Shangri–La: a Medical Case–based Retrieval Tool

Alba G. Seco de Herrera[*], Roger Schaer, Henning Müller
University of Applied Sciences Western Switzerland (HES–SO),
Sierre, Switzerland
Email: albagarcia@nih.gov

November 28, 2016

Abstract

Large amounts of medical visual data are produced in hospitals daily and made available continuously via publications in the scientific literature, representing the medical knowledge. However, it is not always easy to find the desired information and in clinical routine the time to fulfil an information need is often very limited. Information retrieval systems are a useful tool to provide access to these documents/images in the biomedical literature related to information needs of medical professionals. Shangri-La is a medical retrieval system that can potentially help clinicians to make decisions on difficult cases. It retrieves articles from the biomedical literature when querying a case description and attached images. The system is based on a multi-modal retrieval approach with a focus on the integration of visual information connected to text. The approach includes a query-adaptive multi-modal fusion criterion that analyses if visual features are suitable to be fused with text for the retrieval. Furthermore, image modality information is integrated in the retrieval step. The approach is evaluated using the ImageCLEFmed 2013 medical retrieval benchmark and can thus be compared to other approaches. Results show that the final approach outperforms the best multi-modal approach submitted to ImageCLEFmed 2013.

Keywords: Medical visual information retrieval, ImageCLEF, Medical case retrieval, Query adaptive multi–modal fusion, Shangri–La, Classification

[*] Alba G. Seco de Herrera is currently working at the National Library of Medicine (NLM/NIH), Bethesda, MD, USA

Introduction

Images are produced in hospitals in ever–increasing numbers (Akgül et al., 2011) and also with a quickly increasing variety (protocols, different machines, contrast agents, etc.) as they provide crucial information for diagnosis, treatment planning and other tasks. A recent European report estimates that 30% of the global digital storage was occupied by medical image data in 2010 (*Riding the wave: How Europe Can Gain from the Rising Tide of Dcientific Data*, 2010). Besides image production and storage in clinical patient records, images are also made available via biomedical publications in fundamental or clinical research. The number of biomedical articles published grew at a double–exponential pace between 1986 and 2006 according to (Hunter & K. Cohen, 2006), which also underlines the fact that new tools are needed to manage the increasing amount of data that are accessible and represent a large part of medical knowledge.

Many physicians have regular information needs during clinical work, teaching preparation and research activities (W. Hersh, Jensen, Müller, Gorman, & Ruch, 2005; Müller et al., 2006). Most of these needs are expressed via text queries, but also the visual content carries a large part of the medical information stored. Therefore, there is a need for searching through the immense collection of medical images in hospitals and on the World Wide Web, making the data accessible for reuse. Studies show that the time for answering a clinical information need using text information retrieval (IR) systems such as MedLine is around 30 minutes (W. R. Hersh & Hickam, 1998), while clinicians state to have approximately five minutes available (Hoogendam, Stalenhoefand, Robbé, & Overbeke, 2008). Finding relevant information quicker is thus an important task to bring search into clinical routine (Mendelson & Rubin, 2013). Images have an important role, as their content can in general be understood much quicker than text content, particularly for filtering out non–relevant content.

Many tools have been developed for these tasks over the past 20 years (Kalpathy-Cramer et al., 2015). Retrieval and classification of medical images have been explored to get additional information for reading and interpretation of medical cases (Uwimana & Ruiz, 2008) when open questions remain. This helps clinicians in their daily work.

Although text queries are most commonly used, the visual information of the images can enrich the search. Images represent an important part of the content in many publications and searching for medical images has become common in retrieval applications, particularly for radiologists. Visual retrieval has shown to be complementary to text retrieval approaches and images can well help to represent the content of scientific articles, particularly in applications using small interfaces such as mobile phones (Depeursinge, Duc, Eggel, & Müller, 2012). Medical case–based retrieval (taking into account several images, text and potentially other data of the case) has also been proposed by other authors over the past 10 years (Müller et al., 2007; Welter, Deserno, Fischer, Günther, & Spreckelsen, 2011).

In this paper, a web–based retrieval interface called *Shangri–La* is presented in addition to an optimised medical retrieval technology based on previous work. This interface integrates a multi–modal retrieval approach that is also described in this paper. The combination of several techniques shows to increase retrieval performance and it delivered the currently best performance on this publicly available data set. Both, the improved retrieval system combining several techniques and the multimodal query interface constitute the main contributions of this paper. The system was refined based on two user tests with a previous system that were performed in the Khresmoi project (Markonis, Baroz, Ruiz de Castaneda, Boyer, & Müller, 2013; Markonis et al., 2015). Several adaptations to the interface and the retrieval techniques were done for the prototypes described in this paper.

The remainder of the paper is organised as follows. First, it reviews recent medical retrieval systems. A description of the framework used to evaluate the model is given followed by the the experiments carried out. Then, the web–based retrieval interface called *Shangri–La* is presented. Finally, the results are discussed and the conclusions are given.

Related systems

Due to the many challenges in biomedical information retrieval, research has been attracting increasing attention, and many approaches have been proposed (Li, Shi, & D.Frank, 2011). This section presents a few retrieval systems that use multi–modal information for the search. A more detailed overview on systems specialised on biomedical search can be found in Gottlieb et al. (Gottlieb & Marino, 2014).

Well–known retrieval systems such as ARRS GoldMiner[1] or Yottalook[2] retrieve images and articles from peer–reviewed biomedical journals but en-

[1] ARRS GoldMiner provides rapid access to published medical images of the peer–reviewed literature (see `http://goldminer.arrs.org/`).

[2] Yottalook is a medical image search engine that provides decision support at the point of care (see `http://www.yottalook.com/`).

tirely based on text content. On the other hand, there are systems that provide only Content–Based Image Retrieval (CBIR), such as IRMA[3] or img(Anaktisi)[4]. The IRMA system retrieves images visually similar to a query image with respect to a selected set of visual features. Images in the IRMA database are mapped to a classification of the image type containing body part, image modality, biosystem imaged and view, but not diagnosis. This limits its use for practical retrieval tasks. img(Anaktisi) uses a set of visual features designed to be small in terms of size and storage. The features include color and texture information. In addition, img(Anaktisi) includes an Auto Relevance Feedback (ARF) technique to optimally readjust the initial retrieval results based on user feedback.

Regarding multi–modal retrieval systems, the Center of Informatics and Information Technology group (CITI) develops NovaMedSearch[5] as a medical multi–modal search engine that can retrieve either visually similar images or related medical cases (Mourão & Martins, 2013). NovaMedSearch extracts two types of visual features and fuses them with the textual information using inverse square rank fusion. In addition, NovaMedSearch provides an interactive query expansion by suggesting alternative medical terms to the user.

The NLM[6] (National Library of Medicine) provides Open–i[7] (Demner-Fushman, Antani, Simpson, & Thoma, 2012), a service to search and retrieve abstracts and images from the open access literature and other biomedical collections available. Open–i generates enriched article representations, processing the text in the image captions and visual image content independently. In addition, each enriched representation contains meta–information that is used to filter and re–rank the retrieved list.

To improve retrieval quality a successful classification of images into image types (e.g. X–ray, ultrasound, CT, etc) can be applied to filter out irrelevant image types (Rahman et al., 2013), such as general graphs that

[3] IRMA is a project at the Aachen University of Technology (RWTH Aachen), Germany that aims to develop and implement high–level methods for CBIR with prototypical application for medical tasks on an internal image archive with selection case from radiology (see `http://ganymed.imib.rwth-aachen.de/irma/`).

[4] img(Anaktisi) is a web CBIR application that provides retrieval services for various image databases (see `http://orpheus.ee.duth.gr/anaktisi/`).

[5] NovaMedSearch is a multi–modal (text and image) medical search engine designed to find relevant medical images or cases using the Open Access Subset of PubMed Central (see `http://medical.novasearch.org/`).

[6] The NLM maintains and makes available a vast data collection and produces electronic information resources on a wide range of topics (see `http://nlm.nih.gov/`).

[7] Open–i is an open access biomedical search engine (see `http://openi.nlm.nih.gov/`).

are frequent in the medical literature. Already, many web–accessible search systems such as Goldminer or Yottalook allow users to limit the search results to a particular modality (Müller, García Seco de Herrera, et al., 2012), as this is a feature often requested by end users (Markonis et al., 2012). However, the extracted modality information only uses the caption text and not the visual content of the images.

System evaluation

The ImageCLEF 2013 medical retrieval database (García Seco de Herrera, Kalpathy-Cramer, Demner Fushman, Antani, & Müller, 2013) is used in this work to carry out the experimental analysis of the proposed retrieval system (see next section). The database is a subset of 300,000 images of 75,000 articles from the PubMed Central (PMC)[8] database that contained over 4 million images in early 2016 and is growing very quickly.

Tasks

ImageCLEFmed has proposed several tasks over the years since 2004 (Kalpathy-Cramer et al., 2015). Two types of tasks are used to evaluate the system presented in this work: the medical case–based retrieval task and the modality classification task.

Medical Case–based Retrieval

In the case–based retrieval task, a case description is provided as query. The goal is to retrieve articles from the biomedical literature that are useful in differential diagnosis.

Each query topic consists of a case description with patient anamnesis, limited symptoms and test results including imaging studies (but not the final diagnosis). Each of the query topics is accompanied by one to three images. An example of a query topic can be seen in Figure 1. 35 query topics were given to the participants in 2013. For all 35 topics the first N results of participating systems were judged for relevance by physicians to create the ground truth for the task. The ground truth is thus not absolutely complete but covers a large part.

[8] PubMed Central (PMC) is a free full–text archive of biomedical and life sciences literature at the U.S. National Institute of Health's National Library of Medicine (NIH/NLM) (see `http://www.ncbi.nlm.nih.gov/pmc/`).

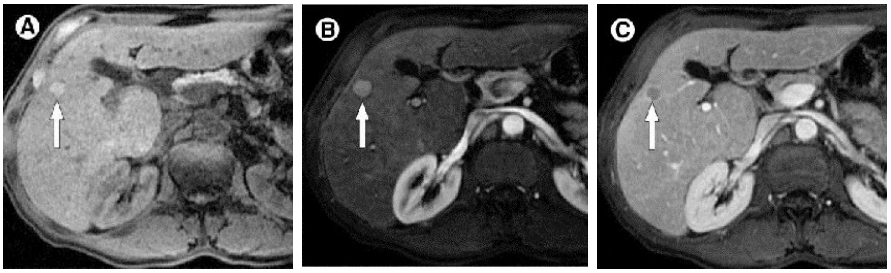


Figure 1: Images from one of the query topics in the medical case–based retrieval task of ImageCLEFmed 2013. They correspond to the textual query "A 56–year–old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase".

The results were analysed with the trec_eval software[9] (version 9.0) following the ImageCLEFmed 2013 practice. The trec_eval software is available to the retrieval research community, so organizations can evaluate their own retrieval systems at any time based on the exact same implementation of performance measures. This software computes a large array of measures including the ones used for ImageCLEFmed 2013: Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP), bpref, precision at 10 (P10) and precision at 30 (P30).

MAP was chosen as the lead measure although all measures cited above were also analysed. Since MAP is the mean of the Average Precision (AP) for all the query topics, it favours systems that return more relevant documents at the top of the list. For a single query topic, the AP approximates the area under the uninterpolated precision–recall curve. Therefore, the MAP is approximately the average area under the precision–recall curve for the given set of queries. However, the maximum MAP that a system can achieve is limited by its recall, and systems can have very high early precision despite having low MAP (Kalpathy-Cramer & Müller, 2011).

When using web–based interfaces, users are interested in how many good results there are on the first page or the first three pages. Precision measures such as P10 or P30 show the ability of a system to present only or mainly relevant items high in the results list. GMAP measures improvements for low–performing query topics by stronger weighting query topics with very

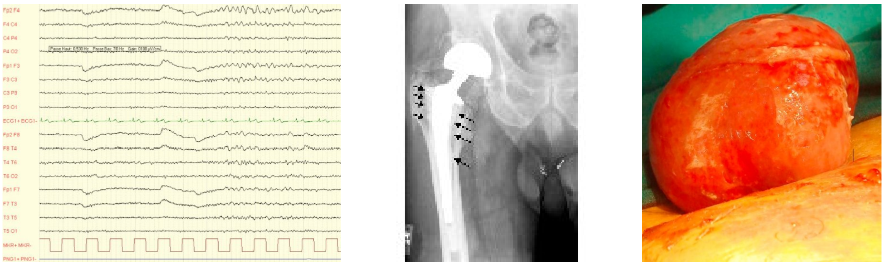
[9] trec_eval is a freely available tool designed for evaluation of various IR systems. It handles streams of documents, queries and relevance judgements (see `http://trec.nist.gov/trec_eval/`)

low AP. The bpref measure is designed for situations where relevance judgements are known to be incomplete. It computes a preference relation of whether items judged relevant are retrieved ahead of items judged irrelevant. When the judgements are complete bpref and MAP are very highly correlated. However, if the judgements are incomplete, rankings of systems by bpref can give a better idea than a ranking by MAP as it takes these non–judged items into account.

For more details on the measures chosen see (Voorhees & Buckland, 2006).

Modality classification

The goal of the modality classification is to classify the images into medical modalities and other image types, such as Computer Tomography (CT), X–ray or general graphs (see Figure 2). An ad hoc hierarchy with 31 classes was used for the modality classification (Müller, Kalpathy-Cramer, Demner-Fushman, & Antani, 2012). In 2013, 2,582 training images and 2,901 test images were provided to the participants.



(a) Printed signals, waves: electroencephalography.

(b) Radiology: x–Ray, 2D Radiography.

(c) Visible light photography: other organs.

Figure 2: Images from three modalities in the modality classification task of ImageCLEFmed 2013.

The number of images per class in the training and test sets of the ImageCLEFmed 2013 classification task varies from fewer than ten to several hundred. Therefore, a training set expansion to improve the classification accuracy based on the image modalities (García Seco de Herrera, Markonis, et al., 2015) was done for the work described here. Finally, a training set with 17,042 images containing manually attached labels is used. The additional images are all part of PubMed Central and are not part of the test data set.

The evaluation of this task is done in terms of classification accuracy, which is the proportion of images for which the classifier can correctly predict the class.

Case–based Retrieval Techniques

ParaDISE (Parallel Distributed Image Search Engine) (Markonis, Schaer, García Seco de Herrera, & Müller, Submitted) is a retrieval engine that allows indexing and searching images using visual features and textual context. The main characteristics of ParaDISE are the scalability, flexibility, expandability and interoperability, allowing the integration of new functionalities.

New components for specific steps and new algorithms for the existing components were added to ParaDISE to develop the medical case–based retrieval system. Earlier versions of the system were used for user tests and based on the comments the current system was developed (Markonis et al., 2015). This section describes the techniques developed to create it.

Basic performance

First, the multi–modal retrieval baseline (Markonis, Eggel, García Seco de Herrera, & Müller, 2011; García Seco de Herrera, Markonis, Eggel, & Müller, 2012; García Seco de Herrera, Markonis, Schaer, Eggel, & Müller, 2013) is created. Figure 3 shows all the principal components of this baseline.

The approach retrieves a ranked list of images instead of articles. The list is converted back to an article list preserving the order derived by the image–based retrieval. Each article receives the score of the highest scoring image that it contains.

The Lucene[10] Information Retrieval library is used to establish the text retrieval baseline. Lucene was chosen for the experiments because it is fast and easy to install and use. Details about the way Lucene is used and configured for the experiments are provided here (mainly using the default settings):

- *English Analyzer* – the English analyser is used for tokenization, stemming and stop word removal of all captions and queries;

- *Multiple boolean operators* – in order to maximise the score of relevant documents, each text query is executed three times: using the OR

[10] Apache Lucene is a project that develops open-source text retrieval software including indexing and search technology (see `http://lucene.apache.org/`).

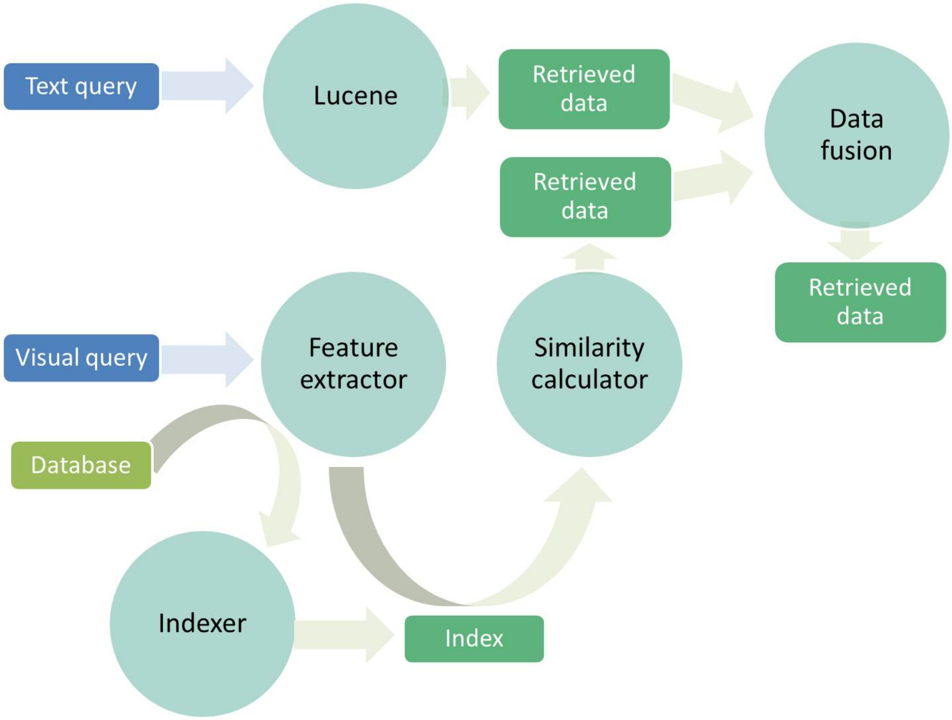


Figure 3: Outline of the principal elements of the multi–modal retrieval baseline.

operator to parse a text query, using the AND operator to parse a text query, and finally putting the query into quotes ("...") to perform an exact phrase search. The three result lists are then fused using a reciprocal rank fusion rule (Cormack, Clarke, & Büttcher, 2009), in this way boosting the ranking of exact matches;

- *Term frequency–inverse document frequency (tf/idf) similarity* – the commonly used tf/idf weighting is applied for ranking images and documents.

For the visual content of the images multiple features are used, as this was a successfully used technique in ImageCLEFmed in the past (Müller, García Seco de Herrera, et al., 2012). A set of low–level visual descriptors is selected from the descriptor bank of ParaDISE (Schaer, Markonis, & Müller, 2014; Markonis et al., Submitted) and their combination is explored (García Seco de Herrera, Markonis, Schaer, et al., 2013) to optimise the outcomes. The following descriptors are chosen after the performance tests on a different imageCLEF data set:

- *Bag of Visual Words (BoVW) using the Scale Invariant Feature Transform (SIFT) (Lowe, 2004) with a spatial pyramid matching (Lazebnik, Schmid, & Ponce, 2006) (BoVW–SPM)* — each image is represented by a histogram symbolizing a set of local descriptors represented in visual words from a previously learned vocabulary; spatial information is added to the BoVW–SIFT descriptor;

- *Bag of Colours (BoC) (García Seco de Herrera, Markonis, & Müller, 2013) with an $n \times n$ spatial grid (Grid BoC)* — each image is represented by a histogram symbolizing the colours from a previously learned vocabulary; spatial information is added to the BoC descriptor;

- *Colour and Edge Directivity Descriptor (CEDD)* (Chatzichristofis & Boutalis, 2008a) — colour and texture information is produced by a 144 bin histogram. Little computation is needed for its extraction;

- *Tamura texture* (Tamura, Mori, & Yamawaki, 1978) — this descriptor extracts six visual properties: coarseness, contrast, directionality, line–likeness, regularity and roughness.

For the visual indexing, the *histogram intersection* (Swain & Ballard, 1991) is used for the similarity comparison for each of the visual descriptors. Histogram intersection has been used successfully as a similarity measure for image retrieval and previous studies have shown that it is robust (Chakravarti & Meng, 2009; Boughorbel, Tarel, & Boujemaa, 2005; Swain & Ballard, 1991).

Data fusion is applied in order to achieve more accurate retrieval results than the retrieval results achieved by single sources (Gkoufas, Morou, & Kalamboukis, 2011). Two types of fusion algorithms are used for the work described here:

- *Visual feature fusion* – results of various visual descriptors are combined;

- *Multi–modal fusion* – information from various sources (images and text) are combined.

To enhance the performance of the medical case–based retrieval task, several fusion strategies were implemented and compared in García et al. (García

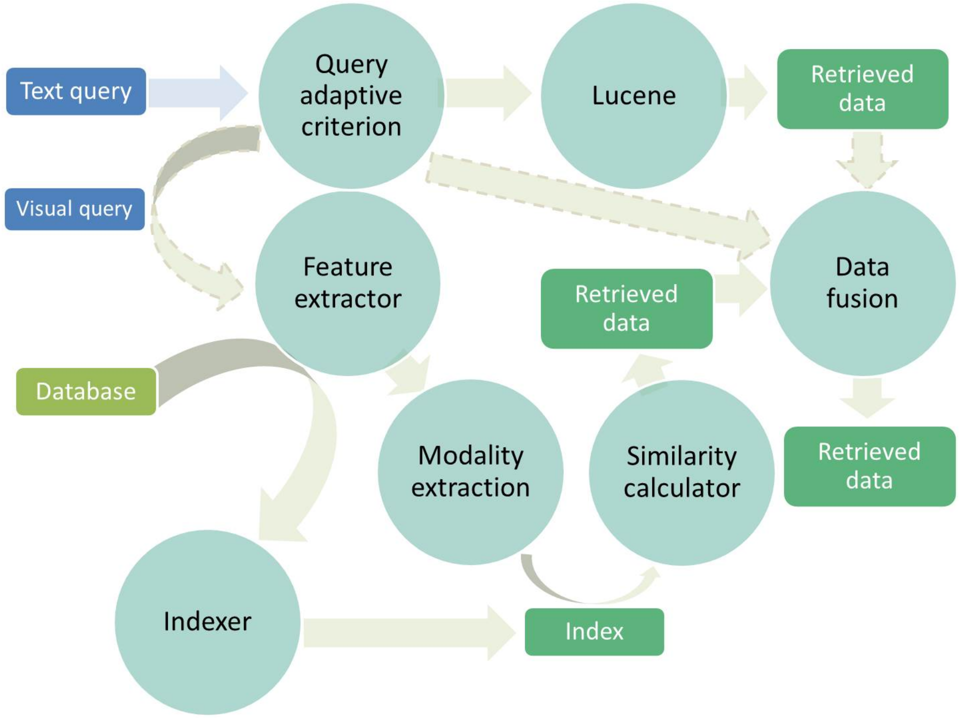


Figure 4: Outline of the multi–modal retrieval including a query–adaptive multi–modal fusion criterion and a modality classification filter.

Seco de Herrera, Schaer, Markonis, & Müller, 2015). The fusion rules in this work are selected based on (García Seco de Herrera, Schaer, et al., 2015). Visual queries are fused with the maximum combination (combMAX) and the visual descriptors for a single query with Borda. A linear combination of text and visual search is applied basing the weights for the linear combination on the MAP obtained in 2011 by the best run.

The rest of this section describes the main retrieval components studied in this work in detail to improve the results obtained with the baseline approach. Figure 4 shows an overview of all the components presented.

Query–adaptive multi–modal fusion

A method for query–adaptive multi–modal fusion is proposed in (García Seco de Herrera, Foncubierta-Rodríguez, & Müller, 2015). The goal is to change the formulation of the retrieval algorithm based on the user query.

For this, Medical Subject Headings (MeSH) terms extracted from the text query are analysed in order to determine the potential use of visual queries as a complementary source. To predict when it is suitable to use visual information in addition to text based on the query, the following criterion is defined:

Query–adaptive fusion criterion Let $\vec{q} \in [0,1]^M$ be the binary histogram of MeSH term occurrences in the textual query. If $\exists i | \vec{q}(i) \neq 0$ and $\exists j | \bar{\mathbf{S}}(i,j) \neq 0$ then the textual query is *suitable* to be fused with a visual query. Where the matrix $\bar{\mathbf{S}}$ shows the synonym relation between text (MeSH terms) and visual features.

Modality classification

A modality classification approach was presented in (García Seco de Herrera, Markonis, et al., 2015). The proposed method uses multi–modal information for the representation of the images. A K–Nearest Neighbour (k–NN) classifier using weighted voting is applied for the image classification. Previous work (García Seco de Herrera, Markonis, et al., 2015) shows that the k–NN algorithm is stable across k choices. In this work $k = 6$ is used based on past results on a different data set.

The text representation of the images with the Lucene search engine is based on the captions of the images.

A set of low–level visual descriptors is selected from the descriptor bank of ParaDISE (Schaer et al., 2014) and their combination is explored (García Seco de Herrera, Markonis, Schaer, et al., 2013). The following descriptors are selected for the modality classification task: CEDD, BoVW, BoC, FCTH and FCH (described also above). In addition to the descriptors defined previously, the following descriptors are used:

- *Fuzzy Colour and Texture Histogram (FCTH)* (Chatzichristofis & Boutalis, 2008b) — this descriptor contains results from the combination of 3 fuzzy systems including colour and texture information in a 192 bin histogram;

- *Fuzzy Colour Histogram(FCH)* (Han & Ma, 2002) — the colour similarity of each pixel's colour associated with all the histogram bins through a fuzzy–set membership function is used.

The approach is trained using an expanded training set of the modality classification task of ImageCLEFmed 2013.

Once the image type information is extracted, the predicted types can be integrated into the search results (for example using filtering) to generate a final result list.

The full image dataset is classified into broad modality types (diagnostic, general or compound figure). The query images of each query topic are also classified and a set of query modalities is produced. Images retrieved in the retrieval step are then *filtered*. When *filtering*, images that are not classified into one of the query modalities are discarded from the results lists.

Results

The data and evaluation scenario used in this section is the ImageCLEFmed 2013 benchmark. This work evaluates the medical case–based retrieval task.

Query Topic Analysis

Query topics are essential for the information retrieval (IR) experiments despite being the most critical element of a collection (Sparck Jones, 1995). Although the ImageCLEFmed 2013 query topics were carefully elaborated, differences between the query topics have implications for the performance. Mandl et al. (Mandl & Womser-Hacker, 2008) assess that the variation between query topics is larger than the variations between systems in most of the evaluation activities. However, ImageCLEFmed 2013 reports system effectiveness as an average over the set of query topics. Table 1 shows the number of documents judged as relevant in the database for each of the query topics. In total, there are only 709 documents judged as relevant for the 35 queries, varying from 1 to 100 relevant articles per query topics. This sparseness and the high differences across topics complicate the task. The ImageCLEFmed 2013 query topics were not proposed by the assessors who judge the documents in the pool (subset of articles that were manually judged for relevance) resulting in few documents considered relevant (Banks, Over, & Zhang, 1999). In addition, most of the submitted runs used only text techniques, only 5 runs were submitted using purely visual techniques. Therefore, most of the documents in the pools that were judged for relevance were retrieved by systems that used only text techniques and are not based on the actual images belonging to the documents. If a run contains relevant articles that were not judged previously its performance has potentially a negative bias (Clough & Sanderson, 2013) in terms of the results. This means that visual retrieval techniques might have a slight disadvantage,

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

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

as more non–judged documents could be retrieved. In the following paragraphs, the query topics are analysed in detail from a visual point of view to better understand the problem before performing further experiments.

Figure 5 shows the Average Precision (AP) per query topic achieved by the run with best MAP submitted to ImageCLEFmed 2013. It is notable that around a third of the query topics obtained an AP of zero. Analysing these query topics in detail, it can be observed that the query images can basically not retrieve the images belonging to the articles judged as relevant based on visual similarity because the images in the relevant articles are visually fairly different from the query images. Figure 6 shows one of these query topics where both query images are visually not very similar to the images of the articles judged as relevant for this query topic. Therefore, no system is likely able to retrieve these articles based only on visual information. In fact, visual information in a multi–modal approach does in this case not contribute to improve the retrieval for these query topics. Many articles contain only graphs, which are not images that are discriminative for a visual search (see Figure 7).

In only two of the analysed query topics there is a single image in the articles judged relevant that could be visually similar to the query images. However, these images are subfigures of a compound image that are not easily accessible for visual analysis. One example is shown in Figure 8 where each of the two query images is visually similar to subfigures of one image in an article judged as relevant.

Relevance judgements in the medical domain can be cognitively demand-

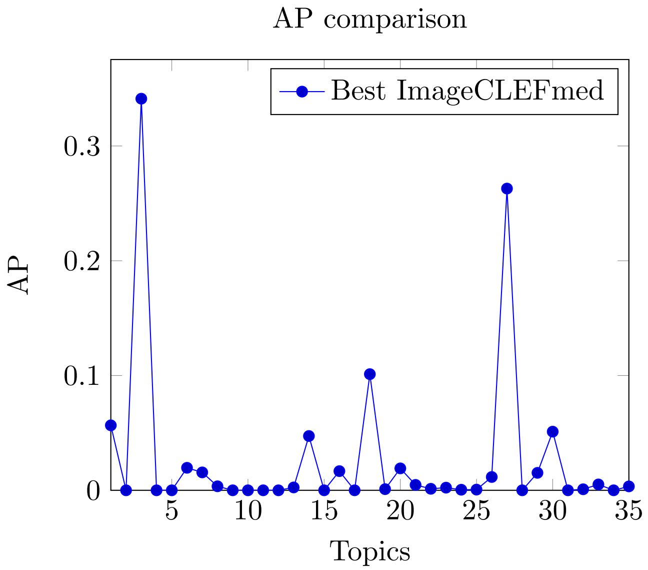
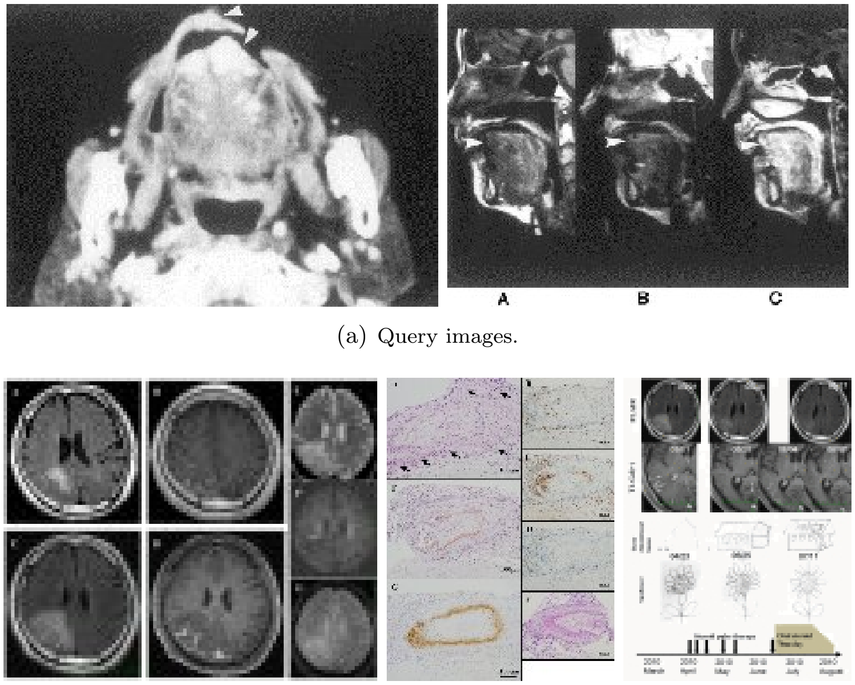


Figure 5: AP for individual query topics achieved by the best visual run submitted to ImageCLEFmed 2013.

ing (Koopman & Zuccon, 2014). In this case, the articles were asked to be relevant if they were useful for a differential diagnosis. After this detailed analysis, it seems that the assessors probably based their decisions mainly on the textual information of the articles and that the visual image content played a less important role. Despite the limitation of the evaluation framework, it provides a good scenario to compare the proposed approaches with the state–of–the–art and with the presented strong baseline. It also gives ideas to maybe change the evaluation scenario slightly in future editions to better evaluate the visual aspects.

Medical case–based retrieval method

The 1,000 highest–ranked articles are retrieved for each query topic in the following experiments. Results are averaged over the total number of queries (35) in order to reproduce the exact setup of ImageCLEFmed 2013. In every experiment, results are compared with the best runs (per type of task) submitted by participants of the ImageCLEFmed 2013 challenge.



(a) Query images.

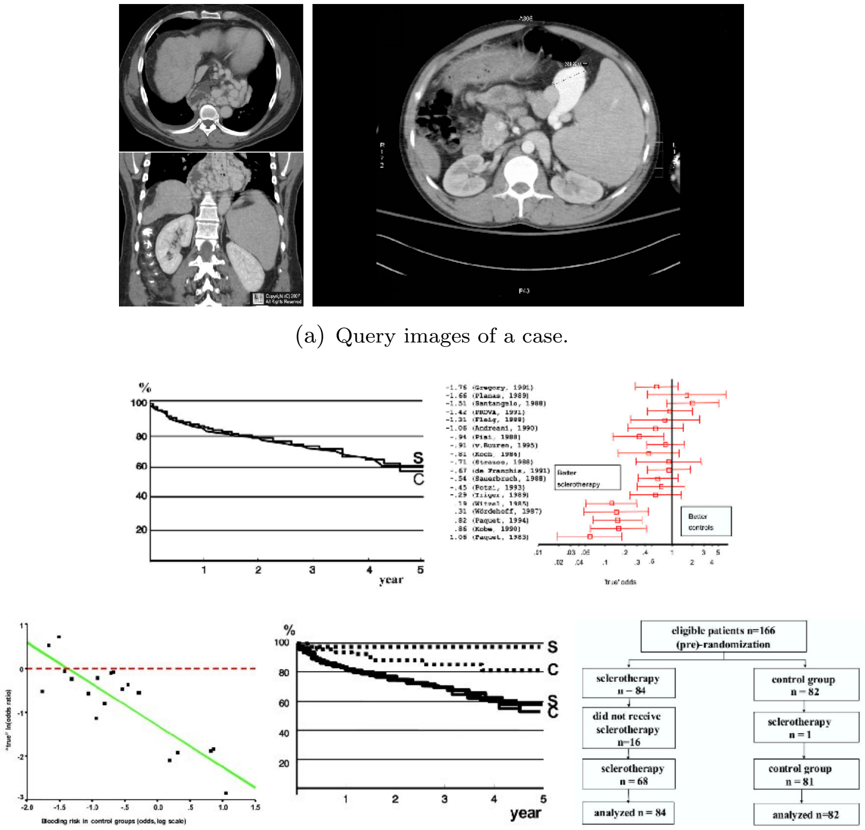
(b) Relevant images.

Figure 6: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to the articles judged as relevant for that query topic. It can be noticed that they are visually not very similar.

Basic performance Table 2 shows results achieved by the text baseline (*RunTB*) and the best runs submitted to ImageCLEFmed 2013 . The presented baseline achieved good results although not as good as the best run submitted in the competition. The best textual approach in the ImageCLEF competition used an external corpus for robust and effective expansion term inference (Sungbin, Lee, & Cho, 2013). The technique described here focuses on visual retrieval and the text baseline is only used to test experimental multi–modal approaches. No such external corpus was used in the technique described here.

Table 3 shows the results achieved on the ImageCLEFmed 2013 case–based task using only visual information.

In CBIR, the selected descriptors are used as a visual baseline for the following experiments. The proposed visual baseline, *RunVB*, performs better than the best visual run submitted in the task in 2013 except at the



(a) Query images of a case.

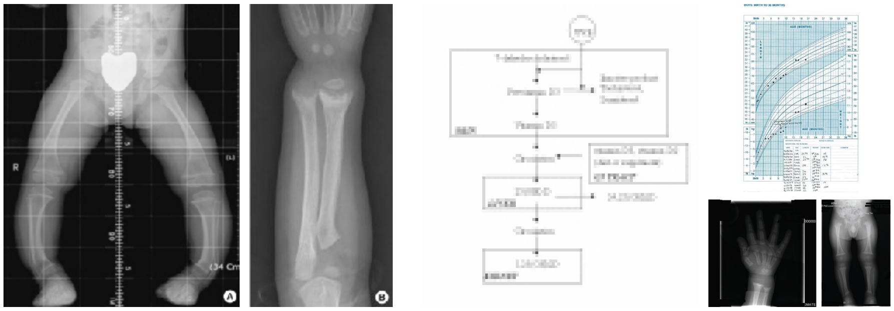
(b) Images of relevant articles.

Figure 7: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to articles judged as relevant for that query topic. All the images in the article are graphs, meaning that they are not discriminative for visual processing.

precision at 10 and 30. The high increase in bpref means that likely more unjudged documents were found by the presented visual run and potentially some of these could be relevant as well, which is not taken into account by the other measures.

Text and visual runs (*RunTB* and *RunVB*) are combined into a single result. The presented baseline approach (*RunMB*) is better than the best mixed (multi–modal) run submitted to ImageCLEFmed 2013.

Table 4 presents the results obtained on the ImageCLEFmed 2013 collection when using the multi–modal approaches included in this article.



(a) Query images of a case.

(b) Images of relevant articles..

Figure 8: Example of (a) query images from a query topic from ImageCLEFmed 2013 and (b) images belonging to the articles judged as relevant for that query topics. There are images in the relevant articles visually similar to the query but as subfigures of a compound figure, which makes visual retrieval very difficult.

Table 2: Results of the approaches of the medical case–based retrieval task when using only text on the ImageCLEFmed 2013 collection.

Run ID	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best text ImageCLEF	**0.2429**	**0.1163**	**0.2417**	**0.2657**	**0.1981**
RunTB	0.1791	0.1107	0.1630	0.2143	0.1581

Query–adaptive multi–modal fusion The synonym matrix of a set of MeSH terms and each visual descriptor is calculated based on a training set of 5,000 random images from the ImageCLEFmed 2013 database that are not part of the test and training sets used here. The choice of the latent value and the percentile is studied in García et al. (García Seco de Herrera, Foncubierta-Rodríguez, & Müller, 2015).

The query–adaptive criterion presented allows the automatic selection of the *text* and *mixed* approach for each of the query topics. The accuracy of correct decisions obtained when applying the proposed approach is 77.15%. For 60% of the query topics, CBIR is not used at all. Using the text approach the correct decision was taken in 54.29% of the cases. Therefore, the proposed criterion prevents the unnecessary use of visual information

Table 3: Performance of the visual baseline approach on the case–based task of ImageCLEFmed 2013 compared with the best visual run submitted in the competition. RunVB is the visual baseline;

Run ID	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best visual ImageCLEF	0.0281	0.0009	0.0335	**0.0429**	**0.0238**
RunVB	**0.0336**	**0.0013**	**0.0666**	0.0343	0.0229

Table 4: Performance of the mixed baseline on the case–based task of ImageCLEFmed 2013 compared with the best mixed results obtained in the task. RunMB is the multi–modal baseline; RunMQ includes the query–adaptive multi–modal fusion; RunMM includes the modality filter; and RunMF is the final multi–modal approach.

Run ID	**MAP**	**GMAP**	**Bpref**	**P10**	**P30**
Best mix ImageCLEF	0.1608	0.0779	0.1426	0.1800	0.1257
RunMB	0.1889	0.1190	0.1720	0.2257	0.1629
RunMQ	0.1885	0.1191	0.1726	**0.2286**	0.1600
RunMM	**0.1904**	**0.1208**	**0.1732**	0.2257	**0.1638**
RunMF	**0.1904**	**0.1208**	**0.1732**	0.2257	**0.1638**

making the system more efficient.

Modality classification for retrieval Table 4 also shows that image modality filtering (*RunMM*) achieves slightly better results than without applying any modality filtering. Moreover, *RunMM* outperforms the best multi–modal approaches in the ImageCLEFmed 2013 contest.

Final combined approach The final approach is executed by combining the techniques studied in this article. The combination of the following steps is applied:

- combination of multiple visual features;

- optimal multi–modal (visual and text information) fusion;
- query–adaptive multi–modal fusion;
- image modality information filtering.

Table 4 shows how the proposed combined approach outperforms the best multi–modal approach submitted to ImageCLEFmed 2013. The combination of the techniques suggested in this work also enabled a more efficient search, limiting the use of visual information only to suitable cases, and reducing the search space thanks to the modality filtering. These aspects are added to the very good results that are better then the best runs in ImageCLEF 2013.

Web–based interface

This section presents a novel web–based retrieval interface, called *Shangri–La*. The goal of the interface is to provide a front–end with which the user can interact and control the underlying medical case–based retrieval system (Brajnik, Mizzaro, & Tasso, 1996). The web–based interface highlights the opportunities and challenges given by the Internet to easily share the developed system. Shangri–La provides multi–modal retrieval functionalities that allow the user to find relevant articles querying the system with a case description consisting of free text and/or visual examples. Currently, the ImageCLEFmed 2013 dataset is accessible from Shangri–La and supported by the proposed system. However, it can easily be extended to other datasets and it can be optimised to specific domains as well, where particularly the visual results can be optimized strongly.

A complete version of the proposed interface is available for testing at the following address: `http://shangrila.khresmoi.eu/`. This system is general and not optimised for any application domain. Machine learning can be applied to optimised it for specific disease areas.

System architecture

The techniques previously described are integrated into the ParaDISE system. Shangri–La is developed as a client–side only application, based entirely on HTML5 and JavaScript.

The interface accesses several web services that use a REST style architecture (Schaer et al., 2014). The used web services are described below (see Figure 9):

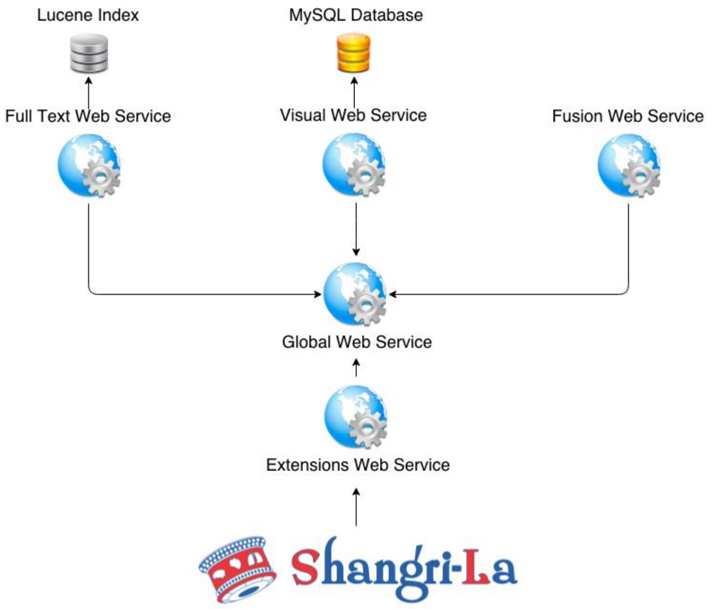


Figure 9: Service layer architecture for the medical case–based retrieval system.

- *Full text web service* — responsible for searching articles in the dataset;
- *Visual web service* — responsible for CBIR;
- *Fusion web service* — responsible for combining results from different sources;
- *Global web service* — facade for client applications, calling the individual web services in succession;
- *Extensions web service* – responsible for all tools that are added to the ParaDISE system. In particular, it is responsible for medical case–based retrieval, dealing with most of the techniques developed.

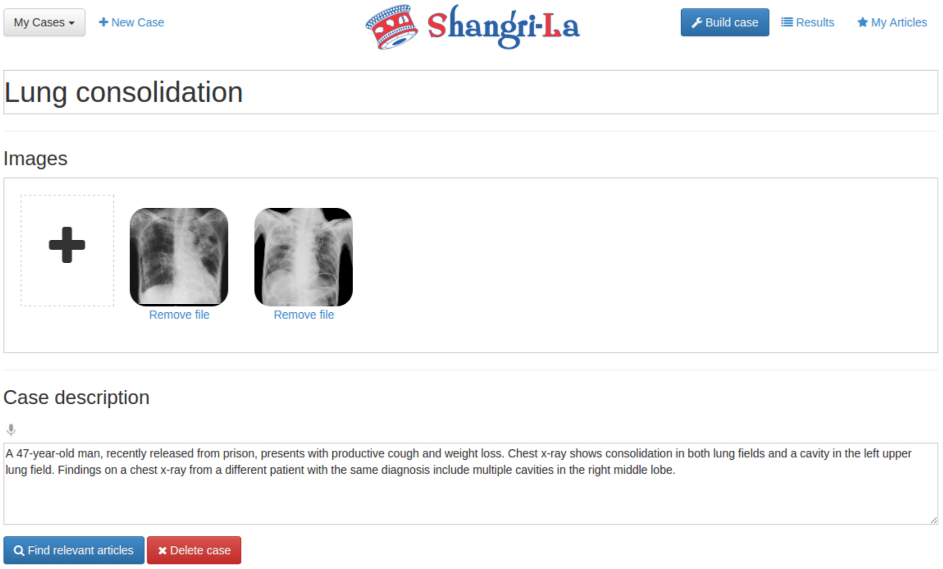
All interactions with the ParaDISE system (that can be hosted on a completely different server, as it is totally independent) use AJAX (Asynchronous JavaScript and XML) to call the extensions web service.

Interface functionality

Shangri–La enables users to interact with the medical case–based retrieval system with a low amount of user effort. The interface hides the complexity of the system implementation, giving users a simple website to collect the desired information. To keep the interaction clear and concise, Shangri–La provides the following three main pages:

- *Build Case* — to formulate a user query;
- *Results* — to provide all of the information needed to support the user's request;
- *My Articles* — to display all of the articles selected by the user.

Links to these three pages are always present to allow the users to return to them easily. The following sections detail each of the three pages contained in Shangri–La.



Lung consolidation

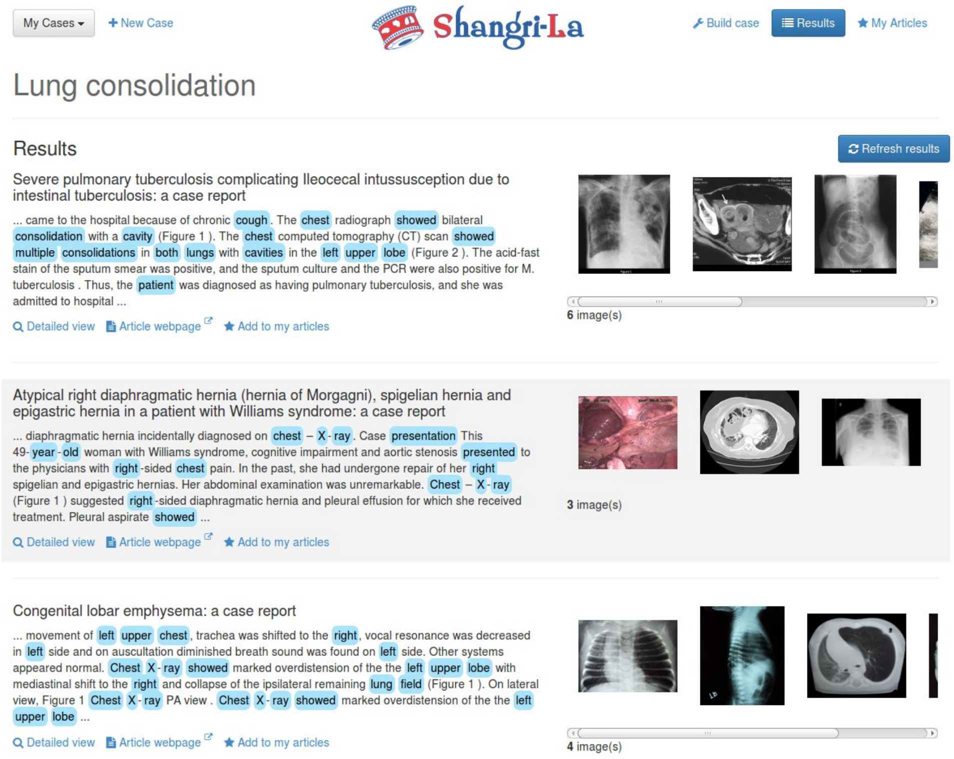
Case description

A 47-year-old man, recently released from prison, presents with productive cough and weight loss. Chest x-ray shows consolidation in both lung fields and a cavity in the left upper lung field. Findings on a chest x-ray from a different patient with the same diagnosis include multiple cavities in the right middle lobe.

Figure 10: Screenshot of the *Build Case* page from the Shangri–La interface. This page shows an example of a simple query.

Build Case page

The goal of the *Build Case* page is to simply and easily capture the user's information need. The medical case–based retrieval system was developed to support inputs including a free text case description and image examples. Query images can be uploaded from a storage device using a file browser dialogue selection method. In addition, drag and drop facilitates query image uploading. A text area is used to enter text that can contain multiple lines of textual information for long case descriptions. Shangri–La also supports real–time speech recognition that transcribes a spoken query into text using the Google Chrome Speech API. However, in the current version many phrases or words are not perfectly well recognised, as the system is not optimised for the medical field. An example of a build case query is shown in Figure 10.



Lung consolidation

Results

Severe pulmonary tuberculosis complicating Ileocecal intussusception due to intestinal tuberculosis: a case report

... came to the hospital because of chronic cough . The chest radiograph showed bilateral consolidation with a cavity (Figure 1). The chest computed tomography (CT) scan showed multiple consolidations in both lungs with cavities in the left upper lobe (Figure 2). The acid-fast stain of the sputum smear was positive, and the sputum culture and the PCR were also positive for M. tuberculosis . Thus, the patient was diagnosed as having pulmonary tuberculosis, and she was admitted to hospital ...

Atypical right diaphragmatic hernia (hernia of Morgagni), spigelian hernia and epigastric hernia in a patient with Williams syndrome: a case report

... diaphragmatic hernia incidentally diagnosed on chest – X - ray . Case presentation This 49- year - old woman with Williams syndrome, cognitive impairment and aortic stenosis presented to the physicians with right -sided chest pain. In the past, she had undergone repair of her right spigelian and epigastric hernias. Her abdominal examination was unremarkable. Chest – X - ray (Figure 1) suggested right -sided diaphragmatic hernia and pleural effusion for which she received treatment. Pleural aspirate showed ...

Congenital lobar emphysema: a case report

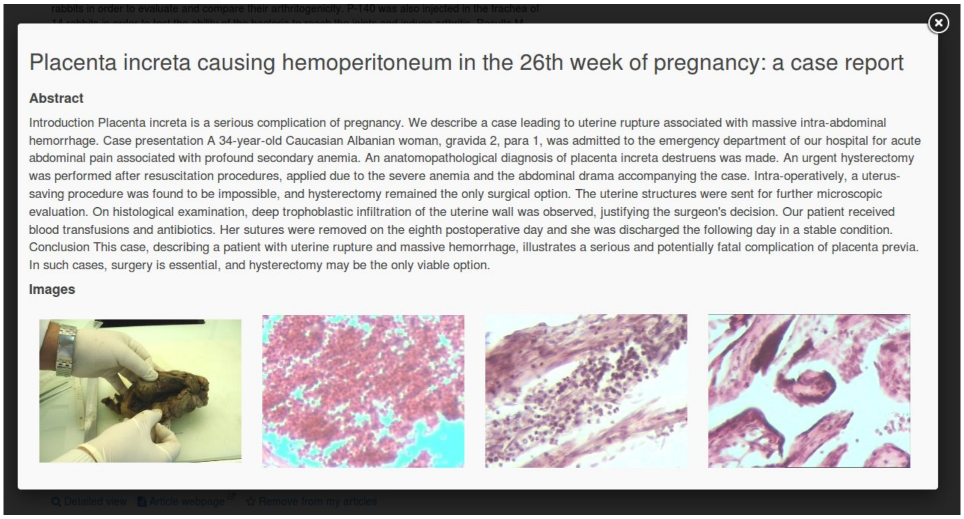
... movement of left upper chest , trachea was shifted to the right , vocal resonance was decreased in left side and on auscultation diminished breath sound was found on left side. Other systems appeared normal. Chest X - ray showed marked overdistension of the the left upper lobe with mediastinal shift to the right and collapse of the ipsilateral remaining lung field (Figure 1). On lateral view, Figure 1 Chest X - ray PA view . Chest X - ray showed marked overdistension of the the left upper lobe ...

Figure 11: Screenshot of the *Results* page from the Shangri–La interface. This page shows a ranked list of articles resulting from a search query.

Results page

The goal of the *Results* page is to provide all of the information needed to support the medical case–based retrieval task. A user study (Drori, 2000) shows that users prefer retrieved results that display also the lines in the document that fulfil the search condition and not the first lines of the document. Therefore, Shangri–La displays the resulting articles of the search in a ranked list, with basic information containing: the title, relevant lines that fulfil the search criteria (such as keywords) taken from the body of the article, and images (if available). In addition, terms contained in the text query are highlighted. An example of an outcome displayed in the *Results* page is shown in Figure 11.

Furthermore, the interface provides a link to the corresponding article as well as a bookmark option. A detailed view of the article is possible without following a link to the original source. It includes its title, abstract and images contained in the article. The user can click on the images contained in a retrieved article to see a larger view (see Figure 12).


rabbits in order to evaluate and compare their arthritogenicity. P-140 was also injected in the trachea of
Detailed view
Article webpage
Remove from my articles

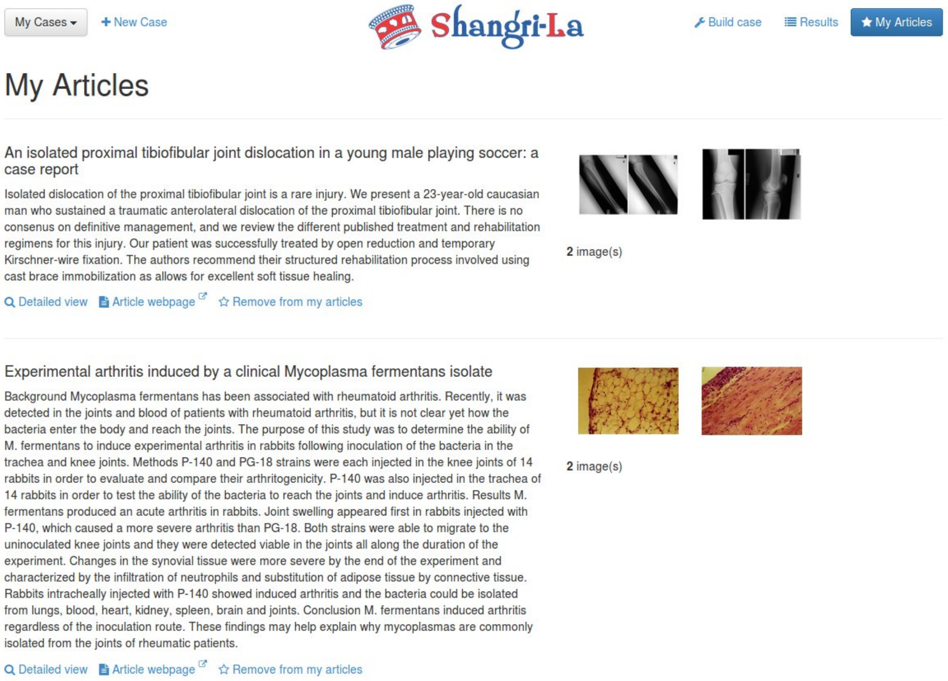
Placenta increta causing hemoperitoneum in the 26th week of pregnancy: a case report

Abstract

Introduction Placenta increta is a serious complication of pregnancy. We describe a case leading to uterine rupture associated with massive intra-abdominal hemorrhage. Case presentation A 34-year-old Caucasian Albanian woman, gravida 2, para 1, was admitted to the emergency department of our hospital for acute abdominal pain associated with profound secondary anemia. An anatomopathological diagnosis of placenta increta destruens was made. An urgent hysterectomy was performed after resuscitation procedures, applied due to the severe anemia and the abdominal drama accompanying the case. Intra-operatively, a uterus-saving procedure was found to be impossible, and hysterectomy remained the only surgical option. The uterine structures were sent for further microscopic evaluation. On histological examination, deep trophoblastic infiltration of the uterine wall was observed, justifying the surgeon's decision. Our patient received blood transfusions and antibiotics. Her sutures were removed on the eighth postoperative day and she was discharged the following day in a stable condition. Conclusion This case, describing a patient with uterine rupture and massive hemorrhage, illustrates a serious and potentially fatal complication of placenta previa. In such cases, surgery is essential, and hysterectomy may be the only viable option.

Images

Figure 12: Screenshot of the *detailed view* of an article from the Shangri–La interface. This page shows the title, abstract and images from a selected article.



My Articles

An isolated proximal tibiofibular joint dislocation in a young male playing soccer: a case report

Isolated dislocation of the proximal tibiofibular joint is a rare injury. We present a 23-year-old caucasian man who sustained a traumatic anterolateral dislocation of the proximal tibiofibular joint. There is no consensus on definitive management, and we review the different published treatment and rehabilitation regimens for this injury. Our patient was successfully treated by open reduction and temporary Kirschner-wire fixation. The authors recommend their structured rehabilitation process involved using cast brace immobilization as allows for excellent soft tissue healing.

Experimental arthritis induced by a clinical Mycoplasma fermentans isolate

Background Mycoplasma fermentans has been associated with rheumatoid arthritis. Recently, it was detected in the joints and blood of patients with rheumatoid arthritis, but it is not clear yet how the bacteria enter the body and reach the joints. The purpose of this study was to determine the ability of M. fermentans to induce experimental arthritis in rabbits following inoculation of the bacteria in the trachea and knee joints. Methods P-140 and PG-18 strains were each injected in the knee joints of 14 rabbits in order to evaluate and compare their arthritogenicity. P-140 was also injected in the trachea of 14 rabbits in order to test the ability of the bacteria to reach the joints and induce arthritis. Results M. fermentans produced an acute arthritis in rabbits. Joint swelling appeared first in rabbits injected with P-140, which caused a more severe arthritis than PG-18. Both strains were able to migrate to the uninoculated knee joints and they were detected viable in the joints all along the duration of the experiment. Changes in the synovial tissue were more severe by the end of the experiment and characterized by the infiltration of neutrophils and substitution of adipose tissue by connective tissue. Rabbits intracheally injected with P-140 showed induced arthritis and the bacteria could be isolated from lungs, blood, heart, kidney, spleen, brain and joints. Conclusion M. fermentans induced arthritis regardless of the inoculation route. These findings may help explain why mycoplasmas are commonly isolated from the joints of rheumatic patients.

Figure 13: Screenshot of the *My Articles* page from the Shangri–La interface. This page shows a selection of articles that the user added to revisit easily.

My Articles page

Bookmarks, also referred to as favourites or hotlist, are a common tool to simplify visiting the same pages at a later moment. A recent survey found that 97% of the users are aware of the bookmark function and 85% regularly save web pages using this method (Shen & Prior, 2013).

Shangri–La allows the user to bookmark articles from the results list of a case search. The *My Articles* page displays the favourite articles that are added by the user. An overview of the selected articles is shown in the same format as in the *Results* page. The user can interact with the article selection from this page. The page allows checking the overview information display, visiting the article or even deleting articles from the list. Figure 13 shows a view of the actual *My articles* page for a selection of articles.

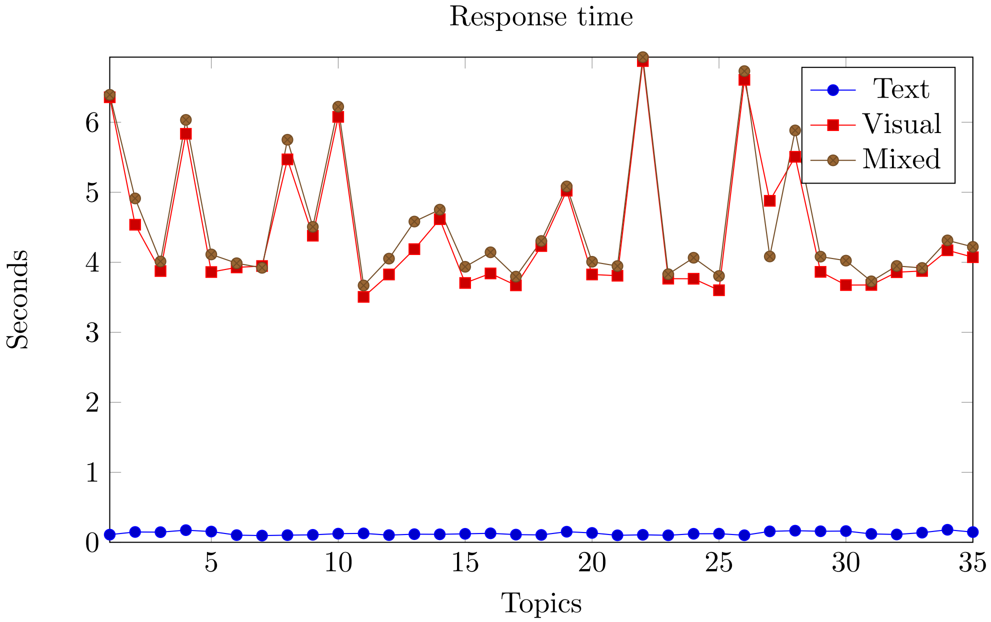


Figure 14: Time in seconds that Shangri–La takes to perform the ImageCLEF queries with text and visual input depending on the query type (visual, text, mixed).

Analysis of the response times

This section analyses the response time of the query system for the same set of 35 queries that were used for the performance evaluation. The response time is not easy to evaluate because it depends on other parameters such as the CPU speed, disk speed and the system workload (Cacheda & Viña, 2002).

For the Shangri–La demo a server with the following features is used:

- 2 x Intel(R) Xeon(R) CPU E7- 4820 @ 2.00GHz processors (64 cores in total);
- 128GB main memory;
- Gigabit Ethernet network connection;
- 550GB Solid State Drive storage.

All 35 query topics provided by the case–based ImageCLEFmed 2013 retrieval task were used for the evaluation. Queries were executed using only the text, only the images in the query or both combined. Each query was run 10 times to limit the influence of other tasks running on the same server. The server is not dedicated for the demo and all group web demos run on the same server, so a dedicated server could be optimized for faster response times. Figure 14 shows the response time averaged over the 10 repetitions for all 35 query topics consisting of text and 2–3 example images.

The response time also depends on the length of the text describing the case and the number of images in each of the query topics (see Table5). The time response needed for text queries is almost instantaneous (an average of 127 ms) and the query length has a limited influence. Visual retrieval scales pretty much linearly with the number of images in a query at a rate of around 2 seconds per query. Mixed queries are dominated by the visual part and are pretty much the same 2 seconds per image.

Table 5: Number of characters, words and images per query topic in the case–based ImageCLEFmed 2013 task.

Topic number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
N. characters	154	311	269	236	417	150	137	194	244
N. words	25	50	49	35	68	20	21	34	34
N. images	3	2	2	2	2	2	2	3	2
Topic number	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
N. characters	330	288	189	215	239	220	300	246	143
N. words	49	45	29	33	35	37	48	38	24
N. images	3	2	2	2	2	2	2	2	2
Topic number	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
N. characters	386	239	254	200	180	253	280	150	319
N. words	64	35	35	33	26	41	43	24	52
N. images	2	2	2	3	2	2	2	3	2
Topic number	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	
N. characters	319	381	355	239	161	349	394	310	
N. words	54	59	58	37	25	63	58	48	
N. images	3	2	2	2	2	2	2	2	

Discussion

The objective of the presented case–based retrieval system is to retrieve articles from the biomedical literature that potentially help clinicians in the decision making. In this work, the main aspect is to bring the visual information available in the medical cases into a retrieval system in addition to the text that already works well. Due to the nature of the data collected in the ImageCLEFmed 2013 campaign, it is difficult to evaluate and to show the improvements that visual information brings to solving the information needs, as most of the tasks are rather text–oriented.

Despite the low performance of the visual search alone, the multi–modal approach is improving the text search and offers complementary functionalities. Results also outperform the best multi–modal runs submitted to ImageCLEFmed 2013 by a weighted linear combination of visual and text retrieval. It demonstrates the effectiveness of the proposed multi–modal baseline framework. The very good retrieval performance combining several techniques in addition to the novel interface and retrieval architecture are the main novelty of this article.

A major challenge is currently the low performance of the visual retrieval. To overcome this, a query–adaptive fusion criterion for the use of multi–modal techniques in medical case–based retrieval is presented. The textual information of MeSH terms is integrated with the visual descriptors creating a matrix of synonym relations between both kinds of features (text and visual). The synonym matrix is then used to decide if a text query is suitable for a multi–modal approach or if text alone would lead to best results. Experimental results indicate that it is indeed effective, showing that correct decisions are taken in 77.15% of the cases. Moreover, by facilitating decision–making the criterion avoids the unnecessary use of visual information. It makes the retrieval system more efficient, as the main response time is related to the processing of the images.

Image classification is applied to enhance the quality of the retrieval system. Modality classification is important in medical image retrieval systems, both for overall retrieval quality and because it is a functionality requested by users. A medical image modality filter is therefore presented to filter out non–relevant images, which has the possibility to remove noise from the results. Image modality filtering improves the performance of simple visual retrieval and multi–modal retrieval. For each query image, descriptors are extracted and compared with the image descriptors stored in the database. Therefore, image filtering reduces the search space focusing the search only on the modalities occurring in the query topic.

The final approach is obtained from a combination of the studied procedures. It outperforms the best multi–modal approach submitted to ImageCLEFmed 2013. Moreover, it improves the effectiveness of the retrieval system by using CBIR only when it is appropriate for the query; and by reducing the search space through a modality filter.

Finally, Shangri–La, a web–based retrieval interface, is implemented to integrate the multi–modal medical case–based retrieval approach proposed in this work.

Conclusions

This work relates the different factors that led to performance improvements of a multimodal medical case–based retrieval system that retrieves biomedical articles with medical cases as queries to find similar cases. After the presentation of a baseline, ImageCLEFmed 2013 query topics are analysed to better understand the task showing that improving retrieval using visual retrieval is limited using the provided ImageCLEFmed 2013 data.

All the techniques presented are implemented for the ImageCLEFmed 2013 database and studied in this work. These techniques are then combined to define a final approach. The final multi–modal approach outperforms the best multi–modal approach submitted to ImageCLEFmed 2013. Moreover, the query–adaptive criterion and the modality filtering contribute to improve the effectiveness of the retrieval system.

Finally, to facilitate the interaction between a user and the medical case–based retrieval system developed in this work, a web–based interface, called Shangri–La, is presented. Such a simple web interface makes it easy for users to work with he system and interact with multimodal data. The interface is based on two iterations of user tests and to fully validate the new interface such user tests could also help in the future. The current data set is also limited in size and there is a larger set of articles available via PubMed Central that can be used in the future.

References

Akgül, C., Rubin, D., Napel, S., Beaulieu, C., Greenspan, H., & Acar, B. (2011). Content–based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, *24*(2), 208–222.

Banks, D., Over, P., & Zhang, N.-F. (1999). Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, *1*(1–2), 7–34.

Boughorbel, S., Tarel, J.-P., & Boujemaa, N. (2005). Generalized histogram intersection kernel for image recognition. In *IEEE international conference on image processing* (Vol. 3, pp. III–161).

Brajnik, G., Mizzaro, S., & Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: A case study on user support. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 128–136).

Cacheda, F., & Viña, A. (2002). Performance evaluation of web information retrieval systems and its application to e–business. In *Challenges and achievements in e-business and e-work contents: International conference* (Vol. 1).

Chakravarti, R., & Meng, X. (2009). A study of color histogram based image retrieval. In *Sixth international conference on information technology:new generations ITNG* (pp. 1323–1328).

Chatzichristofis, S. A., & Boutalis, Y. S. (2008a). CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Lecture notes in computer sciences* (Vol. 5008, pp. 312–322).

Chatzichristofis, S. A., & Boutalis, Y. S. (2008b). 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* (pp. 191–196).

Clough, P., & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, *18*(2).

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). 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* (pp. 758–759). New York, NY, USA: ACM.

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

Depeursinge, A., Duc, S., Eggel, I., & Müller, H. (2012, January). Mobile medical visual information retrieval. *IEEE Transactions on Information Technology in BioMedicine*, *16*(1), 53–61.

Drori, O. (2000). *Improving display of search results in information retrieval* *systems–users' study* (Tech. Rep.). Center for Research in Computer Science of the Leibniz.

García Seco de Herrera, A., Foncubierta-Rodríguez, A., & Müller, H. (2015). Medical case–based retrieval: Integrating query MeSH terms for query–adaptive multi–modal fusion. In *SPIE medical imaging.*

García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., & Müller, H. (2013, September). Overview of the ImageCLEF 2013 medical tasks. In *Working notes of CLEF 2013 (cross language evaluation forum).*

García Seco de Herrera, A., Markonis, D., Eggel, I., & Müller, H. (2012). The medGIFT group in ImageCLEFmed 2012. In *Working notes of CLEF 2012.*

García Seco de Herrera, A., Markonis, D., Joyseeree, R., , Schaer, R., Foncubierta-Rodríguez, A., & Müller, H. (2015). Using semi–supervised learning for image modality classification. In *Multimodal retrieval in the medical domain (MRMD) 2015.* Springer.

García Seco de Herrera, A., Markonis, D., & Müller, H. (2013, October). Bag of colors for biomedical document image classification. In H. Greenspan & H. Müller (Eds.), *Medical content–based retrieval for clinical decision support* (pp. 110–121). Lecture Notes in Computer Sciences (LNCS).

García Seco de Herrera, A., Markonis, D., Schaer, R., Eggel, I., & Müller, H. (2013, September). The medGIFT group in ImageCLEFmed 2013. In *Working notes of CLEF 2013 (cross language evaluation forum).*

García Seco de Herrera, A., Schaer, R., Markonis, D., & Müller, H. (2015). Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task. *Computerized Medical Imaging and Graphics*, *39*, 46–54.

Gkoufas, Y., Morou, A., & Kalamboukis, T. (2011). Combining textual and visual information for image retrieval in the medical domain. *The Open Medical Informatics Journal*, *5*, 50–57.

Gottlieb, K., & Marino, G. (2014). *Diagnostic endosonography: A case–based approach.* Springer Berlin Heidelberg.

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

Hersh, W., Jensen, J., Müller, H., Gorman, P., & Ruch, P. (2005). A qualitative task analysis for developing an image retrieval test collection. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation* (pp. 11–16). Vienna, Austria.

Hersh, W. R., & Hickam, D. H. (1998). How well do physicians use electronic information retrieval systems? *Journal of the American Medical Association*, *280*(15), 1347–1352.

Hoogendam, A., Stalenhoefand, A. F., Robbé, P. d. V. F., & Overbeke, A. J. (2008). Answers to questions posed during daily patient care are more likely to be answered by uptodate than pubmed. *Journal of Medical Internet Research*, *10*(4).

Hunter, L., & K. Cohen, B. (2006, Mar). Biomedical language processing: What's beyond PubMed? *Molecular Cell*, *21*(5), 589–594.

Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., & Müller, H. (2015). Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. *Computerized Medical Imaging and Graphics*, *39*(0), 55 - 61.

Kalpathy-Cramer, J., & Müller, H. (2011). Systematic evaluations and ground truth. In *Biomedical image processing* (pp. 497–520). Springer.

Koopman, B., & Zuccon, G. (2014). Why assessing relevance in medical IR is demanding. In *SIGIR 2014, medical information retrieval (MedIR) workshop.*

Lazebnik, S., Schmid, C., & Ponce, J. (2006). 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* (pp. 2169–2178). Washington, DC, USA: IEEE Computer Society.

Li, Y., Shi, N., & D.Frank, H. (2011). Fusion analysis of information retrieval models on biomedical collections. In *Proceedings of the 14th international conference on information fusion.* IEEE Computer Society.

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

Mandl, T., & Womser-Hacker, C. (2008). Analyzing information retrieval results with a focus on named entities. *Computational Linguistics and Chinese Language Processing*, *13*(1), 121–140.

Markonis, D., Baroz, F., Ruiz de Castaneda, R. L., Boyer, C., & Müller, H. (2013). User tests for assessing a medical image retrieval system: A pilot study. In *Medinfo 2013.*

Markonis, D., Eggel, I., García Seco de Herrera, A., & Müller, H. (2011). The medGIFT group in ImageCLEFmed 2011. In *Working notes of CLEF 2011.*

Markonis, D., Holzer, M., Baroz, F., Ruiz de Castaneda, R. R., Boyer, C., Langs, G., … Depeursinge, A. (2015). User-oriented evaluation of a medical image retrieval system for radiologists. *International journal of medical informatics*.

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

Markonis, D., Schaer, R., García Seco de Herrera, A., & Müller, H. (Submitted). The Parallel Distributed Image Search Engine (ParaDISE). *Multimedia Tools and Applications*.

Mendelson, D. S., & Rubin, D. L. (2013). Imaging informatics: Essential tools for the delivery of imaging services. *Academic radiology*, *20*(10), 1195–1212.

Mourão, A., & Martins, F. (2013). NovaMedSearch: a multimodal search engine for medical case–based retrieval. In *Proceedings of the 10th conference on open research areas in information retrieval* (pp. 223–224).

Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., & Geissbuhler, A. (2006, aug). Health care professionals' image use and search behaviour. In *Proceedings of the medical informatics europe conference (MIE 2006)* (pp. 24–32). Maastricht, The Netherlands.

Müller, H., García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., & Eggel, I. (2012, September). Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In *Working notes of CLEF 2012 (cross language evaluation forum).*

Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., & Antani, S. (2012). Creating a classification of image types in the medical literature for visual categorization. In *SPIE medical imaging.*

Müller, H., Zhou, X., Depeursinge, A., Pitkanen, M., Iavindrasana, J., & Geissbuhler, A. (2007). Medical visual information retrieval: State of the art and challenges ahead. In *2007 IEEE international conference on multimedia and expo* (pp. 683–686). IEEE.

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

Riding the wave: How europe can gain from the rising tide of dcientific data. (2010, October). Submission to the European Comission, available online at `http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf`. Retrieved from `http://cordis.europa`

`.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf`

Schaer, R., Markonis, D., & Müller, H. (2014). Architecture and applications of the Parallel Distributed Image Search Engine (ParaDISE). In *FoRESEE 2014, 1st international workshop on future search engines at informatik 2014.*

Shen, S.-T., & Prior, S. D. (2013). My favorites (bookmarks) schema: One solution to online information storage and retrieval. In *Proceedings of the 2013 international conference on information systems and design of communication* (pp. 33–40).

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Management*, *31*(3), 291–314.

Sungbin, C., Lee, J., & Cho, J. (2013, September). SNUMedinfo at ImageCLEF 2013: Medical retrieval task. In *Working notes of CLEF 2013 (cross language evaluation forum).*

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Tamura, H., Mori, S., & Yamawaki, T. (1978, June). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, *8*(6), 460–473.

Uwimana, E., & Ruiz, M. E. (2008). Integrating an automatic classification method into the medical image retrieval process. In *AMIA annual symposium procreedings* (pp. 747–751).

Voorhees, E. M., & Buckland, L. P. (Eds.). (2006, November). *The fifteenth Text REtrieval Conference (TREC 2006) proceedings* (Vol. Special Publication 500-272). National Institute of Standards and Technology (NIST).

Welter, P., Deserno, T. M., Fischer, B., Günther, R. W., & Spreckelsen, C. (2011). Towards case–based medical learning in radiological decision making using content–based image retrieval. *BMC Medical Informatics and decision Making*, *11*(68).