NovaSearch on medical ImageCLEF 2013

André Mourão, Flávio Martins and João Magalhães

Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia,
Caparica, Portugal,
a.mourao@campus.fct.unl.pt, flaviomartins@acm.org, jm.magalhaes@fct.unl.pt

Abstract. This article presents the participation of the Center of Informatics and Information Technology group CITI in medical ImageCLEF 2013. This is our first participation and we submitted runs on the modality classification task, the ad-hoc image retrieval task and case retrieval task. We are developing a system to integrate textual and visual retrieval into a framework for multimodal retrieval. Our approach for multimodal case retrieval achieved best global performance using Segmented (6×6 grid) Local Binary Pattern (LBP) histograms and Segmented HSV histograms for images and Lucene with query expansion (using the first top 3 results). In modality classification we achieved one of the largest MAP gains in the multimodal classification task, resulting in the third best team result.

Keywords: medical retrieval, case-based retrieval, multimodal fusion, medical modality classification

1 Introduction

ImageCLEF 2013 AMIA: Medical task is the CLEF benchmark focused on the retrieval and classification of medical images and articles from PubMed. This is our first participation on ImageCLEF and the Medical track in particular, so we tried to design a system to participate on every task (excluding Compound figure separation). However we delved more in the case-based retrieval task and results lists fusion.

Our system is divided in image retrieval, textual retrieval and results fusion. For image retrieval, we focused on top performing features from previous editions [10] (CEDD [3], FCTH [4], Local Binary Pattern histograms [1], color histograms).

For text retrieval, we used Lucene[1] from the Apache project paired with our implementation of the BM25L [8] ranking function for Lucene. For modality classification, we combined text and image descriptors (early fusion) and performed classification using Vowpal Wabbit[2]. Our training dataset included only the images provided on the 2013 edition [7]. We did not perform any type of

[1] http://lucene.apache.org/core/

[2] https://github.com/JohnLangford/vowpal_wabbit/wiki

dataset augmentation. On the case-based retrieval task, we also tested a variety of late fusion approaches, and came up with a variant of Reciprocal Rank (RR) [11] we named Inverse Square Rank (ISR). Our fusion method provided the best performance for our case-based image and textual runs.

2 Techniques

2.1 Text retrieval

In text retrieval, article text is indexed using Lucene and retrieved using the BM25L [8] retrieval function. We expanded the initial query with preferred and alternative terms sourced from a SKOS formatted version of MeSH using Lucene-SKOS. This improves precision metrics at the top ranks, so we then exploit these top documents to perform pseudo-relevance feedback.

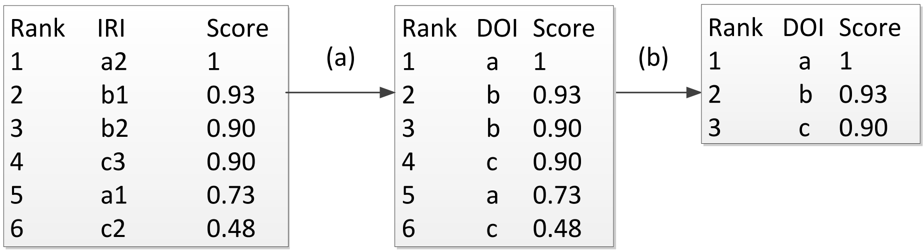
The indexed fields depend on the task: For image retrieval, we achieved good results indexing and searching only on the image captions, title and abstract. For case retrieval, we indexed and searched on the full text (all chapters including image captions), abstract and title. We ran pseudo-relevance feedback using the top 3 results retrieved using the initial query. We added a maximum of 25 new query terms to the initial query. We purged any candidate words that did not appear in a minimum of 5 documents.

2.2 Visual retrieval

For visual retrieval, we extracted a set of descriptors effective in medical images retrieval (CEDD, FCTH, Local Binary Pattern (LBP) histograms and color histograms). We used this image features in pairs. For LPB and color histograms, we extracted both descriptors for 6×6 image grid and concatenated the results. For CEDD and FCTH, we extracted and concatenated both feature descriptors.

CEDD (Color and Edge Directivity Descriptor) expresses color and edge directivity information from the image into a compact descriptor (54 bytes per image). It combines "real color" histograms from the HSV color space (e.g. bins Light Magenta, Dark Green) with MPEG-7 Edge Histogram Descriptor. FCTH (Fuzzy Color and Texture Histogram) combines the same color histograms as CEDD with texture information extracted with multiple energy wavelets using fuzzy logic techniques. Color histograms are histograms of the individual components of the HSV color space. Local Binary Patters are texture descriptors based on thresholding a pixel with its neighbors to detect texture variations. The thresholded values are concatenated into an histogram to represent image texture. The features of all images in the corpus are stored in a FLANN [9] L_2 index.

The image retrieval results are sorted by their similarity, with the score being the normalized inverse of the L_2 distances between the query image and the indexed result images. For case-base retrieval, an additional step must be performed: the image id (IRI), must be converted into a document id (DOI) (Figure 1 (a)) and the duplicate results must be merged to have an unique document list (Figure 1 (b)). More details are present in section 2.3.



Rank	IRI	Score
1	a2	1
2	b1	0.93
3	b2	0.90
4	c3	0.90
5	a1	0.73
6	c2	0.48

Rank	DOI	Score
1	a	1
2	b	0.93
3	b	0.90
4	c	0.90
5	a	0.73
6	c	0.48

Rank	DOI	Score
1	a	1
2	b	0.93
3	c	0.90

Fig. 1. Case based retrieval step. (a) get document id (DOI) from image id (IRI); (b) combine multiple document results into one (unique) document list. The example uses CombMAX fusion for simple visualization.

2.3 Results fusion

Result fusion aims at combining ranked lists from multiple sources into a single combined ranked list. Consider these two use cases: combine the results from queries with multiple images and combine the results from text and images queries.

There are two main approaches for late fusion: score based and rank based. Score based approaches (CombSUM, CombMAX and CombMNZ) combine the normalized scores given by the individual searches (e.g. visual and textual) as a basis to the create the new ranked list. The studied variant that achieves the best performance [2] is CombMNZ, but ranked based fusion is gaining momentum, and can outperform score based fusion under most conditions [5, 6]. For each document i, the score after fusion can be computed as:

$$\text{combSUM}(i) = \sum_{k=1}^{N(i)} S_k(i), \tag{1}$$

$$\text{combMAX}(i) = \max(S), \forall S \subset D_i, \tag{2}$$

$$\text{combMNZ}(i) = N(i) \times \text{combSUM}(i), \tag{3}$$

where $S_k(i)$ is the score of the i document on the k result list.

N(i) refers to the number of times a document appears on a results list. A result list k does not contain all documents. Documents with a zero score or a very high rank can be safely ignored. Thus, *N(i)* varies between 0 (the document i does not appear on any list) and the total number of results list (the document i appears on all lists). For example, in our experiments, there are two results lists: one for visual search and other for textual search, limited to 1000 results each.

Rank based fusion methods consider the inverse of the rank of each document in each one of the individual lists as the score. Reciprocal Rank and Reciprocal Rank Fusion are the two methods we evaluated:

$$\text{RR}(i) = \sum_{k=1}^{N(i)} \frac{1}{R_k(i)}, \tag{4}$$

$$\mathrm{RRF}(i) = \sum_{k=1}^{N(i)} \frac{1}{h + R_k(i)}, \text{with}\, h = 60. \tag{5}$$

where $R_k\,(i)$ is the rank of document i on the k rank.

After analyzing both score and rank based approaches, we combined elements from both to improve precision. Inverted Squared Rank (ISR) combines the inverse rank approaches of RR and RRF (using the squared rank to improve precision at top results) with the frequency component of combMNZ (results that appear on multiple lists are boosted):

$$\mathrm{ISR}(i) = N(i) \times \sum_{k=1}^{N(i)} \frac{1}{R_k(i)^2}. \tag{6}$$

3 Results

3.1 Modality classification

In the modality classification task, we used the CEDD and FCTH descriptors and (stemmed) text from the corresponding caption and article title. In the image and textual runs, we used the corresponding descriptors from all images in the provided training dataset to create a model based on stochastic gradient descent with Vowpal Wabbit. These runs correspond to the *CEDD_FCTH* for images and *words* from text in Table 1.

The multimodal run concatenates the descriptors described above and performs the stochastic gradient descent with Vowpal Wabbit with the combined descriptors (CEDD, FCTH, caption and title words). This run correspond to the *All* in Table 1.

The dataset provided as training data for modality classification was unbalanced in the quantity of images per class. For example, the "Compound or multipane images" (COMP) category contains over 1,105 training images while the "PET" (DRPE) category contains 16 training images . Also, images in the COMP may not be visually distinct; they consist on the combination of images from other categories. Thus, we have decided to send runs where we ignored the COMP in training and classification. These run correspond to the runs ended by *noComb* in Table 1.

Regarding the performance of our algorithm, we were able to be the 3$^{\mathrm{rd}}$ best team overall with our multimodal *All* run, classifying 72.92% of the images correctly. This is only 8.76% bellow the best run. Our visual run classified 57.62% of the images correctly (23.17% behind the best run) and our best textual run classified 62.35% of the images correctly (1.82% behind the best run). The most interesting fact is the big improvement of the results from single modality runs in the multimodal run. We improved our best single modality run by over 10%, while the best team only improved their best run by 0.89% in the multimodal approach. Our *noComb* approaches did not perform well, as the testing dataset contained a lot of COMP images.

Table 1. Modality classification performance comparison by retrieval type. All our runs and the best runs are present. If our run is the best, the second best is present

Run Name	Type	Correctly classified
IBM_modality_run1	T	**64.17%**
words	T	62.35%
words_noComb	T	32.80%
IBM_modality_run4	V	**80.79%**
CEDD_FCTH	V	57.62%
CEDD_FCTH_NoComb	V	32.49%
IBM_modality_run8	M	**81.68%**
All	M	72.92%
All_NoComb	M	44.61%

3.2 Case retrieval

For case retrieval, we submitted one visual run (with segmented LPB Histograms and HSV Color Histograms) and two textual runs (one with MeSH expansion and one without expansion). The run with MeSH expansion (with *_MSH_* on the run name) outperformed the run without expansion in all metrics, with a special emphasis on MAP (12% increase) and P@10 (11% increase).

We achieved our best result in the expanded textual run, with a MAP 22%, GM-MAP 11.8% and P@10 26%, very close to the best textual run (MAP: 24%, GM-MAP: 11.6% and P@10: 27%). The visual run achieved the best result in class, outperforming the second best result by a factor of 10. Our run achieved a MAP of 3% and P@10 of 4%. Our best multimodal run (using ISR) also achieved the best result in class with a MAP of 16% and P@10 of 18%. The combMNZ based run achieved much worse results (MAP: 8% and P@10: 14%), following our idea that rank-based fusion is better that score-based fusion in multimodal retrieval. Although these results are worse that the textual only results, our rank-based fusion algorithms improved existing algorithm by a small margin.

Table 2. Case based runs performance comparison by retrieval type. All our runs and the best runs are present. If ours is the best in the modality, the second best is present

Run Name	Type	MAP	GM-MAP	bpref	P@10	P@30
SNUMedinfo9	T	**0.2429**	0.1163	**0.2417**	**0.2657**	**0.1981**
FCT_LUCENE_BM25L_MSH_PRF	T	0.2233	**0.1177**	0.2000	0.2600	0.1800
FCT_LUCENE_BM25L_PRF	T	0.1992	0.0964	0.1874	0.2343	0.1781
FCT_SEGHIST_6x6_LBP	V	**0.0281**	**0.0009**	**0.0300**	**0.0429**	**0.0238**
medgift_visual_nofilter_casebased	V	0.0029	0.0001	0.0035	0.0086	0.0067
FCT_CB_MM_rComb	M	**0.1608**	0.0779	**0.1400**	0.1800	0.1257
medgift_mixed_nofilter_casebased	M	0.1467	**0.0883**	0.1318	**0.1971**	**0.1457**
FCT_CB_MM_MNZ	M	0.0794	0.0035	0.0800	0.1371	0.0810

Fusion In addition to the submitted runs, we compared the performance of the fusion algorithms using the best textual and visual runs for case-based retrieval. Performance was evaluated using trec_eval and the provided relevance judgments (Table 3).

With our data, rank-based approaches outperformed score based approaches by a factor of 2. One of the reasons is the scoring differencs between text and images. Even though both visual and text scores have the same normalization (the [0...1] interval), the distribution of the results in the score space is different. Rank based approaches can handle multi-modality better, because the scores are ignored.

Regarding the differences between RR, RRF and ISR: ISR performed as well or better in our experiments (Table 3 and 4) in most measures, with a significant performance boost on P@10 for case-based retrieval. The polynomial component promotes top ranking results to the top of the list, offering a better precision at top results (e.g. P@10).

Table 3. Fusion comparison for the medical ImageCLEF case based queries

Run Name	Comb	MAP	GM-MAP	bpref	P@10	P@30
FCT_CB_MM_rComb	**ISR**	**0.1608**	0.0779	**0.14**	**0.1800**	**0.1257**
(not submitted)	RRF	0.1597	**0.0787**	0.13	0.1571	0.1248
(not submitted)	RR	0.1582	0.0779	**0.14**	0.1771	0.1238
(not submitted)	combSUM	0.0804	0.0039	0.09	0.1429	0.0790
FCT_CB_MM_MNZ	combMNZ	0.0794	0.0035	0.08	0.1371	0.0810
(not submitted)	combMAX	0.0292	0.0013	0.03	0.0457	0.0248

Table 4. Ad-hoc image retrieval fusion performance.

Run Name	Comb	MAP	GM-MAP	bpref	P@10	P@30
nlm-se-image-based-mixed (best global result)		0.3196	0.1018	0.2983	0.3886	0.2686
(not submitted)	RRF	**0.1496**	**0.0331**	**0.1549**	**0.2200**	**0.1495**
(not submitted)	ISR	0.1457	0.0329	0.1504	0.2057	0.1476
(not submitted)	RR	0.1448	0.0326	0.1492	0.2086	0.1457
(not submitted)	combMMZ	0.0488	0.0013	0.0735	0.1457	0.0752

3.3 Image retrieval

In ad-hoc image retrieval, we submitted one visual run (with segmented LPB Histograms and HSV Color Histograms) and two textual runs (one with MeSH expansion and one without expansion). As with case retrieval, the run with MeSH expansion (FCT_SOLR_BM25L_MSH) outperformed the run without expansion

(FCT_SOLR_BM25L) in all metrics, although to a lesser degree: MAP increased 5% and P@10 increased 10%.

We achieved our best result in the expanded textual run, with a MAP of 23% and P@10 of 30%, well behind the best textual (and best overall) run with a MAP of 32% and P@10 of 39%. As expected, our visual runs performed worst with a MAP of 0.7% and P@10 of 3%, about half of the performance of the best visual run (MAP: 1.9% and P@10: 6%).

After the competition, we tested the same fusion algorithms as with case-based retrieval for multimodal retrieval. Our conclusion are similar to the ones discussed in the case-based retrieval section: rank based approaches are much better for our data, with ISR and RR being the best for fusion.

Table 5. Ad-hoc image based runs performance comparison by retrieval type. All our runs and the best runs are present.

Run Name	Type	MAP	GM-MAP	bpref	P@10	P@30
nlm-se-image-based-textual	T	**0.3196**	**0.1018**	**0.2982**	**0.3886**	**0.2686**
FCT_SOLR_BM25L_MSH	T	0.2305	0.0482	0.2316	0.2971	0.2181
FCT_SOLR_BM25L	T	0.22	0.0476	0.228	0.2657	0.2114
DEMIR4	V	**0.0185**	**0.0005**	**0.0361**	**0.0629**	**0.0581**
FCT_SEGHIST_6x6_LBP	V	0.0072	0.0001	0.0151	0.0343	0.0267

4 Conclusions

We would like to emphasize our results in the visual and multimodal case retrieval tracks, where we achieved the best MAP and bpref. Our fusion algorithm (ISR) achieved slightly better performance than existing fusion algorithms and helped us achieving the best result on the multimodal case retrieval track.

Our results in the modality classification are also noteworthy. In the multimodal run, we increased the best single modality result by 17%, ending with the third best team result.

We were not able to submit all the desired combination of features and fusion algorithms due to limited time. We hope that next year we can submit all the desired runs. We will also focus on improving multimodal fusion using other fusion approaches (e.g early fusion in case and image based retrieval) and algorithms. Other technique we will study is the integration of modality as a feature in the retrieval tasks.

Overall, we are satisfied with our first participation in ImageCLEF and hope that we improve our performance in the following editions.

References

1. T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–41, Dec. 2006.
2. N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, 31(3):431–448, May 1995.
3. S. A. Chatzichristofis and Y. S. Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems*, ICVS'08, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.
4. S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis, and N. Papamarkos. Accurate Image Retrieval Based on Compact Composite Descriptors and Relevance Feedback Information. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 24(2):207 – 244, 2010.
5. G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR '09*, pages 758–759, New York, NY, USA, 2009. ACM.
6. D. Frank Hsu and I. Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retr.*, 8(3):449–480, May 2005.
7. A. Garcia Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, and H. Müller. Overview of the imageclef 2013 medical tasks. In *Working notes of CLEF 2013*, 2013.
8. Y. Lv and C. Zhai. When documents are very long, bm25 fails! In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1103–1104. ACM, 2011.
9. M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, volume 340, pages 331–340. INSTICC Press, 2009.
10. H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF 2012 working notes*, 2012.
11. M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, and L. Zhao. Expansion-based technologies in finding relevant and new information: Thu trec2002 novelty track experiments. In *the Proceedings of the Eleventh Text Retrieval Conference (TREC*, pages 586–590, 2002.