PhD Exposé

Medical Case Retrieval

Mario Taschwer

Supervisor: Prof. Laszlo Böszörmenyi, AAU

Co-Supervisor: Prof. Oge Marques, FAU, Florida

March, 2013

Alpen-Adria-Universität Klagenfurt (AAU)

Abstract

The proposed PhD project addresses the problem of medical case retrieval (MCR), where a medical case is represented by a multimedia document describing a certain disease or a patient's history. The ImageCLEF evaluation campaign poses a yearly MCR task using a heterogeneous dataset of more than 75,000 medical publications consisting of text and images. The best results achieved by participants of the ImageCLEF MCR task in 2012 are moderate and call for improvement. Interestingly, approaches based on visual retrieval perform significantly worse than text-only retrieval, even if combined with text retrieval. This project therefore aims at designing an MCR model that is able to deliver a substantially better retrieval performance on the ImageCLEF dataset. Moreover, the potential of further improvement by leveraging the feedback of medical expert users for long-term learning will be investigated.

Contents

1 Introduction **3**

2 State of the Art **4**
2.1 Multimedia information retrieval 6
2.1.1 Text Retrieval 6
2.1.2 Content-based Visual Retrieval 6
2.1.3 Data Fusion 9
2.1.4 Relevance Feedback 10
2.2 Knowledge Representation 10

3 Aim and Objectives **11**

4 Previous Work **13**

5 Evaluation **13**

6 Methodology **14**

7 Project Management **15**

References **15**

1 Introduction

Clinical decision support systems provide clinicians with patient-specific assessments or recommendations to aid clinical decision making. Several features of such systems have been shown to improve clinical practice significantly [33]: automatic provision of decision support as part of clinician workflow, provision of recommendations rather than just assessments, provision of decision support at the time and location of decision making, and computer-based decision support. Depending on the degree of decision support expected from a computer system, these features may pose demanding requirements on the effectiveness and efficiency of used technology. Following the paradigm of *evidence-based medicine* [56], clinical decision making needs to integrate the physician's individual clinical expertise and the best available external clinical evidence from systematic research. Computer-based decision support systems may help to provide external evidence, learn from individual expertise, and possibly provide recommendations for diagnosis or treatment.

A well-known approach to designing a decision support system is the method of *case-based reasoning* (CBR), developed by the artificial intelligence research community [1, 5]. Its main objective is to solve a new problem by applying previous experiences adapted to the current situation. For clinical decision support, the problem is represented by a patient's symptoms, and the solution is a decision about diagnosis and treatment. A problem and its solution are called a *case*, and cases are retained in

a case library for subsequent reasoning about new problems. The process of case-based reasoning can be divided into four main tasks [1]: (1) for a given new problem, retrieve similar cases from the case library; (2) reuse the most relevant cases to propose a solution for the new problem; (3) revise the proposed solution to adapt it to the current problem; (4) retain the new case in the case library. Although successful CBR systems for different narrow medical application domains have been built and evaluated on a few hundred cases [5], general methods to design a CBR system applicable to larger and heterogeneous medical datasets still present an open research problem.

This PhD project addresses task (1) of a medical CBR system: given a description of patient symptoms (called a *query*), retrieve the most relevant medical cases from the case library. Both descriptions of symptoms and cases contain text and images, and the case library is not restricted to a particular medical domain. Relevance of cases is ultimately defined by medical experts, but the retrieval system is supposed to implement a relevance model that allows for automatic retrieval. As the CBR process usually involves interaction with a medical expert (typically in task (3)), the retrieval system should also be able to learn from its expert users in order to improve its relevance model.

Medical case retrieval tasks are issued every year by the ImageCLEF evaluation campaign[1] [44] since 2009, allowing researchers to evaluate their systems using a common large dataset. In 2012, the dataset consisted of ca. 75,000 case descriptions, which are in fact publications available in the PubMed Central[2] full-text archive of biomedical literature, together with about 300,000 referenced image files. The retrieval performance achieved by 6 participating research teams in 2012 [45] lets room for improvement (numbers of 2011 task [32] in parentheses, MAP = mean average precision [3]):

- best result achieved by textual retrieval only: MAP $\approx$ 17% (13%);
- best result using combined textual and visual retrieval: MAP $\approx$ 10% (8%);
- best result using visual retrieval only: MAP $\approx$ 3.7% (2.0%).

These results confirm that medical case retrieval on general large datasets is still an open problem that needs further research. Moreover, the results contradict the expectation that fusing retrieval of different feature sets (text and visual features) should improve retrieval performance, giving rise to additional research objectives. In addition to its use in decision support systems, medical case retrieval is also a relevant problem in medical education and research, because it allows to select interesting cases for students and to retrieve datasets for studies meeting case-based criteria.

2 State of the Art

The narrow research field of medical case retrieval (MCR) can be positioned at the intersection of three larger areas of artifical intelligence research:

[1] `http://imageclef.org/`
[2] `http://www.ncbi.nlm.nih.gov/pmc/`

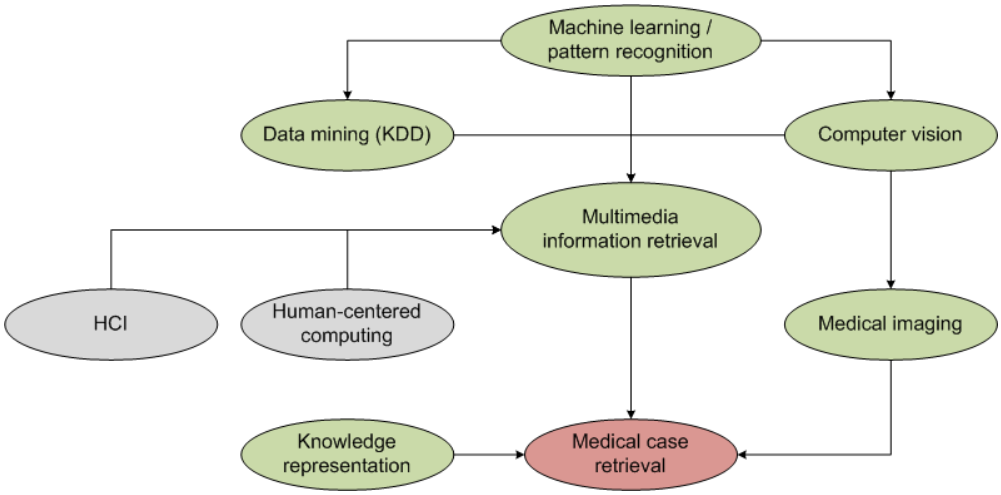

Figure 1: Research fields related to medical case retrieval. KDD = Knowledge discovery in databases, HCI = Human-computer interaction.

- *Multimedia information retrieval:* Indexing and retrieving multimedia documents requires techniques from classical information retrieval, hereafter called *text retrieval*, from content-based image and video retrieval, referred to as *content-based visual retrieval*, and from the *data fusion* literature dealing with the combination of several information retrieval systems or information sources. Utilizing user interactions for improving retrieval results is known as *relevance feedback*.
- *Knowledge representation:* Retrieval of medical multimedia documents seems to exhibit limited performance when relying on information extracted from the document corpus only. So approaches incorporating *external knowledge* into the retrieval process have been proposed, often representing expert knowledge by *medical ontologies*.
- *Computer vision:* When utilizing images for content-based retrieval, computer vision methods are needed to extract discriminative features and detect semantic concepts. For diagnostic images, more specific techniques developed by the *medical imaging* research community may be required, such as image registration or segmentation.

The research fields related to MCR are depicted in Figure 1. Multimedia information retrieval does not only apply computer vision methods, but also techniques from machine learning, pattern recognition, and data mining (in the sense of knowledge discovery in databases). Naturally, multimedia information retrieval, and hence MCR, involves user interaction when deployed in a real world setting. So the research fields of human-computer interaction and human-centered computing play an important role for designing a complete MCR system. However, the focus of this PhD project is on

automatic retrieval and system evaluation methods without or little user interaction, so these two research fields are presently ignored.

The following subsections give an overview of literature relevant for MCR in the research fields described above. The literature review is not complete, in particular we do not review the fields of computer vision and medical imaging explicitly, because their techniques are used in nearly all publications related to medical image retrieval. However, we are confident that the presented overview reflects the current state of the art and does not miss substantial advances in the MCR field – otherwise they would have been applied to the ImageCLEF medical tasks in 2012.

2.1 Multimedia information retrieval

Although multimedia information retrieval (MIR) has emerged as a separate research field only in the last decade [37, 20], its concepts and techniques originate from other, more traditional fields of information retrieval, most notably from the text retrieval, content-based visual retrieval, and data fusion domains. We therefore present contributions of these fields to MCR as subfields of MIR, although the different research areas of information retrieval are usually not conceived as such. We deliberately ignore content-based audio retrieval, because this is not yet a subject of current research in MCR and not of this PhD project.

2.1.1 Text Retrieval

Text retrieval is the main subject of classical information retrieval. There are many textbooks on this topic; a recent book with 1800 references covering many aspects of information retrieval is [3]. The book by Hersh [28] focuses on retrieval of health care and biomedical information.

Two standard models of text retrieval are the *vector space model* [57] and the *probabilistic model* [54], combined with TF-IDF [64, 52, 73] or BM25 [53] term weighting. These methods are able to deliver state-of-the-art text retrieval performance, and mature open-source implementations are available, most notably Lucene[3] and Indri[4] [43].

The best results for medical case retrieval in ImageCLEF 2011 were obtained by text-only retrieval using fulltext indexing and standard techniques [72, 41]. In 2012, again fulltext indexing using Lucene gave best results [14].

2.1.2 Content-based Visual Retrieval

Datta et al. [13] give a comprehensive overview of research in *content-based image retrieval* (CBIR) during the last decade. The authors define CBIR as "any technology that in principle helps to organize digital picture archives by their visual content". The search paradigm most commonly considered in CBIR research contributions is *query by example*, meaning that an image is available to be used as a query to retrieve relevant "similar" images from a large picture archive. Typically, the user of a CBIR

[3] `http://lucene.apache.org/`
[4] `http://www.lemurproject.org/indri/`

system expects a semantic similarity of images relevant to the query, which depends on the user context and application domain and may not be directly related to the visual similarity of images. This discrepancy is known as the *semantic gap* [62], which is still an open problem in many application domains of CBIR.

The medical imaging domain provides some opportunities that may help bridging the semantic gap, like better defined imaging semantics, rich metadata, and existing knowledge representations. But there are also additional challenges like its interdisciplinary nature, integration of different information sources, and limited availability of training data [79]. A review of CBIR in medical applications and its clinical benefits is given by Müller et al. [46].

From the many facets of CBIR research identified by Datta et al. [13], we focus on the core techniques supporting the basic CBIR process: (1) *feature extraction* represents an image by one ore more vectors of numbers capturing visual properties that are able to discriminate between relevant and non-relevant images, but are also invariant under irrelevant image transformations (e.g. rotation); (2) pattern recognition techniques are used to build *visual signatures* from feature vectors that reduce their dimensionality and aim at representing the desired image semantics, in an effort to bridge the semantic gap; (3) *similarity measures* are applied to visual signatures in order to retrieve (and rank) images that are most similar to a given query image.

A wealth of different image features and corresponding extraction algorithms has been proposed for CBIR [13]. Deselaers et al. [15] performed extensive experimental comparisons between 19 image features on different datasets, including the IRMA dataset of 10,000 medical images. Feature types can be categorized into *global features* describing the visual properties of the entire image by a single feature vector, and *local features* extracted from certain locations or regions in the image. The visual properties captured by feature extraction methods include color, texture, and shape, and many proposed image features represent a combination of these. Among other mathematical models, wavelet transforms are used to represent texture features [18]. A more recent composite image descriptor capturing brightness and texture characteristics for medical image retrieval has been proposed by Chatzichristofis et al. [9].

Whereas global feature vectors are often used directly as visual signatures, local feature vectors of an image need to be summarized to form a signature. The *bag of features* approach applies clustering of local feature vectors of an image collection to construct a codebook of cluster centers (*visual words*), and every image is represented by a term vector of visual words [60], in analogy to text retrieval. Iakovidis et al. [31] build the visual signature by clustering wavelet coefficients and estimating the distributions of clusters using Gaussian mixture models and an expectation-maximization algorithm. They obtain promising medical image retrieval results on the IRMA dataset. Quellec et al. [49] extend the wavelet-based visual signatures of Do and Vetterli [18] by adapting the wavelet basis in order to optimize retrieval performance for a given image collection. They evaluate their approach successfully on two specific homogeneous medical image datasets as well as on a face image dataset.

Another attempt to reduce the semantic gap is to express visual signatures in terms of *semantic concepts* automatically detected in images using pattern recognition techniques. A comprehensive and detailed discussion of concept-based video retrieval is

given by Snoek and Worring [63]. Most of the techniques described there can also be applied to image retrieval. A well-known categorization scheme for diagnostic images is the *IRMA code* [36], classifying the visual content along four dimensions: image modality (e.g. X-ray, ultrasound, CT, MR), body orientation, body region, and biological system. IRMA categories may serve as concepts to build semantically meaningful visual signatures.

Rahman et al. [51] proposed a concept-based image retrieval framework utilizing class probabilities of multiple classifiers as visual signatures and cosine similarity for retrieval. Class probabilities are estimated from binary SVM classifiers. For different low-level visual feature spaces, concept-based similarity values are calculated separately and fused using a linear combination scheme where weights are optimized adaptively for each query. Weight optimization incorporates automatic relevance estimation based on classifier fusion over low-level feature spaces, but may also include user relevance feedback. The framework was evaluated on the ImageCLEF 2006 medical dataset using 116 IRMA categories and 4 low-level visual features (MPEG-7 Edge Histogram and Color Layout, GLCM-based texture features, and block-based gray values). In 2011, the authors proposed a similar retrieval scheme [50].

The visual signature of a query image needs to be compared to that of images in the collection to retrieve the "most similar" ones. The underlying assumption is that similarity of visual signatures is correlated with semantic relevance. Failure of this assumption indicates that the semantic gap has not been bridged sucessfully. Similarity of visual signatures is computed by applying an appropriate *similarity measure*. Eidenberger [19] conducted an extensive experimental comparison and analysis of many well-known similarity measures used for CBIR.

Güld et al. [26] describe a generic *framework for medical image retrieval systems* developed by the IRMA project [35]. The framework aims at enabling flexible and efficient development and deployment of retrieval algorithms in a distributed environment with web-based user interfaces. Demo applications using this framework are available online[5].

Zhou et al. [79] propose a framework for content-based medical image retrieval on a semantic level. They emphasize the need for a scalable semantic retrieval system (e.g. easily adaptable to different image modalities and anatomical regions) and for incorporating external knowledge. An architecture for integrated (symbolic and sub-symbolic) image feature extraction and semantic reasoning is proposed. As a prototype implementation, they describe a semantic anatomy tagging engine called ALPHA, which employs a novel approach to deformable image segmentation by combining hierarchical shape decomposition and CBIR.

For the *ImageCLEF medical case retrieval task*, purely visual retrieval generally gave poor performance in recent years. In 2011, the best approach [41] achieved only 2% MAP. It used the GNU Image Finding Tool[6] (GIFT) [65] and was based on color histogram intersection and texture features obtained from Gabor filters, weighted using a standard TF-IDF scheme. The results from querying multiple images per case were

[5] `http://irma-project.org/onlinedemos.php`
[6] `http://www.gnu.org/software/gift/`

fused using a score-based combSUM strategy [77]. In 2012, best visual retrieval was achieved using late inverse rank fusion of SIFT-based bag-of-visual-words (BoVW) and bag-of-colors (BoC) features [14], yielding 3.7% MAP.

A Java library supporting content-based text and image retrieval is LIRE[7] [39, 40]. It provides a number of different global and local image feature extractors and efficient indexing techniques for images and text based on Lucene[8].

2.1.3 Data Fusion

Data fusion (also known as information fusion or meta-search) is a well-known research field of information retrieval. The main objective is to combine multiple information sources to improve retrieval performance. Depending on the phase of the retrieval process chain where the combination happens, different *fusion levels* can be distinguished [67]: signal level, feature level, and decision level. Signal- and feature-level fusion are also called *early fusion*, whereas decision-level fusion is also known as *late fusion*, which aims at combining the results of multiple retrieval systems.

In the context of MCR, late fusion is of particular interest, because it allows for *multi-modal fusion* of text and visual retrieval systems. Late fusion approaches can be categorized into *score-based* and *rank-based* methods, according to which information from retrieval result lists is used (score or rank). Wu [74] gives a concise overview of known methods of both categories and proposes a new weight optimization method for linear score combination based on the multiple linear regression technique. Moreover, the author addresses another important issue of score-based data fusion systems, namely how to obtain reliable scores from score or rank information provided by component systems (*score normalization*). The logistic and cubic regressions models are found to provide reliable solutions to the score normalization problem. The proposed approach is evaluated on several text retrieval datasets of recent TREC challenges.

Zhou et al. [77] investigated and generalized the classical score combination methods combMAX, combSUM, and combMNZ [23] for single- and multi-modal fusion of the 8 best runs submitted to the ImageCLEF medical image retrieval tasks in 2008 and 2009. They conclude that logarithmic rank penalization is the most stable score normalization strategy, but there is no significant difference between the various score combination methods considered.

Gkoufas et al. [24] evaluated linear combination methods using multi-field textual retrieval and visual retrieval built into LIRE [39] on the MCR datasets of ImageCLEF 2009 and 2010. Fusion of textual and visual retrieval could not improve retrieval performance (MAP) over fulltext-only retrieval on the ImageCLEFmed 2009 and 2010 datasets, only precision at 5 and 10 increased slightly.

In ImageCLEF 2011, the best result of fusing text and visual retrieval decreased performance with respect to text-only retrieval (see Section 1). The fusion strategy was combSUM. A similar result was obtained in 2012 for combining Lucene fulltext indexing with the visual retrieval approach based on BoVW and BoC features [14].

[7] `http://www.semanticmetadata.net/lire/`
[8] `http://lucene.apache.org/`

A different approach to data fusion is *filtering*, where component retrieval systems are used in a pipeline fashion such that a subsequent system works on a reduced dataset that has been filtered by the previous system (i.e. irrelevant documents have been filtered out). Usually, filtering is applied in combination with other fusion techniques. Such an approach for medical image retrieval using an IRMA code classifier for filtering is proposed by Rahman et al. [50]. A more general approach using text retrieval as the filtering stage and locality-sensitive hashing for visual retrieval is proposed by Zhang et al. [76].

2.1.4 Relevance Feedback

Relevance feedback (RF) has been an active research field in multimedia information retrieval for several decades [78, 13], because it attempts to address the semantic gap problem by incorporating relevance judgments from users. Algorithmic approaches to RF can be categorized as *short-term learning* and *long-term learning* techniques [78], depending on the desired effect of user feedback on retrieval results: short-term learning affects the current query only [34], whereas long-term learning aims at improving retrieval performance for future queries [11]. More recent approaches include a probabilistic RF framework processing multiple image queries consisting of both positive and negative samples [2] for short-term learning, and a semi-supervised long-term learning algorithm [75].

Many RF methods utilize relevance judgments of users as additional training data for machine learning. Depending on whether also unlabeled training data are used for learning, *inductive* (using only labeled training data) and *transductive* methods can be distinguished. A prominent technique for transductive RF is manifold-ranking [27], a recent extension using random walks has been proposed by Rota Bulò et al. [55].

RF learning methods have to cope with the small sample size problem, because the number of training samples provided by relevance feedback is usually too small to reliably improve prediction performance for most learning algorithms. It is therefore desireable that the system selects samples for relevance feedback that, when labeled by the user, yield maximal performance improvement for the learning algorithm with respect to some optimization criterion. This is exactly the problem addressed by the *active learning* literature [70, 59]. However, choosing the most informative samples will most likely not coincide with the most positive samples the user is interested in, so active learning techniques applied to iterative short-term learning often rely on the user's patience [78]. Active learning may therefore be more interesting for long-term learning.

2.2 Knowledge Representation

In the context of information retrieval, *external knowledge* is an information source that is not available in the dataset or query, but could be utilized to improve retrieval performance. There are two main techniques to achieve this aim: *query expansion* [6, 8] and *multi-label annotation* [66]. Both techniques may incorporate external knowledge in the form of an *ontology* [25], which specifies concepts, relationships, and other

distinctions that are relevent for modeling a domain. In the medical domain, many ontologies have been developed to store and classify medical knowledge [4, 16]. Some of them are UMLS[9] [7], SNOMED, ICD, RadLex, and MeSH[10].

Query expansion using the MeSH ontology has been applied to MCR with varying success. Diaz-Galiano et al. [16] observed a significant increase in retrieval performance on the ImageCLEF 2005 and 2006 MCR datasets, whereas Mata et al. [42] could not using the ImageCLEF 2011 dataset. However, the latter authors reported a more successful approach in [12].

Multi-label annotation [66] employs machine learning techniques to automatically assign several, possibly related semantic concepts to (multimedia) documents. This can improve retrieval performance if the annotated concepts are relevant to the query and add information to documents (i.e. the annotated label is not already contained in text documents). If the possible labels are organized in a tree structure, multi-label classification specializes to *hierarchical multi-label classification* (HMC). Dimitrovski et al. [17] propose an HMC classifier for medical image annotation based on ensembles of predictive clustering trees. They evaluate their approach on the ImageCLEF 2007 and 2008 medical image annotation datasets (using IRMA code labels), outperforming both non-hierarchical multi-label classifiers based on support vector machines (SVMs) and single-classifier HMC approaches. Fan et al. [21] propose an HMC classifier for video concept annotation using a concept ontology. They evaluate their approach in the domain of surgery education videos, where concepts are linked to features derived from salient objects [38].

3 Aim and Objectives

As explained in the previous sections, *medical case retrieval* (MCR) is a relevant problem in computer science whose known solutions for large and heterogeneous datasets are too ineffective to be of practical value. This PhD project therefore aims at designing a model for MCR that is able to deliver a substantially better retrieval performance on such datasets than known solutions. Moreover, to be usable in practice, the proposed techniques need to be both efficient and robust to be applicable to large medical datasets of diverse content.

Considering the retrieval methods used for the ImageCLEF MCR task in 2012 (see Section 1), the most obvious potential for improvement seems to be in data fusion methods combining different modalities of case representation. The most informative modality is *text*, but due to their frequency, also different *image modalities* (e.g. diagram, X-ray, computer tomography, magnetic resonance, ultrasound) should be used for retrieval. Moreover, *external knowledge* in the form of medical ontologies or *expert user feedback* could be used as additional information sources, which can be considered as separate modalities of information. The focus of this work will therefore be on *multimodal* approaches to MCR. However, in order to achieve successful data fusion

[9] `http://www.nlm.nih.gov/research/umls/`
[10] `http://www.nlm.nih.gov/mesh/`

with visual modalities, improvement of visual-only retrieval performance is a necessary goal, too.

The main difficulty of solving the MCR problem (and other multimedia information retrieval problems) is to bridge the *semantic gap* between the low-level case representation and the high-level meaning of case similarity (see Section 2.1.2). One way to address this problem is to automatically detect semantically meaningful *concepts* from low-level document features using pattern recognition techniques. It is hoped that semantic case similarity can be expressed in a simpler and more robust way in terms of concepts than in terms of document features.

Medical case retrieval on heterogeneous datasets faces some difficulties that are not so severe in other application domains of multimedia information retrieval (cf. [79]):

- Semantically similar cases are related to images with very different low-level visual properties. On the other hand, visually similar images may belong to semantically different cases. Concept detection consequently has to cope with a high intra-class variability and a low inter-class variability.
- Annotation of medical training samples is expensive, so training datasets for supervised machine learning approaches are small and often imbalanced.
- Heterogeneous datasets contain a large number of medical concepts needed to describe the semantics of medical cases. So hand-crafted detection of a few concepts is not an option.

Pattern recognition and, in particular, concept detection supporting MCR systems should therefore also try to leverage the expert knowledge of its users (usually physicians). This may happen by improving retrieval results of the current query using *relevance feedback* techniques, also known as *short-term learning* (see Section 2.1.4). In light of the case-based reasoning process described in Section 1, however, *long-term learning* techniques using input from expert users to improve retrieval for future queries appear to be even more relevant.

Motivated by the previous considerations, this PhD project will address the following research objectives:

O1 Determine the reasons for the moderate retrieval performance of current multimodal techniques on the ImageCLEF MCR dataset.

O2 Design a novel MCR model combining different modalities of case representations and information sources (without relevance feedback) to enable a substantial improvement of retrieval performance on the ImageCLEF MCR dataset, achieving at least 30% MAP.

O3 Using the system resulting from O2, investigate the potential of further improvement of retrieval performance by long-term learning from medical expert users.

As there is no comprehensive comparative study of known MCR techniques evaluated on the ImageCLEF dataset in the literature, pursuing **O1** requires selecting, implementing, evaluating, and analyzing some of the most promising known approaches to

MCR. Moreover, to understand the reasons for their retrieval performance, statistical properties of features extracted from the ImageCLEF dataset need to be investigated.

Designing a better MCR model according to **O2** includes the sub-problems of choosing appropriate features to represent visual information from images and detecting medical concepts such that visual-only retrieval improves over known approaches (see Section 2.1.2). Additionally, the utilization of medical ontologies as external information source is an important sub-problem (Section 2.2). The combination of different modalities needs to be addressed by appropriate data fusion methods (Section 2.1.3).

Proper evaluation of a long-term learning system according to **O3** would require a user study with several (e.g. 20) representative medical experts and an appropriate experimental design to derive statistically significant results. Due to the difficulty and costs of enlisting so many medical expert users, evaluation of the system according to **O3** will simulate experts by using part of the relevance judgments (ground truth) provided with the ImageCLEF MCR dataset.

4 Previous Work

We built a baseline case-based retrieval system for the ImageCLEFmed 2012 dataset, using LIRE and the Lucene text retrieval engine (see Section 2.1.2). Our fulltext-only retrieval achieved 16.3% MAP, which is close to the best retrieval performance achieved for the ImageCLEFmed 2012 case-based retrieval task (17%).

5 Evaluation

In the information retrieval research field, *evaluation methods* have a long history [68]. The system evaluation methodology (as opposed to user evaluation) used by the Text Retrieval Evaluation Campaign (TREC) [69] since the 1990s dates back to the Cranfield paradigm [10] developed in the 1950s. The basic evaluation process is: acquire a dataset, publish search topics to participants, use manual relevance judgments and precision-recall-based metrics to evaluate submitted ranking lists. This methodology is followed by the majority of evaluation campaigns in the (multimedia) information retrieval field [61]. Notable benchmarking campaigns besides TREC and its spin-offs (e.g. TRECVID) are CLEF (Cross Lingual Evaluation Forum), ImageCLEF, ImageEval, MediaEval, and INEX (Initiative for the Evaluation of XML Retrieval). General benefits and disadvantages of evaluation campaigns are discussed by Smeaton et al. [61]. A more general literature survey on retrieval evaluation using test collections is given by Sanderson [58]. Precision-recall metrics are critically treated by Huijsmans and Sebe [30]. They explain the limitations of precision-recall graphs and develop additional improved performance indicators for information retrieval.

The systems to be developed for pursuing the research objectives described in Section 3 will be evaluated using the TREC methodology as required by the ImageCLEF organizers. Retrieval runs are evaluated using NIST's `trec_eval` tool[11] and the pro-

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

vided experts' relevance judgments (ground truth). From the various performance metrics computed by `trec_eval`, the single *mean average precision* value (MAP, see e.g. [3]) will be the most important to compare different retrieval systems.

When evaluating the long-term learning system according to objective **O3**, part of the relevance judgments will be used as training data, so evaluation must use only the remaining relevance judgments. To achieve more robustness of results with respect to the choice of training samples, a *cross-validation method* will be employed [71].

6 Methodology

As described in Section 3, this PhD project aims at designing a "better" model of medical case retrieval. So the anticipated research output will be an artifact (the model) and a report on proper evaluation and comparison with known methods. This kind of research is common in computer science and is the subject of *design science research methodologies* (DSRM) developed by the information systems research community [29, 47]. The DSRM proposed by Peffers et al. [47] divides the research process into six activities:

1. *Problem identification and motivation:* define the specific research problem and justify the value of a solution.
2. *Objectives for a solution:* infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible.
3. *Design and development:* create the artifact (constructs, models, methods, or instantiations).
4. *Demonstration:* demonstrate the use of the artifact to solve one or more instances of the problem.
5. *Evaluation:* observe and measure how well the artifact supports a solution to the problem.
6. *Communication:* communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences.

The authors also propose to structure the research output (publications) according to these activities, and provide a number of case studies to demonstrate that existing publications can be interpreted using this DSRM model.

We want to apply this research methodology to the proposed PhD project and already structured this proposal according to the DSRM model described above. The research problem is identified and motivated in Section 1. Defining the objectives (Section 3) requires knowledge of the current state of the art, which is therefore described before (Section 2). Design and development will be the main workload of the PhD project, but a baseline system has already been built (Section 4). The created model

Table 1: PhD project time table.

Activity	**Deadline**
Topic selection	Sep 2012
PhD proposal	Mar 2013
Investigate known MCR techniques (**O1**)	Nov 2013
Develop better MCR model (**O2**)	Feb 2015
Incorporate long-term learning (**O3**)	Feb 2016
Complete PhD thesis	Sep 2016

will be implicitly demonstrated by system evaluation (Section 5), which is a common pattern observable in many information retrieval publications. The research output will be communicated by means of the PhD thesis and intermediate publications.

7 Project Management

A rough project time table is shown in Table 1. The applicable version of PhD curriculum at AAU is 2009W, which expires in Sep 2017. There is a risk of limited improvement of MCR techniques due to properties of the dataset (too heterogeneous). In this case, a different, less heterogeneous dataset will be chosen, e.g. by filtering the ImageCLEF MCR dataset.

During the project, repeated participation in the yearly ImageCLEF MCR tasks is planned. Every major project phase (**O1**–**O2**) is expected to emit a workshop or conference publication.

References

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[2] M. Arevalillo-Herráez, F. J. Ferri, and J. Domingo. A naive relevance feedback model for content-based image retrieval using multiple similarity measures. *Pattern Recogn.*, 43(3):619–629, Mar. 2010.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2011.

[4] M. Batet, D. Sánchez, and A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1):118–125, 2011.

[5] S. Begum, M. Ahmed, P. Funk, N. Xiong, and M. Folke. Case-based reasoning systems in the health sciences: A survey of recent trends and developments. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(4):421–434, July 2011.

[6] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manag.*, 43(4):866–886, July 2007.

[7] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl. 1):D267–D270, 2004.

[8] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, Jan. 2012.

[9] S. Chatzichristofis and Y. Boutalis. Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools and Applications*, 46:493–519, 2010.

[10] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 3–12, New York, NY, USA, 1991. ACM.

[11] M. Cord and P. H. Gosselin. Image retrieval using long-term semantic learning. In *Image Processing, 2006 IEEE International Conference on*, pages 2909–2912. IEEE, 2006.

[12] M. Crespo, J. Mata, and M. Maña. Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. *Journal of the American Medical Informatics Association [online]*, Sept. 2012. DOI 10.1136/amiajnl-2012-000943.

[13] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008.

[14] A. G. S. de Herrera, D. Markonis, I. Eggel, and H. Müller. The medGIFT group in ImageCLEFmed 2012. In Forner et al. [22].

[15] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11(2):77–107, Apr. 2008.

[16] M. C. Díaz-Galiano, M. Martín-Valdivia, and L. A. Ureña López. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Comput. Biol. Med.*, 39(4):396–403, Apr. 2009.

[17] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Deroski. Hierarchical annotation of medical images. *Pattern Recogn.*, 44(10-11):2436–2449, Oct. 2011.

[18] M. N. Do and M. Vetterli. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *Image Processing, IEEE Transactions on*, 11(2):146–158, 2002.

[19] H. Eidenberger. Evaluation and analysis of similarity measures for content-based visual information retrieval. *Multimedia Systems*, 12(2):71–87, 2006.

[20] H. Eidenberger. *Handbook of Multimedia Information Retrieval*. Books on Demand, Norderstedt, Germany, 2012.

[21] J. Fan, H. Luo, Y. Gao, and R. Jain. Incorporating concept ontology for hierarchical video classification, annotation, and visualization. *Trans. Multimedia*, 9(5):939–957, Aug. 2007.

[22] P. Forner, J. Karlgren, and C. Womser-Hacker, editors. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012.

[23] E. A. Fox and J. A. Shaw. Combination of multiple searches. *NIST special publication*, (500215):243–252, 1994.

[24] Y. Gkoufas, A. Morou, and T. Kalamboukis. Combining textual and visual information for image retrieval in the medical domain. *Open Medical Informatics Journal*, 5:50–57, 2011.

[25] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5):907–928, 1995.

[26] M. O. Güld, C. Thies, B. Fischer, and T. M. Lehmann. A generic concept for the implementation of medical image retrieval systems. *International Journal of Medical Informatics*, 76(23):252 – 259, 2007.

[27] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 9–16. ACM, 2004.

[28] W. R. Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Health informatics. Springer, 3rd edition, 2009.

[29] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Q.*, 28(1):75–105, Mar. 2004.

[30] D. P. Huijsmans and N. Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):245–251, Feb. 2005.

[31] D. Iakovidis, N. Pelekis, E. Kotsifakos, I. Kopanakis, H. Karanikas, and Y. Theodoridis. A pattern similarity scheme for medical image retrieval. *Information Technology in Biomedicine, IEEE Transactions on*, 13(4):442 –450, july 2009.

[32] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika. Overview of the CLEF 2011 medical image classification and retrieval tasks. In Petras et al. [48].

[33] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *British Medical Journal*, 330(7494):765, 2005.

[34] A. Kushki, P. Androutsos, K. N. Plataniotis, and A. N. Venetsanopoulos. Query feedback for interactive image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):644–655, 2004.

[35] T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. B. Wein. Content-based image retrieval in medical applications. *Methods of Information in Medicine*, 43(4):354–361, 2004.

[36] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein. The IRMA code for unique classification of medical images. In *Medical Imaging 2003*, pages 440–451. International Society for Optics and Photonics, 2003.

[37] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, Feb. 2006.

[38] H. Luo, J. Fan, Y. Gao, and G. Xu. Multimodal salient objects: General building blocks of semantic video concepts. In P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. Smeulders, editors, *Image and Video Retrieval, Proc. CIVR*, volume 3115 of *Lecture Notes in Computer Science*, pages 374–383. Springer, 2004.

[39] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible Java CBIR library. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 1085–1088, New York, NY, USA, 2008. ACM.

[40] M. Lux and O. Marques. Visual information retrieval using java and lire. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(1):1–112, Jan. 2013.

[41] D. Markonis, I. Eggel, A. G. S. de Herrera, and H. Müller. The medGIFT group in ImageCLEFmed 2011. In Petras et al. [48]. `http://clef2011.org/resources/proceedings/Markonis_Clef2011.pdf`, visited in July 2012.

[42] J. Mata, M. Crespo, and M. J. Maña. Using MeSH to expand queries in medical image retrieval. In *Proc. MICCAI, Medical Content-Based Retrieval for Clinical Decision Support*, MCBR-CDS'11, pages 36–46. Springer, 2012.

[43] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing & Management*, 40(5):735 – 750, 2004. Special Issue on Bayesian Networks and Information Retrieval.

[44] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *ImageCLEF – Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer Berlin Heidelberg, 2010.

[45] H. Müller, A. G. S. de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In Forner et al. [22].

[46] H. Mller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applicationsclinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1 – 23, 2004.

[47] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *J. Manage. Inf. Syst.*, 24(3):45–77, Dec. 2007.

[48] V. Petras, P. Forner, and P. D. Clough, editors. *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*, 2011.

[49] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux. Wavelet optimization for content-based image retrieval in medical databases. *Medical Image Analysis*, 14(2):227 – 241, 2010.

[50] M. Rahman, S. Antani, and G. Thoma. A learning-based similarity fusion and filtering approach for biomedical image retrieval using SVM classification and relevance feedback. *IEEE Trans. Inf. Tech. Biomedicine*, 15(4):640–646, July 2011.

[51] M. M. Rahman, B. C. Desai, and P. Bhattacharya. Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. *Computerized Medical Imaging and Graphics*, 32(2):95 – 108, 2008.

[52] S. Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.

[53] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, Apr. 2009.

[54] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[55] S. Rota Bulò, M. Rabbi, and M. Pelillo. Content-based image retrieval with relevance feedback using random walks. *Pattern Recognition*, 44(9):2109–2122, 2011.

[56] D. Sackett, W. Rosenberg, J. Gray, R. Haynes, and W. Richardson. Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72, 1996.

[57] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[58] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[59] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of WisconsinMadison, Jan. 2010.

[60] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[61] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[62] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, Dec. 2000.

[63] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2(4):215–322, Apr. 2009.

[64] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[65] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(1314):1193 – 1198, 2000. Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99.

[66] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[67] L. Valet, G. Mauris, and P. Bolon. A statistical overview of recent literature in information fusion. *IEEE Aerospace and Electronic Systems Magazine*, 16(3):7–14, 2001.

[68] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370, London, UK, UK, 2002. Springer-Verlag.

[69] E. M. Voorhees and D. K. Harman, editors. *TREC : experiment and evaluation in information retrieval*. MIT Press, Cambridge, Mass., 2005.

[70] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.*, 2(2):10:1–10:21, Feb. 2011.

[71] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

[72] H. Wu and C. Tian. UESTC at ImageCLEF 2011 medical retrieval task. In Petras et al. [48]. `http://clef2011.org/resources/proceedings/Wu_Clef2011.pdf`, visited in July 2012.

[73] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3):13:1–13:37, June 2008.

[74] S. Wu. Linear combination of component results in information retrieval. *Data Knowl. Eng.*, 71(1):114–126, Jan. 2012.

[75] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):723–742, 2012.

[76] N. Zhang, K. L. Man, T. Yu, and C.-U. Lei. Text and content based image retrieval via locality sensitive hashing. *Engineering Letters*, 19(3):228–234, 2011.

[77] X. Zhou, A. Depeursinge, and H. Müller. Information fusion for combining visual and textual image retrieval. In *Proceedings of the 20th International Conference on Pattern Recognition (2010)*, ICPR '10, pages 1590–1593, Washington, DC, USA, 2010. IEEE Computer Society.

[78] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544, 2003.

[79] X. S. Zhou, S. Zillner, M. Moeller, M. Sintek, Y. Zhan, A. Krishnan, and A. Gupta. Semantics and CBIR: a medical imaging perspective. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 571–580, New York, NY, USA, 2008. ACM.