# Comparing Fusion Techniques for the ImageCLEF 2013 Medical Case Retrieval Task

Alba G. Seco de Herrera, Roger Schaer, Dimitrios Markonis,
Henning Müller

*University of Applied Sciences Western Switzerland (HES–SO)
Sierre, Switzerland*

## Abstract

Retrieval systems can supply similar cases with a proven diagnosis to a new example case under observation to help clinicians during their work. The ImageCLEFmed evaluation campaign proposes a framework where research groups can compare case–based retrieval approaches.

This paper focuses on the case–based task and adds results of the compound figure separation and modality classification tasks. Several fusion approaches are compared to identify the approaches best adapted to the heterogeneous data of the task. Fusion of visual and textual features is analyzed, demonstrating that the selection of the fusion strategy can improve the best performance on the case–based retrieval task.

*Keywords:* Medical Case–based retrieval, Multimodal Fusion, Visual Reranking, ImageCLEF, medGIFT

## 1. Introduction

Hospitals and medical institutions generate thousands of imaging studies per day [1], which lead in the Geneva University hospitals to a production

*Email address:* `{alba.garcia, roger.schaer, dimitrios.markonis,henning.mueller}@hevs.ch` (Alba G. Seco de Herrera, Roger Schaer, Dimitrios Markonis,
Henning Müller)

*URL:* `http://medgift.hevs.ch/` (Alba G. Seco de Herrera, Roger Schaer, Dimitrios Markonis,
Henning Müller)

of over 300,000 images per day in 2013. Scientific articles carry much of the medical knowledge both in their textual content and the large number of images they contain. Articles can be very valuable for the clinical routine, research and education [2] where up–to–date medical knowledge is needed. However, it is not always easy to find the desired information in this large amount of data  [3] and in clinical routine the time to fulfill an information need is often very limited. As a consequence, there is a requirement to manage and retrieve these documents in the most efficient and effective way [4]. Information access and retrieval systems are a useful tool to respond to information needs of medical professionals and to provide access to the biomedical literature related to these information needs. Clinicians regularly use information retrieval systems and benefits to decision making and patient care are reported in the literature [5].

ImageCLEF[1] [6] is a benchmark on cross–language image annotation and retrieval in various domains. Since 2004, the medical task of ImageCLEF (ImageCLEFmed) aims at evaluating the performance of medical image retrieval systems [7, 8]. Since 2009, a case–based retrieval task has been running as part of ImageCLEFmed. In the case–based retrieval task, the retrieval unit is a medical article describing a case. The goal of this task is to evaluate systems which, given a case including images and a textual description (anamnesis), retrieve articles describing cases that are useful for a differential diagnosis or match the exact diagnosis of the query [7].

The combination of various single search modalities (such as text and visual image features) makes it possible to use cross–modal relationships and thus improve the performance beyond the performance of single components [9]. However, the improvement of the performance of these multimodal systems has long been considered difficult due to the richness of multimedia [10] and the complexity to extract meaningful information from visual documents in a large domain automatically. Fusing the retrieval results of visual and textual resources into a final ranking is a popular approach for multimodal retrieval.

Several fusion models are applied in the literature to combine multimodal sources. Pham et al. [11] combine text and visual features by normalizing and concatenating them to generate the feature vectors. Then, Latent

[1] `http://www.imageclef.org/` the image–based retrieval task of the Cross Language Evaluation Forum

Semantic Analysis (LSA) is applied on these features for image retrieval. Cao et al. [12] represents the features from different modalities as a multi–dimensional matrix and incorporate these feature vectors using an extended LSA model. Gkoufas et al. [13] increase the retrieval performance by applying linear methods to combine visual and textual sources of images. Mourão et al. [14] introduce a new fusion technique, Inverted Squared Rank (ISR), a variant of the Reciprocal Rank Fusion (RRF).

Despite the existence of many fusion strategies for multimodal information [15], ImageCLEFmed has shown that most multimodal retrieval systems obtained low results in the case–based retrieval task, often below the performance of purely textual runs. The best Mean Average Precision (MAP) achieved by a combined visual/textual run in 2013 was 0.1608 (lower than the best textual run), while in the ad–hoc image retrieval task multimodal retrieval had a much higher performance with a MAP of 0.3196 [6] and this was better than the purely textual runs.

To improve retrieval quality, a successful classification of images into types (e.g. X–ray, ultrasound, computer tomography, etc) can be applied to filter out irrelevant images [2]. Already many web–accessible search systems such as OpenI[2] [16], Goldminer[3] or Yottalook[4] allow users to limit the search results to a particular modality [8] as this is a feature often requested by end users [17]. ImageCLEFmed has been running an image modality classification task since 2010. A class hierarchy was proposed including diagnostic images, generic biomedical illustrations and compound or multi–plane images with several sub categories [7].

Another necessary step is to automatically separate compound or multi–plane figures (figures consisting of several sub figures) in the biomedical literature. A very large portion of the images found in the biomedical literature are in fact compound figures. For the used PubMed Central database this concerns approximately 40% of all images. When data of articles is made available digitally, often the sub figures are not separated but made available in a single block. Information retrieval systems for images should be capable of distinguishing the parts of compound figures that are relevant to a given query. Compound figure separation is therefore a required first step

[2] http://openi.nlm.nih.gov/

[3] http://goldminer.arrs.org/

[4] http://www.yottalook.com/

to retrieving focused figures and as a consequence also cases from the literature. Therefore, in ImageCLEFmed 2013 a specific track on compound figure separation was added [7].

In this paper, the medGIFT group[5] presents extended experiments on the case–based retrieval task to improve results obtained in ImageCLEFmed 2013 [18]. To improve the precision of the system, an analysis of past experiments is performed as a first step. The article focuses on the investigation of standard fusion techniques of visual descriptors and multimodal approaches. Multimodal results overcome the best results achieved in the ImageCLEFmed case–based retrieval task when choosing the right fusion techniques. The multimodal approach achieves a MAP of 0.1795. Successful image modality classification (69.63%) and compound figure separation (84.64%) tools are also presented in this paper for their future integration into the case–based retrieval system.

The paper is organized as follows. The database used to evaluate the proposed methods is described in Section 2. Section 3 outlines the medGIFT participation in ImageCLEFmed 2013. Section 4 presents the techniques applied in this study. The obtained results are presented in Section 5 and the article concludes with Section 6.

## 2. Image database used

The data and evaluation scenario used in this text is reused from the ImageCLEFmed 2013 benchmark [7]. The 2013 collection consists of over 300,000 images of 75,000 articles of the biomedical open access literature. This corpus is a subset of PubMed Central[6] containing in total over 1.5 million images and being updated with new data regularly. The distributed PubMed subset contains only articles allowing redistribution.

Each of the query topics contains a case description with patient anamnesis, limited symptoms and test results including imaging studies but not the final diagnosis. Table 1 shows the number of of relevant documents in the database for each of the topics, in total there are only 709 relevant documents for the 35 queries, which complicates the task.

In the following experiments, the 1000 best–ranked articles are retrieved for each query topic. Results are averaged over the 35 queries for the case–

[5] `http://medgift.hevs.ch/`

[6] `http://www.ncbi.nlm.nih.gov/pmc/`

Table 1: Number of relevant articles per topic in the case–based ImageCLEFmed 2013 task.

| **Topic number** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **N. of relevant articles** | 21 | 3 | 3 | 4 | 34 | 54 | 33 | 40 | 3 | 1 |
| **Topic number** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| **N.of relevant articles** | 1 | 3 | 24 | 58 | 5 | 2 | 1 | 10 | 17 | 32 |
| **Topic number** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** |
| **N.of relevant articles** | 32 | 53 | 38 | 11 | 3 | 101 | 8 | 7 | 15 | 41 |
| **Topic number** | **31** | **32** | **33** | **34** | **35** | **Total** | | | | |
| **N.of relevant articles** | 2 | 26 | 4 | 9 | 10 | 709 | | | | |

based retrieval task in order to reproduce the exact setup of ImageCLEF.

## 3. Analysis of the ImageCLEFmed 2013 submission of the medGIFT group

In this section an analysis of the medGIFT runs in the ImageCLEFmed 2013 case–based task is provided. The analysis was used to better understand the medGIFT system to improve its results.

In 2013, the medGIFT group submitted three runs to the case–based retrieval task: one textual, one visual and one multimodal run [18]. The Lucene[7] information retrieval library was used to establish the text retrieval baseline. Provided below are some details about the way Lucene was used and configured:

- EnglishAnalyzer – In Lucene, an analyzer is used for tokenization (splitting text into parts), stemming (keeping only the root of a word) and stop word removal (excluding common words from the index). The EnglishAnalyzer that was used filters out a list of common English stop words (and, or, is, ...) and performs stemming based on rules specific to the English language (removing the letter ”s” at the end of words, removing common endings like ”-ing”, ”-er”, etc.).

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

- Multiple boolean operators – When parsing a text query, Lucene uses a boolean operator for terms separated by a space character (AND, OR). In order to maximize the score of relevant documents, each text query was executed three times : using the OR operator, using the AND operator and finally putting the query into quotes ("...") to perform an exact phrase search. The three result lists were then fused using a reciprocal rank fusion rule [19], boosting this way the ranking of exact matches.

- Term frequency–inverse document frequency (TF–IDF) similarity – several similarity measures are implemented in Lucene. The commonly used TF–IDF weighting was used.

For the visual retrieval, a combination of the following six visual features was applied:

- color and edge directivity descriptor (CEDD) [20];

- bag of visual words using SIFT, Scale Invariant Feature Transform, (BoVW) [21];

- fuzzy color and texture histogram (FCTH) [22];

- bag of colors (BoC) [23];

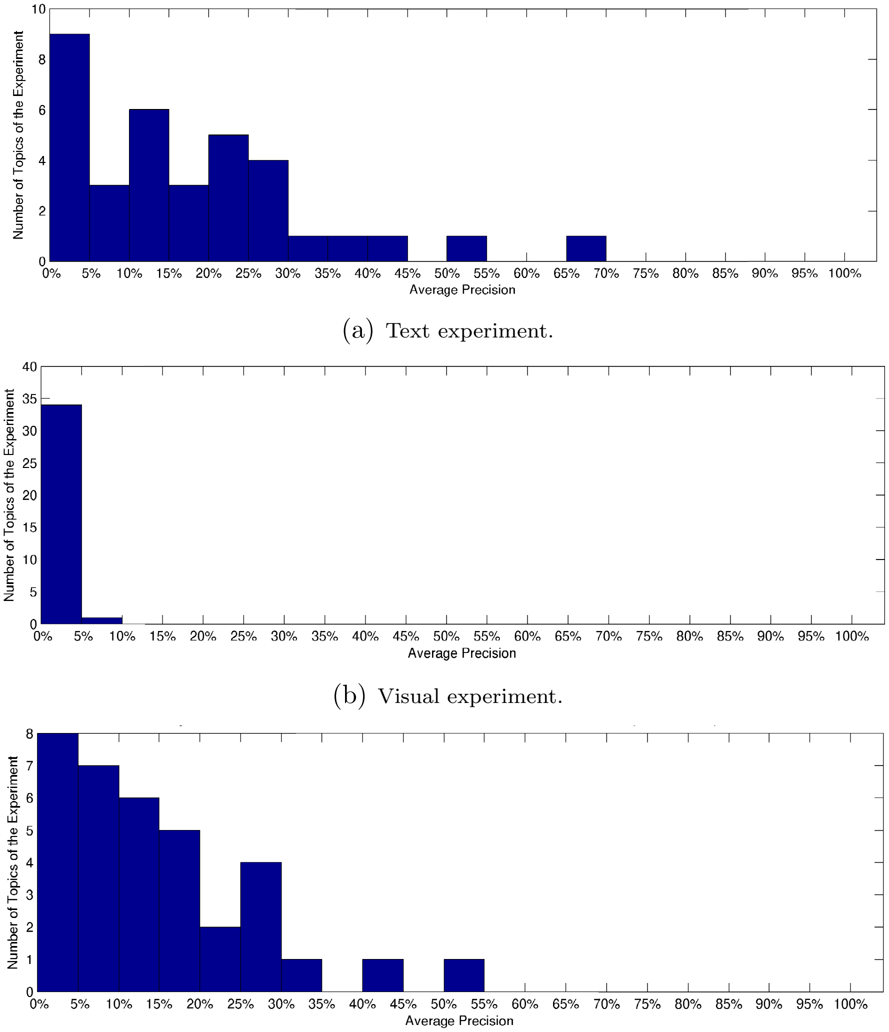
- BoVW with a spatial pyramid matching [24] (BoVW–SPM);

- BoC with $n \times n$ spatial grid (Grid BoC).

García Seco de Herrera et al. [18] describes in more detail the techniques used in the runs.

Figure 1 shows the distribution in terms of average precision of the topics of the textual, visual and mixed runs submitted in 2013. Textual and mixed experiments score really well for only a handful of topics while the visual experiment obtains fairly low scores. One main problem seems to be that the visual results with low performance do not manage to add any information to the text retrieval results. The goal of this paper is to improve the fusion of text and visual search to achieve better performance with multimodal retrieval.

We focus on the text retrieval for a more detailed analysis per topic. Figure 2 shows the difference in the performance of the text retrieval run



(a) Text experiment.

(b) Visual experiment.

(c) Mixed experiment.

Figure 1: The average precision obtained by the text, visual and mixed runs of medGIFT and the number of topics that were in a specific range of results.

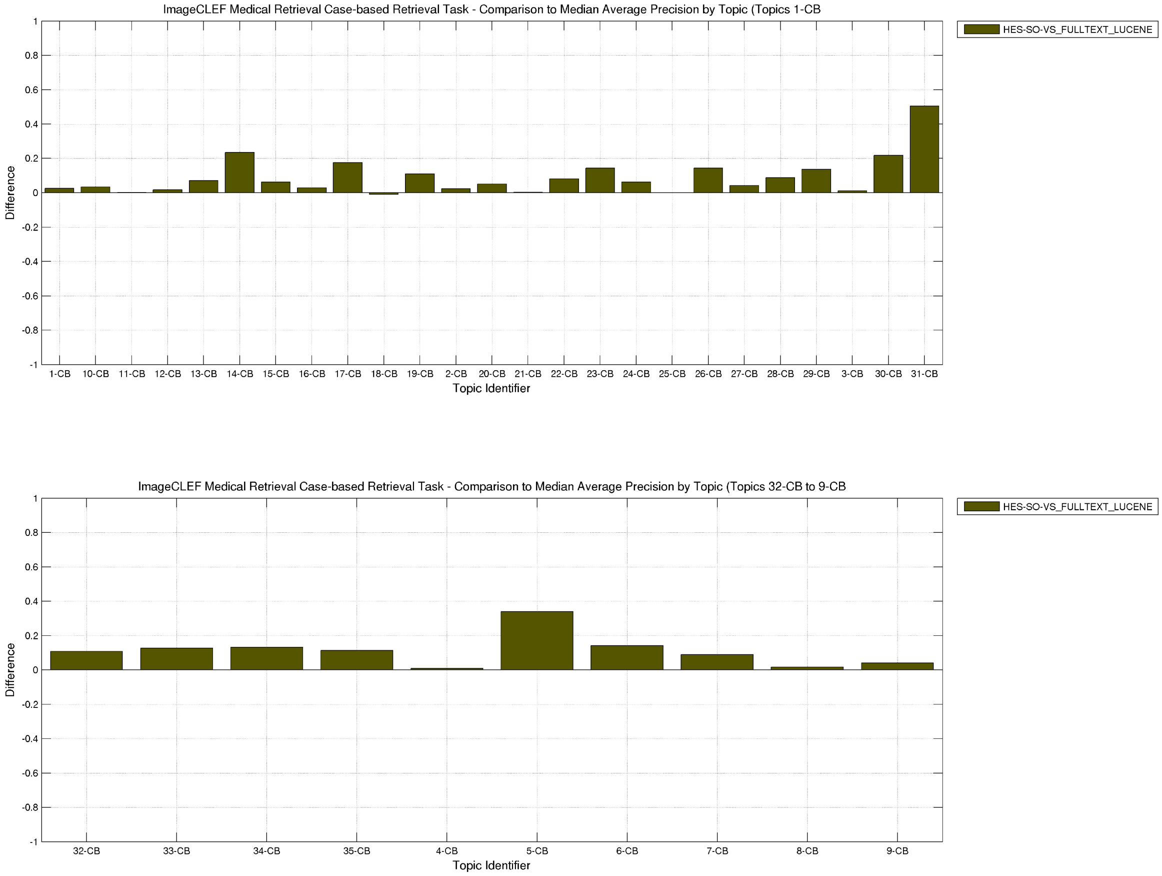


Figure 2: Average precision of the textual run submitted by medGIFT in ImageCLEFmed 2013 compared to the median average precision of all runs submitted.

of medGIFT relative to the median average precision of other runs on the same topic. This difference is positive for most of the topics, meaning that the run is better than the median for virtually all topics. This run is worse than the median only for topic 18 (see Figure 3) where it retrieves 6 relevant documents out of 10. Looking at the R–precision per topic of all the runs submitted to the case–based task in 2013 (see Figure 4), it is visible that topic 18 gets a high variety of scores between 0 and 0.30. Figure 4 also shows that there are some ”difficult” topics for all groups (systems obtain low scores) [25].

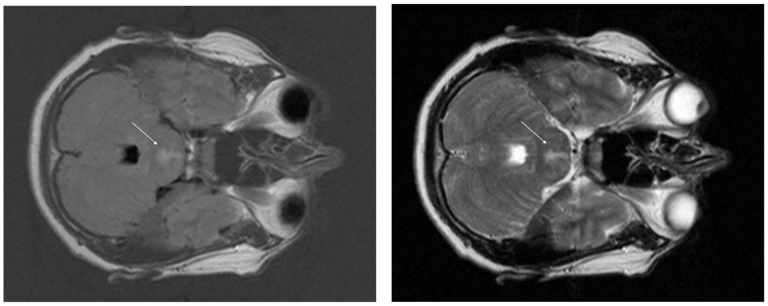


Figure 3: Description of Topic 18 in the case–based retrieval task with its images: 'A 51–year–old woman with HIV who lives in New England has new–onset, focal encephalitis. MRI scan shows a diffuse lesion centrally in the pons.'

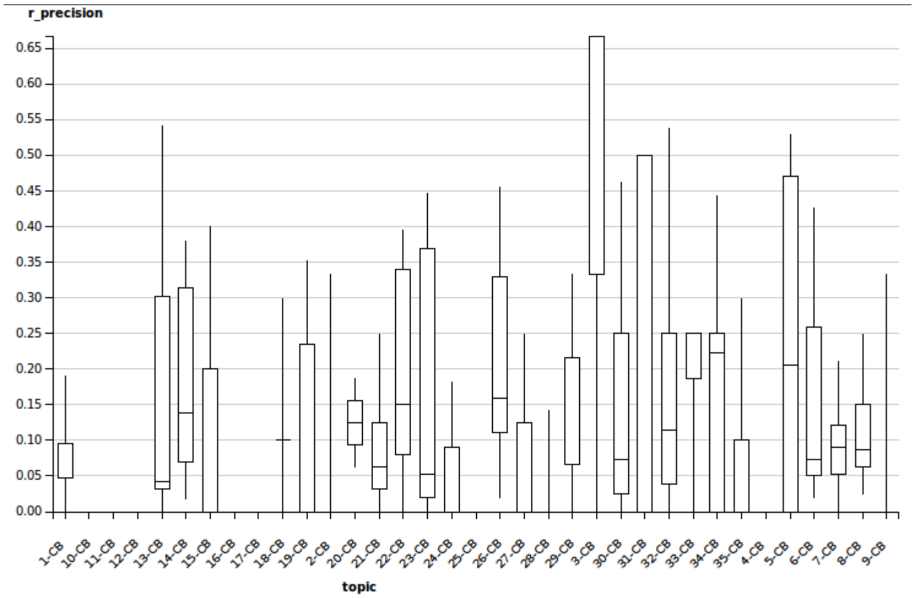


Figure 4: R–precision per topic over all the approaches submitted by the ImageCLEFmed participants.

## 4. Techniques

For this paper, the results obtained in the challenge are used as a baseline, which are also analyzed in Section 3. To enhance the performance of the case–based retrieval task, several fusion strategies were implemented. This section focuses on the description of the fusion strategies as well as a modality classification and a compound figure separation approach. See García Seco de Herrera et al. [18] for further details of the techniques used.

### *4.1. Visual search*

Several fusion strategies were tested to combine results of each of the query images and of several visual descriptors of the same image (see Table 2) to improve visual retrieval. To combine the results/features of multiple query images into a single ranked list, two main fusion strategies were used: early and late fusion [26]. For early fusion, Rocchio's algorithm was applied to merge all feature vectors into a single vector:

$$\vec{q}_m = \alpha\vec{q}_o + \beta\frac{1}{|I_r|}\sum_{\vec{i}_j \in I_r}\vec{i}_j - \gamma\frac{1}{|I_{nr}|}\sum_{\vec{i}_j \in I_{nr}}\vec{i}_j \tag{1}$$

where $\alpha, \beta$ and $\gamma$ are weights, $\vec{i}_m$ is the modified query, $\vec{i}_o$ is the original query, $I_r$ is the set of relevant images and $I_{nr}$ is the set of non–relevant images. In our scenario there are no non–relevant images and we consider the set of relevant images as the original query. Thus, only the second term of the right part of the equation is used [26].

In late fusion, the ranked lists of retrieval results are fused and not the features. The following fusion rules were used for our experiments:

- combSUM

  $$combSUM(i) = \sum_{j=1}^{N_j} S_j(i) \tag{2}$$

  with $N_j$ being the number of descriptors to be combined and $S(i)$ is the score assigned to image $i$;

- combMNZ

  $$combMNZ(i) = F(i) * combSUM(i) \tag{3}$$

  where $F(i)$ is the frequency of image $i$ being returned by one input system with a non–zero score;

- combMAX

$$combMAX(i) = \arg \max_{j=1:N_j} (S_j(i)) \tag{4}$$

- combMIN

$$combMIN(i) = \arg \min_{j=1:N_j} (S_j(i)) \tag{5}$$

- Reciprocal rank fusion:

$$RRF\texttt{score}(i) = \sum_{r \in R} \frac{1}{k + r(i)} \tag{6}$$

  where, $R$ is the set of rankings assigned to the images and $k = 60$ for our study [19];

- Borda

$$Borda(i) = \sum_{r \in R} r(i) \tag{7}$$

For further details on the fusion rules see also [15].

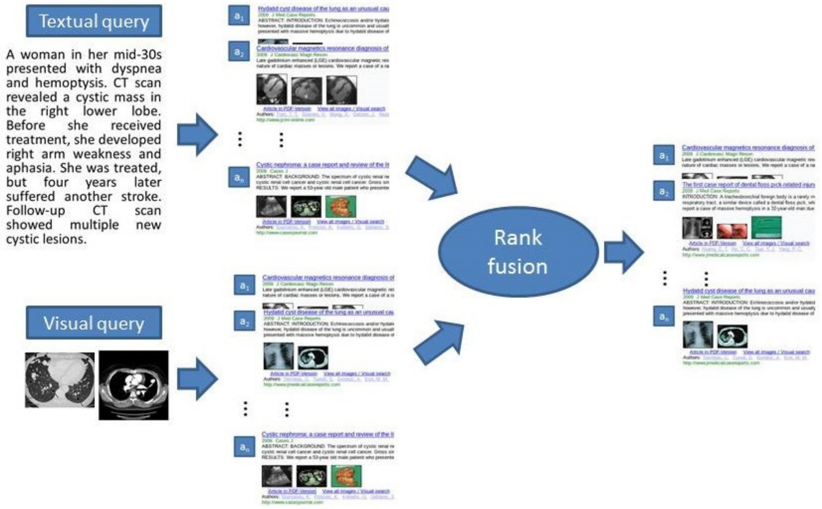
### *4.2. Fusion of visual and textual search*

In Section 3 it was observed that the fusion of visual and textual search was not optimal for many runs submitted to the case–based retrieval task 2013. In this section, several combination strategies are applied.

There are several ways of combining visual and textual retrieval [27]. For the medGIFT experiments two approaches were tested: (1) performing both visual and textual retrieval and then combining the results of the two runs; and (2) using textual retrieval as basis and then rerank results based on visual retrieval.

#### *4.2.1. Combination of visual and textual search*

The Lucene text retrieval system is described in Section 3. The visual search was done by extracting the descriptors mentioned in Section 3. To combine visual and textual ranks, the techniques described in Section 4.1 were applied: Borda; combMAX; combMIN; combMNZ and reciprocal rank. A linear combination of the ranks of the textual and visual runs was also used. Similar to the approach presented by Ramhan et al. [2], the weight of



A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.

Figure 5: The final rank is obtained by combining both visual and text search.

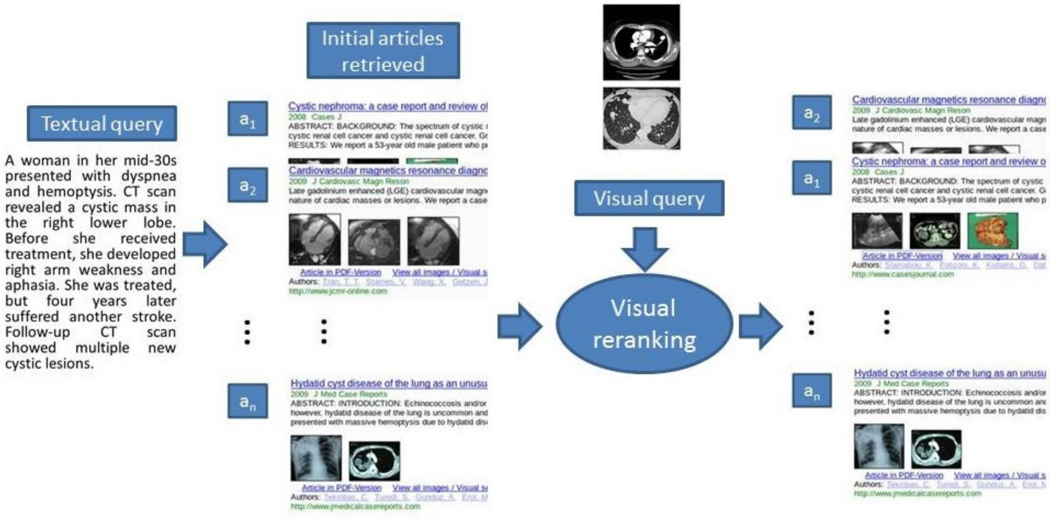
each rank was defined by a function of their performance in terms of Mean Average Precision (MAP):

$$\omega_v = \frac{MAP(T)}{MAP(T) + MAP(V)}; \omega_t = \frac{MAP(V)}{MAP(T) + MAP(V)} \tag{8}$$

where the best MAP scores obtained using text ($MAP(T) = 0.1293$) and visual ($MAP(V) = 0.0204$) search in ImageCLEFmed 2011 [28] were employed. The scores were not taken from the ImageCLEFmed 2012 campaign because some of their queries are reused for ImageCLEFmed 2013 (the database used in this paper). Figure 5 shows the fusion process followed in this section.

### *4.2.2. Visual Reranking*

The reranking method that was used reorders the initial text search results based on the visual descriptors. An initial text search using Lucene returns a ordered set $\mathcal{T} = a_1, ..., a_{1500}$ of the 1500 articles with the largest score values $S(a)$ assigned to the articles $a$, so more than the 1000 required for the final results list. Instead of accepting these results, the articles' images belonging to $\mathcal{T}$ are used to re–rank the results. In the visual re–ranking process the retrieved result list of articles $\mathcal{T}$ is substituted by a set of the images associated with the retrieved articles. Content–based image retrieval is performed using the topics' query images within this image set using the


Textual query
Initial articles retrieved
Visual query
Visual reranking
a1
a2
an
Cystic nephroma: a case report and review of
Cardiovascular magnetics resonance diagno
Hydatid cyst disease of the lung as an unusu

A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.

Figure 6: The proposed visual reranking reorders articles based on images extracted from the initial text search results.

visual features mentioned above. A sorted list of result images is retrieved and is converted back to the article list preserving the order derived by the content–based retrieval. The result is a ordered set $\mathcal{V}$ where an article $a \in \mathcal{V}$ if and only if $a \in \mathcal{T}$ but in a different order. This new order is based on the scores values determined by the visual features extracted from the visual information.

The visual re–ranking process is illustrated in Figure 6.

### *4.3. Modality Classification*

The automatic classification of the image type can be useful in retrieval pipelines. For example, a step of automatic modality filtering can potentially improve the precision of the search by reducing the search space to the set of relevant modalities [4]. Moreover, users often want to restrict search within specific image modalities [17].

The best medGIFT run in ImageCLEFmed 2013 was used as a baseline [18] in this study. The modality classification approach applied in this paper combines visual and text information. The text classification is based on Lucene. For the visual strategy the following descriptors were used: CEDD; BoVW; FCTH; BoC and fuzzy color histogram (FCH) [29].

### *4.4. Compound figure separation*

Separation of compound figures into subparts can improve retrieval accuracy by enabling comparison of images with less noise [30].

The medGIFT approach applies an automated separation process using the technique described in [31]. This approach uses a set of tunable rules that can be learned using a training subset of the figure dataset provided in ImageCLEFmed 2013. It uses solely visual information to detect and separate sub figures in any type of compound figure.

## 5. Experimental results

The main objective of this text is to evaluate the effectiveness of the various fusion methods for the medical case–based retrieval task. Fusion is performed in two cases in the retrieval pipelines: to handle multiple query images in content–based image retrieval and to combine the various visual and textual features. The results were computed with the trec_eval software[8] (version 9.0) following the ImageCLEFmed practice. In this paper four measures with complementary information are shown: MAP; bpref; precision at 10 (P10) and precision at 30 (P30).

Three experiments were conducted for the case–based retrieval task. The first experiment combines the various query and feature fusion techniques when using only visual information. The results of these combinations are shown in Table 2. The visual features fusion are first combined to refer some distinguishable properties of the images as color or texture. Best results on visual runs were achieved when the queries are fused with RRF and the descriptors with combMNZ (see Section 4.1 for more details on the techniques). The experiments indicate that using combMNZ to fuse the various visual features always outperforms other fusing rules in terms of MAP and P10. For the purpose of analysis, the sets queries of each topic were combined into one and computed an overall score.

[8] `http://trec.nist.gov/trec_eval/`

Table 2: Results of the approaches for the case–based retrieval task when using various fusion strategies for visual retrieval. Several query (QF) and descriptor (DF) fusion techniques are combined in the table.

| **Run ID** | **QF** | **DF.** | **MAP** | **Bpref** | **P10** | **P30** |
|---|---|---|---|---|---|---|
| Run1 | Rocchio | Borda | 0.0004 | 0.0092 | 0 | 0 |
| Run2 | Rocchio | combMAX | 0.0004 | 0.0096 | 0 | 0.0029 |
| Run3 | Rocchio | combMIN | 0.0002 | 0.0093 | 0 | 0.0019 |
| Run4 | Rocchio | combMNZ | 0.0008 | 0.0084 | 0.0029 | 0.0048 |
| Run5 | Rocchio | combSUM | 0.0006 | 0.0084 | 0.0029 | 0.0038 |
| Run6 | Rocchio | RRF | 0.0005 | 0.0085 | 0 | 0.0038 |
| Run7 | Borda | Borda | 0.0005 | 0.0060 | 0 | 0.0019 |
| Run8 | Borda | combMAX | 0.0004 | 0.0066 | 0 | 0.0019 |
| Run9 | Borda | combMIN | 0.0002 | 0.0124 | 0 | 0 |
| Run10 | Borda | combMNZ | 0.0009 | 0.0055 | 0.0029 | 0.0038 |
| Run11 | Borda | combSUM | 0.0005 | 0.0060 | 0.0029 | 0.0029 |
| Run12 | Borda | RRF | 0.0012 | 0.0061 | 0.0086 | 0.0057 |
| Run13 | combMAX | Borda | 0.0006 | 0.0062 | 0.0066 | 0.0019 |
| Run14 | combMAX | combMAX | 0.0006 | 0.0089 | 0.0057 | 0.0029 |
| Run15 | combMAX | combMIN | 0.0003 | **0.0156** | 0 | 0.0019 |
| Run16 | combMAX | combMNZ | 0.0036 | 0.0077 | **0.0114** | 0.0057 |
| Run17 | combMAX | combSUM | 0.0021 | 0.0077 | 0.0086 | 0.0067 |
| Run18 | combMAX | RRF | 0.0013 | 0.0066 | 0.0086 | 0.0048 |
| Run19 | combMIN | Borda | 0.0005 | 0.0077 | 0.0029 | 0.0029 |
| Run20 | combMIN | combMAX | 0.0006 | 0.0091 | 0.0086 | 0.0038 |
| Run21 | combMIN | combMIN | 0.0003 | 0.0172 | 0 | 0.0019 |
| Run22 | combMIN | combMNZ | 0.0032 | 0.008 | 0.0086 | 0.0057 |
| Run23 | combMIN | combSUM | 0.0015 | 0.0079 | 0.0057 | 0.0057 |
| Run24 | combMIN | RRF | 0.0011 | 0.0060 | 0.0086 | 0.0067 |
| Run25 | combMNZ | Borda | 0.0005 | 0.0061 | 0.0029 | 0.001 |
| Run26 | combMNZ | combMAX | 0.0004 | 0.0077 | 0 | 0.0038 |
| Run27 | combMNZ | combMIN | 0.0001 | 0.0111 | 0 | 0.001 |
| Run28 | combMNZ | combMNZ | 0.0029 | 0.0058 | 0.0086 | 0.0067 |
| Run29 | combMNZ | combSUM | 0.0011 | 0.0053 | 0.0057 | 0.0057 |
| Run30 | combMNZ | RRF | 0.0008 | 0.0055 | 0.0029 | 0.0038 |
| Run31 | combSUM | Borda | 0.0005 | 0.006 | 0.0029 | 0.0019 |
| Run32 | combSUM | combMAX | 0.0005 | 0.0084 | 0.0057 | 0.0038 |

Continued on next page...

...continued from previous page

| **Run ID** | **QF** | **DF** | **MAP** | **Bpref** | **P10** | **P30** |
|---|---|---|---|---|---|---|
| Run33 | combSUM | combMIN | 0.0002 | 0.0127 | 0 | 0.0019 |
| Run34 | combSUM | combMNZ | 0.0033 | 0.0075 | 0.0086 | **0.0076** |
| Run35 | combSUM | combSUM | 0.0014 | 0.0067 | 0.0086 | 0.0067 |
| Run36 | combSUM | RRF | 0.0009 | 0.0051 | 0.0029 | 0.0048 |
| Run37 | RRF | Borda | 0.0005 | 0.0057 | 0 | 0.0019 |
| Run38 | RRF | combMAX | 0.0004 | 0.0070 | 0 | 0.0038 |
| Run39 | RRF | combMIN | 0.0002 | 0.0121 | 0 | 0 |
| Run40 | RRF | combMNZ | **0.0037** | 0.0129 | 0.0086 | 0.0067 |
| Run41 | RRF | combSUM | 0.0011 | 0.0060 | 0.0086 | 0.0067 |
| Run42 | RRF | RRF | 0.0010 | 0.0047 | 0.0029 | 0.0057 |

The result achieved with the text approach is shown in Table 3. Since the case–based task has been running, textual approaches always achieved better results than visual or multimodal runs.

Table 4 shows the performance of the combination of textual and visual information. To carry out these experiments the best visual approach was used (see Table 2) to help in better task accomplishment. The fusion used for Run40 was applied because it obtained the best results in terms of MAP and good results in terms of Bpref, P10 and P30.

The results of the second experiment (Runs 44–50) show the performance of the fusion of the independent textual and visual results. The fusion rules described in Section 4.2.1 were applied for this experiment. The best result was obtained by Run50 using a linear combination of text and visual search (MAP=0.1795). Linear combination is one of the simplest and most widely used fusion methods [32] and betters the fusion more than the other used approaches. Runs 45, 48 and 50 also outperform the best mixed run submitted to this task in ImageCLEFmed 2013 (MAP=0.1608).

In the third experiment (Run51), text retrieval is used to extract a subset of all potential relevant images. In this experiment the retrieval performance was poor, potentially because this visual approach is not optimal for a task where the number of relevant articles is very low.

As described in Sections 4.3 and 4.4, modality classification and compound figure separation can be integrated into the retrieval system to enhance the performance. We implemented both, modality classification and

Table 3: Results of the approaches at the case–based retrieval task when using only text.

| **Run ID** | **MAP** | **Bpref** | **P10** | **P30** |
|---|---|---|---|---|
| Run43 | 0.1791 | 0.1630 | 0.2143 | 0.1581 |

Table 4: Results of the approaches at the case–based retrieval task when using various fusion strategies to combine visual and textual information ('Multimodal fusion').

| **Run ID** | **Multimodal fusion** | **MAP** | **Bpref** | **P10** | **P30** |
|---|---|---|---|---|---|
| Best ImageCLEF | Only textual | 0.1608 | 0.1426 | 0.1800 | 0.1257 |
| Run44 | Borda | 0.1302 | 0.1230 | 0.1371 | 0.1105 |
| Run45 | combMAX | 0.1770 | 0.1625 | 0.2143 | **0.1571** |
| Run46 | combMIN | 0.1505 | 0.157 | 0.2171 | 0.1438 |
| Run47 | combMNZ | 0.1197 | 0.1257 | 0.1714 | 0.1133 |
| Run48 | combSUM | 0.1741 | 0.1609 | **0.2229** | 0.161 |
| Run49 | RRF | 0.1084 | 0.1011 | 0.1543 | 0.1114 |
| Run50 | Linear | **0.1795** | **0.1627** | 0.2086 | **0.1571** |
| Run51 | Visual reranking | 0.0012 | 0.0214 | 0.0114 | 0.0067 |

compound figure separation but the results are currently only evaluated separately and the inclusion into case-base retrieval has not yet been finalized. The modality classification run submitted by medGIFT in 2013 obtained an accuracy of 69.63% for the 31 classes used among the best runs in ImageCLEF 2013. Moreover, the compound figure separation approach achieved the best accuracy of all the participants (84.64%) [18]. By limiting the search of similar cases to cases that have images of the same modality we think that retrieval quality can be increased. By omitting all non-clinical images from the results list, the noise present should also be removed, leading potentially to better results. In the same way, the presence of compound figures makes retrieval more difficult as those figures may contain clinically relevant images of the same type as the queries, but as they are only a small part of the overall figure this information can be missed. The query topics themselves only contain clinical images and no compound figures, but in the results set the entire spectrum of images can be found.

## 6. Conclusions

To address the ImageCLEF medical case–based retrieval task, a multimodal approach was applied in this text. Experimental results demonstrate the impact of the types of fusion rules used on the retrieval performance. The retrieval performance can be enhanced more effectively when there is a sufficient number of relevant articles, which is not the case for all the presented topics. Despite the low performance of the visual search, the effectiveness of the multimodal approaches is improving and provides evidence that multimodal medical case–based retrieval systems can obtain good performance. Results outperformed the best multimodal runs submitted to ImageCLEFmed 2013 by a weighted linear combination of visual and text retrieval using ranks. Moreover, the fusion method applied in this paper can be easily reproduced by other researchers and may serve for further investigation on the fusion of multimodal search.

A major challenge is the low performance of the visual retrieval approach for the case–based task. To overcome this, a medical image modality classification is also presented in this paper to filter out non–relevant images, which has the possibility to remove some noise from the results. In addition, a compound figure separation method is introduced for distinguishing the parts of images relevant to a given query and therefore focusing the search on the sub figures. The two techniques are not fully integrated with the cases–based retrieval yet, but they have the potential to increase performance and remove part of the noise. Since the visual retrieval performance is rather low, future work will concentrate more on the extraction of proper visual descriptors and ways to increase the visual performance. Better visual runs will also improve the overall performance as currently visual information adds relatively little to the overall performance. The text runs can also still be improved, for example by query expansion methods using external knowledge such as MeSH term co–occurrence. Such mapping to MeSH or UMLS has often obtained the best performance in past years and again multimodal performance should also profit from performance increase in each of the parts.

## 7. Acknowledgments

This work was partly supported by the EU $7^{th}$ Framework Program in the context of the Khresmoi project (FP7–257528).

## References

[1] G. Csurka, S. Clinchant, G. Jacquet, Medical image modality classification and retrieval, in: International Workshop on Content–Based Multimedia Indexing, CBMI, IEEE, 2011, pp. 193–198.

[2] M. M. Rahman, D. You, M. S. Simpson, S. K. Antani, D. Demner-Fushman, G. R. Thoma, Multimodal biomedical image retrieval using hierarchical classification and modality fusion, International Journal of Multimedia Information Retrieval 2 (3) (2013) 159–173.

[3] M. C. Díaz-Galiano, M. T. Martín-Valdivia, L. A. Ureña López, M. A. García-Cumbreras, SINAI at ImageCLEF 2009 medical task, in: Working Notes of the 2009 CLEF Workshop, Corfu, Greece, 2009.

[4] J. Kalpathy-Cramer, W. Hersh, Multimodal medical image retrieval: image categorization to improve search precision, in: Proceedings of the international conference on Multimedia information retrieval, MIR '10, ACM, New York, NY, USA, 2010, pp. 165–174.

[5] J. I. Westbrook, E. W. Coiera, A. S. Gosling, Do online information retrieval systems help experienced clinicians answer clinical questions?, Journal of the American Medical Informatics Association 12 (3) (2005) 315–321.

[6] B. Caputo, H. Müller, B. Thomee, M. Villegas, R. Paredes, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. Martinez Gomez, I. Garcia Varea, C. Cazorla, ImageCLEF 2013: the vision, the data and the open challenges, in: Working Notes of CLEF 2013 (Cross Language Evaluation Forum), 2013.

[7] A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, in: Working Notes of CLEF 2013 (Cross Language Evaluation Forum), 2013.

[8] H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, I. Eggel, Overview of the ImageCLEF 2012 medical image retrieval and classification tasks, in: Working Notes of CLEF 2012 (Cross Language Evaluation Forum), 2012.

[9] R. Yan, A. G. Hauptmann, The combination limit in multimedia retrieval, in: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA '03, ACM, New York, NY, USA, 2003, pp. 339–342.

[10] A. Hanjalic, R. Lienhart, W.-Y. Ma, J. R. Smith, The holy grail of multimedia information retrieval: So close or yet so far away?, in: Proceedings of the IEEE, Vol. 96, 2008, pp. 541–547.

[11] T.-T. Pham, N. E. Maillot, J.-H. Lim, J.-P. Chevallet, Latent semantic fusion model for image retrieval and annotation, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM'07, ACM, New York, NY, USA, 2007, pp. 439–444.

[12] Y. Cao, Y. Li, H. Müller, C. E. Kahn, Jr., E. Munson, Multi–modal medical image retrieval, in: SPIE Medical Imaging, 2011.

[13] Y. Gkoufas, A. Morou, T. Kalamboukis, IPL at ImageCLEF 2011 medical retrieval task, in: Working Notes of CLEF 2011, 2011.

[14] A. Mourão, F. Martins, J. a. Magalhães, Assisted query formulation for multimodal medical case-based retrieval, in: Proceedings of ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine, 2013.

[15] A. Depeursinge, H. Müller, Fusion techniques for combining textual and visual information retrieval, in: H. Müller, P. Clough, T. Deselaers, B. Caputo (Eds.), ImageCLEF, Vol. 32 of The Springer International Series On Information Retrieval, Springer Berlin Heidelberg, 2010, pp. 95–114.

[16] D. Demner-Fushman, S. Antani, M. S. Simpson, G. R. Thoma, Design and development of a multimodal biomedical information retrieval system, Journal of Computing Science and Engineering 6 (2) (2012) 168–177.

[17] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, S. Kriewel, H. Müller, A survey on visual information search behavior and requirements of radiologists, Methods of Information in Medicine 51 (6) (2012) 539–548.

[18] A. García Seco de Herrera, D. Markonis, R. Schaer, I. Eggel, H. Müller, The medGIFT group in ImageCLEFmed 2013, in: Working Notes of CLEF 2013 (Cross Language Evaluation Forum), 2013.

[19] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2009, pp. 758–759.

[20] S. A. Chatzichristofis, Y. S. Boutalis, CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval, in: Lecture notes in Computer Sciences, Vol. 5008, 2008, pp. 312–322.

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

[22] S. A. Chatzichristofis, Y. S. Boutalis, FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval, in: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Service, 2008, pp. 191–196.

[23] A. García Seco de Herrera, D. Markonis, H. Müller, Bag of colors for biomedical document image classification, in: H. Greenspan, H. Müller

(Eds.), Medical Content–based Retrieval for Clinical Decision Support, MCBR–CDS 2012, Lecture Notes in Computer Sciences (LNCS), 2013, pp. 110–121.

[24] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Washington, DC, USA, 2006, pp. 2169–2178.

[25] S. Mizzaro, The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation?, in: Advances in Information Retrieval, Lecture Notes in Computer Science, Springer, 2008, pp. 642–646.

[26] A. García Seco de Herrera, D. Markonis, I. Eggel, H. Müller, The medGIFT group in ImageCLEFmed 2012, in: Working Notes of CLEF 2012, 2012.

[27] J. Glasgow, I. Jurisica, Integration of case–based and image–based reasoning, in: AAAI Workshop on Case–Based Reasoning Integrations, AAAI Press, Menlo Park, California, 1998, pp. 67–74.

[28] D. Markonis, A. García Seco de Herrera, I. Eggel, H. Müller, The medGIFT group in ImageCLEFmed 2011, in: Working Notes of CLEF 2011, 2011.

[29] J. Han, K.-K. Ma, Fuzzy color histogram and its use in color image retrieval, IEEE Transactions on Image Processing 11 (8) (2002) 944–952.

[30] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, Creating a classification of image types in the medical literature for visual categorization, in: SPIE medical imaging, 2012.

[31] A. Chhatkuli, D. Markonis, A. Foncubierta-Rodríguez, F. Meriaudeau, H. Müller, Separating compound figures in journal articles to allow for subfigure classification, in: SPIE Medical Imaging, 2013.

[32] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: A survey, Multimedia Systems 16 (6) (2010) 345–379.