scale (1000 training images produce approximately 2'500'000 descriptors) and high dimensional (SIFT descriptors are 128–dimensional). The DENCLUE algorithm is highly efficient for clustering large–scale datasets, can detect arbitrarily shaped clusters and handles outliners and noise well. Moreover, opposed to other density–based clustering algorithms it performs well for high–dimensional data. However, when using a density–based clustering algorithm care has to be taken for data sets containing clusters of different densities. To deal with this, the parameter ξ that controls the significance of the candidate cluster in respect to its density was set to zero.

In order to create a pipeline for easy component–based evaluation for this method the outputs of every intermediate step were stored in CSV files and mySQL tables. These use a large amount of storage resources but speed up the procedure of tuning and evaluating components of the method once the ground truth is available. Due to the characteristics of this architecture it was possible to use only vocabularies with a small number of visual words ($\tilde{1}00$) and a $n \times n$ partition was used with maximum $n = 2$.

The third submitted approach combines the modality classification and the image retrieval tasks. Using the GIFT assignment of modalities the histograms of visual words where indexed in mySQL tables based on their classes. In this indirect manner, the approaches of GIFT and bag–of–words were combined as well. The visual word histogram of the query image was first compared to the indexed histograms of the training set using a histogram intersection. The classes of the 5 nearest neighbors were acquired. Then, the same histogram was compared again but only with the indexed histograms in the tables of these classes. The 1000 nearest images were acquired as the results for the topic images. For topics that contained more than one query image the combSUM technique was used as is explained in the next section.

Fusion Techniques In 2009, the ImageCLEF@ICPR fusion task was organized to compare fusion techniques using the best ImageCLEFmed visual and textual results [8]. Studies such as [9] show that combSUM (1) and combMNZ(2) proposed by [10] 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. In general, rank–based

[5] `http://fiji.sc/wiki/index.php/Fiji`

fusion worked better than score–based fusion.

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

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

where $F(i)$ is the frequency 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 ImageCLEFmed2011, the fusion approach using scored–based combSUM was used in three cases:

- fusing textual and visual runs to produce mixed runs;
- fusing results from various images which belong to the same topic for the bag–of–visual–word approaches, (GIFT handles queries with several images automatically);
- fusing GIFT and bag–of–visual–word approaches.

2.2 Image Collection

230'089 medical images were available for ImageCLEFmed 2011. Among them 1'000 images with modality labels were used as training data and another 1'000 images were selected as test data for the modality classification. Details about the setup and collections of the ImageCLEFmed tasks can be found in the overview paper [11].

3 Results

This section describes our results for the three medical tasks.

3.1 Modality Classification

One run was submitted to the modality classification task using GIFT. For runs of various natures (textual, visual, mixed) the best accuracy and average accuracy are shown in Table 1. It can be observed that GIFT, using 1NN classification

Table 1. Results of the runs for the modality classification task.

run	best accuracy	average accuracy
mixed runs	0.8691	0.7188
textual runs	0.7041	0.5903
visual runs	0.8359	0.6878
GIFT_1NN	0.6220	

performed worse than the average accuracy. This was expected, as neither k for

the kNN was optimal, nor the optimal GIFT feature configuration was used, due to the dataset size. Results also show that visual runs achieve performance close to the mixed runs showing the importance of visual characteristics in modality classification. The analysis of text results are not absolutely reliable though, as only two exclusively textual runs were submitted.

3.2 Image–based Retrieval

In total 8 runs were submitted to the image–based retrieval task by the medGIFT group. Using the GIFT baseline and the 2 textual baselines, 2 mixed runs were produced using the combSUM approach. One run was the fusion of GIFT and the 2–step approach described in Section 2.1. Results are shown in Table 2. Mean average precision (MAP), binary preference (Bpref), and early precision (P10, P30) are shown as measures. For the full text retrieval, the score of a text was extended to all images of this text, for the caption–based retrieval it was extended to all images of this caption. In terms of mean average precision

Table 2. Results of the medGIFT runs and the best runs for the image–based topics.

run	run_type	MAP	P10	P30	Rprec	Bpref	num_rel_ret
best mixed run	Manual	0.2372	0.3933	0.3550	0.2881	0.2738	1597
mixed_captions_ib	Automatic	0.1176	0.2800	0.2100	0.1575	0.1614	705
mixed_full_ib	Automatic	0.0857	0.2900	0.2700	0.1300	0.1308	830
best visual run	Automatic	0.0338	0.1500	0.1317	0.0625	0.0717	717
gift_visual_ib	Automatic	0.0274	0.1467	0.1367	0.0581	0.0807	731
visual_ib	Automatic	0.0252	0.1267	0.1200	0.0554	0.0752	709
bovw_visual_ib	Automatic	0.0126	0.0867	0.0800	0.0315	0.0437	324
bovw_s2_visual_ib	Automatic	0.0076	0.0900	0.0650	0.0182	0.0279	213
best textual run	Automatic	0.2172	0.3467	0.3017	0.2369	0.2402	1471
image-based_captions	Automatic	0.1742	0.3000	0.2683	0.2096	0.2179	1261
image-based_fulltext	Automatic	0.0921	0.2167	0.2150	0.1264	0.1506	1211

(MAP), the best textual run (0.2172) outperforms the best visual run (0.0338) by a factor of 7, which shows a big performance gap between the two approaches. However, it is significantly smaller than the gap in ImageCLEF 2010.

The average score of all textual runs is 0.1644, whereas the average score of all visual retrieval runs is 0.0146. The performance of the baseline produced by Apache Lucene based on image caption information (HES–SO–VS_CAPTIONS) is slightly above the average. On the other hand, GIFT performed surprisingly well, considering the non–optimal configuration and age of the tool. The bag–of–visual–words approaches did not demonstrate good results, most likely due to the lack of parameter tuning and lack of using training data. For the 2–step approach the initial results were not as good as the initial results, but with parameter tuning, already better results could be obtained. However, using the

component–based architecture that was developed, further research will be easier to perform.

Merging of textual runs with visual runs reduces the performance of the textual runs, which is again due to the non–optimal technique using scores and not ranks. The two mixed runs submitted by the medGIFT group are based on a simple merging approach and are punished by the large performance gap between textual and visual runs.

Case–Based Retrieval The medGIFT group submitted four visual runs, one textual run and one mixed run for the case–based retrieval task. The visual runs were obtained by processing a case–based fusion of the results of querying all images of a case using the combSUM strategy. Text runs were performed using the full text and for the caption–based retrieval the results of all captions of a text were combined using combSUM. Based on visual and textual runs, mixed runs were produced by using the combSUM strategy. Table 3 shows the MedGIFT runs and if our run was not the best also the performance of the best run.

Table 3. Results of the medGIFT runs and the best runs for the case–based topics.

run	run_type	MAP	P10	P20	Rprec	Bpref	num_rel_ret
mixed_GIFTLucene_full	Automatic	0.0754	0.1667	0.1556	0.1227	0.0958	121
best textual run	Automatic	0.1297	0.1889	0.1500	0.1588	0.1212	144
case_based_fulltext	Automatic	0.1293	0.2000	0.1444	0.1509	0.1122	141
case_based_captions	Automatic	0.0437	0.1111	0.0833	0.0816	0.0540	90
gift_visual	Automatic	0.0204	0.0444	0.0333	0.0336	0.0292	45

Best performance in terms of MAP (0.1297) was obtained by purely textual retrieval. The Lucene baseline (fulltext) is the second best run (0.1293) among all automatic runs and the difference to the best runs is statistically not significant. MedGIFT was the only lab that submitted purely visual runs and even though the best result (0.0204 by GIFT) is lower than the best textual run, the difference is not as bad as for the image–based task. The mixed run of GIFT and the Lucene fulltext achieved the best results (0.0754) in the mixed runs. This run also has the best P5 so very early precision of all all runs but on the other hand its P10 is already lower than the best textual run for P10, which is also a run of the medGIFT group. This combination decreased the performance of the corresponding textual run for MAP, so there is still a potential in improving the current systems by not using a direct fusion but rather a reordering of the results.

4 Conclusions

Based on the results of the medGIFT participation several lessons can be learned, often similar or at least in line with previous years. The baseline run of Lucene using captions performed better in the image–based task while the fulltext–based approach showed good results in case–based retrieval task. In general, the baseline of GIFT performs well in image–based and case–based retrieval although on a lower level than the text retrieval approaches. The same cannot be said for the modality classification but this was probably due to the poor classification rule that was extremely simple without any use of training data. For visual classification several very good and optimized systems exist that reach a much better performance. As the datasets grow larger, aspects of system scalability such as the trade–off of memory usage, speed and quality have to be taken into account for future content–based image retrieval systems.

Concerning the bag–of–visual–words runs, further testing and work is required to fully exploit the advantages of the methods used. While larger vocabularies and finer partitions may improve the result, a better classifier can enhance the 2–step approach, which already delivered better results than the approach presented in this text.

Finally, we can see that the majority of the mixed runs decreased the performance compared to the textual runs when combined. This indicates that special care needs to be taken for the fusion of unbalanced runs in terms of performance, as the textual and visual runs are obtaining very different performance. In the based, rank–based measure have shown to be better than score–based approaches and this was mistakenly not taken into account.

5 Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme under grant agreement 257528 (KHRESMOI), 249008 (Chorus+) and 258191 (Promise).

References

1. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. ICCV '03, Washington, DC, USA, IEEE Computer Society (2003) 1470–1477
2. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. CVPR '06, Washington, DC, USA, IEEE Computer Society (2006) 2169–2178
3. Philbin, J., Chum, O., Isard, M.: Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA (June 2007) 1–8

4. 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
5. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Said, R., Bakke, B., Kahn Jr., C.E., Hersh, W.: Overview of the CLEF 2010 medical image retrieval track. In: Working Notes of CLEF 2010 (Cross Language Evaluation Forum). (September 2010)
6. 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.
7. Hinneburg, A., Keim, D.A.: An efficient approach to clustering in large multimedia databases with noise. In: Conference on Knowledge Discovery and Data Mining (KDD). Volume 5865., AAAI Press (1998) 58–65
8. 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.
9. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval. In: International Conference on Pattern Recognition, ICPR'10, Los Alamitos, CA, USA, IEEE Computer Society (2010)
10. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Text REtrieval Conference. (1993) 243–252
11. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (September 2011)